# Principe MDL et choix de modèles

Aurélien Garivier

Université Paris Sud

Orsay

# Plan de l'exposé

- Théorie de l'information et MDL
- Chaînes de Markov d'ordre variable
- HMM à émission continue

# Codage source

- Source $\mathbb{P}$ stationnaire sur l'alphabet $A$.
- Pb: transmettre les messages issus de cette source:
- Fonction de codage $\phi_n : A^n \longrightarrow \{0,1\}^*$, on veut minimiser la longueur de code moyenne des messages

$$\mathbb{E}_{\mathbb{P}}\left[|\phi_n(x)|\right]$$

# Entropie

- Théorème de Shannon ('48) :

$$\mathbb{E}_{\mathbb{P}}\left[|\phi_n(x)|\right] \geqslant H_n(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[-\log \mathbb{P}(X_1^n)\right],$$

  à peu près atteignable.

- Par ailleurs, si la source est ergodique

$$\frac{1}{n} H_n(\mathbb{P}) \longrightarrow H(\mathbb{P})$$

  *taux entropique* de la source = nb de bits nécessaires au codage de chaque caractère émis.

- Code $\phi_n(x) \leftrightarrow$ loi de codage $q_n = 2^{-|\phi_n(x)|}$
  $\Rightarrow -\log q_n(x)$=*longueur de code*

# Codage universel

- Codeur et décodeur savent juste $\mathbb{P} \in \mathcal{S} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.
- Ex: Markov chains, general stationary ergodic processes.
- Il faut *une seule* loi de codage $q_n$ pour toutes les sources $\mathbb{P}_\theta$
  $\Rightarrow$ surcoût appelé *redondance*

$$R_n(q_n, \theta) = \mathbb{E}_{\mathbb{P}_\theta}\left[|\phi_n(X)|\right] - H_n(\mathbb{P}_\theta) = KL(\mathbb{P}^n | q_n)$$

# Redondance minimax

- L'universalité du codeur $q_n$ est mesurée par la redondance dans le pire des cas.
- La meilleure possible est la *redondance minimax*

$$R_n(\mathcal{S}) = \inf_{q_n} \sup_{\theta \in \Theta} R_n(q_n, \theta)$$

- Th: pour les classes $\mathcal{S}$ paramétriques à $k$ degrés de liberté, Rissanen (1984) : $R_n(\mathcal{S}) = \frac{k}{2} \log n + O(1)$ $\Rightarrow$ il faut au moins

$$nH(\mathbb{P}_\theta) + \frac{k}{2} \log n$$

bits en moyenne pour coder un message de taille $n$.

# Choix de modèle

On dispose d'une séquence $x$, réalisation d'une va dont la loi est dans un des modèles $\mathcal{S}_0$, $\mathcal{S}_1 \ldots$

Pb: identifier ce modèle à partir de l'observation.

Exemple :

- chaîne ADN x=ACCACTGACTAGACCT. . . vient d'une chaîne de Markov d'ordre 0, 1, 2, . . . ?
- suite de nombre réels provenant d'un mélange d'un nombre inconnu de composantes gaussiennes.

# Principe MDL

Minimum Description Length : “choisis le modèle qui donne *la plus courte description des données*” (Rasoir d’Occam).

$\Rightarrow$ Soit $k_i = \dim \mathcal{S}_i$, à partir de $x$ :

- on calcule $\widehat{\theta}_i(x)$ = estimateur du paramètre dans le modèle $\mathcal{S}_i$ (ex: maximum de vraisemblance).
- on choisit le modèle minimisant

$$nH(\mathbb{P}_{\widehat{\theta}_i(x)}) + \frac{k_i}{2} \log n$$

- $\approx$ maximum de vraisemblance pénalisé. Pénalité BIC.

# Plan de l’exposé

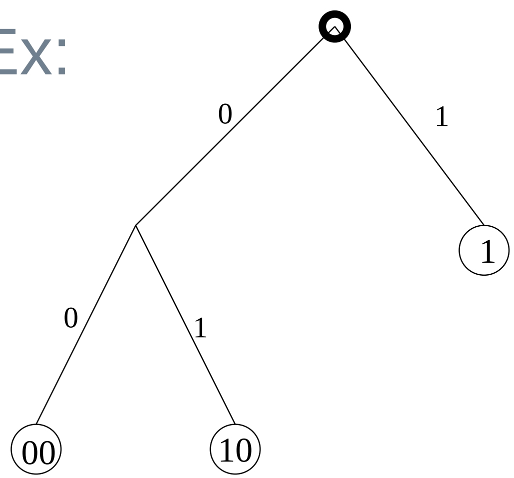
- Théorie de l’information et MDL
- Chaînes de Markov d’ordre variable
- HMM à émission continue

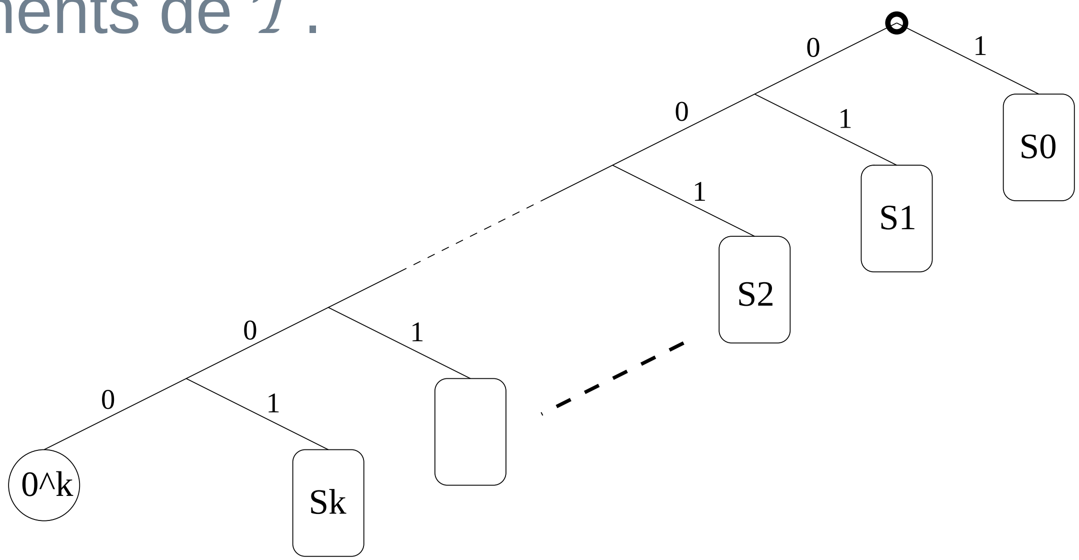
# Dictionnaire de suffixes complet

- $\mathcal{T}$ est un *Dictionnaire de suffixes complet* (DSC) si

$$\forall x_{-\infty}^{0} \in A^{\mathbb{Z}-}, \exists! k \in \mathbb{N} : x_{-k}^{0} \in \mathcal{T}.$$

- Si $x_{-\infty}^{0} \in A^{\mathbb{Z}-}$, on note $\mathcal{T}(x)$ son suffixe dans $\mathcal{T}$.
- Un DSC peut être représenté par un trie dont les feuilles sont les éléments de $\mathcal{T}$.
- Ex:




$$\mathcal{T} = \{00, 10, 1\} \qquad \mathcal{T} = \{0^k\} \cup \{10^j : 0 \leq j < k\}$$

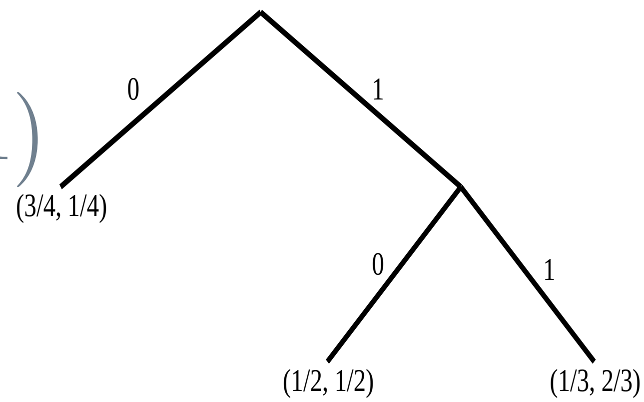
# VLMC

- Soit $\mathcal{T}$ un DSC et soit $p = (p(.|w))_{w \in \mathcal{T}}$ un $|\mathcal{T}|$-uplet de lois de probabilité sur $A$.
- La *chaîne de Markov d'ordre variable* $\mathbb{P}_{\mathcal{T},p}$ est la distribution stationaire sur $A^{\mathbb{Z}}$ définie par

$$\mathbb{P}_{\mathcal{T},p}\left(X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0\right) = p\left(x_1 | \mathcal{T}\left(x_{-\infty}^0\right)\right).$$

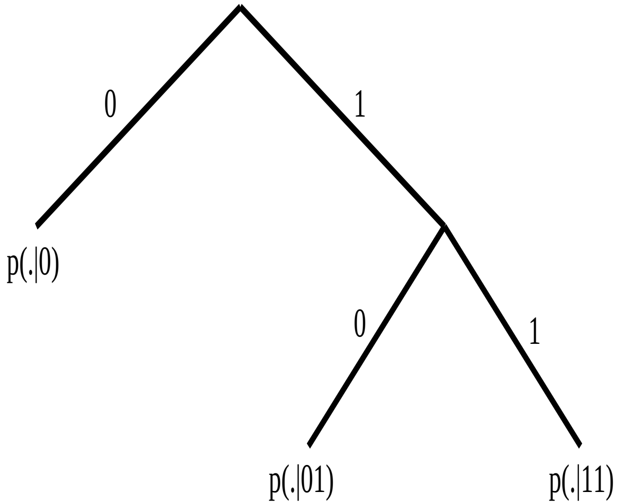
- Ex:

$$\mathbb{P}\left(X_1^4 = 1001 | X_{-\infty}^0 = \ldots 01\right)$$
$$= \frac{1}{2} \times \frac{1}{3} \times \frac{3}{4} \times \frac{1}{4}$$



# Les VLMC sont des chaînes de Markov

- Profondeur du trie = ordre markovien



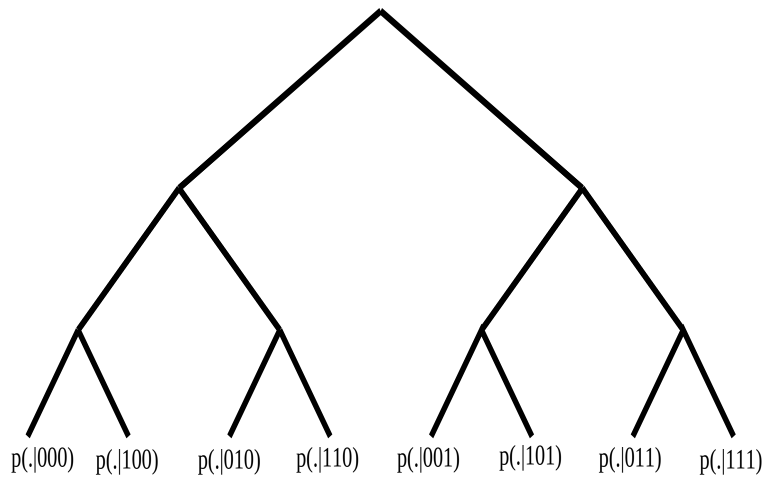

$$\rightarrow M = \begin{pmatrix} p\left(.\middle|0\right) \\ p\left(.\middle|0\right) \\ p\left(.\middle|01\right) \\ p\left(.\middle|11\right) \end{pmatrix}$$

- *Variable Length* Markov Chains : potentiellement moins de paramètre pour une mémoire donnée

# Les chaînes de Markov sont des VLMC…

- …correspondant à un arbre complet:

$$M = \begin{pmatrix} p\,(.|000) \\ p\,(.|100) \\ \vdots \\ p\,(.|111) \end{pmatrix} \longrightarrow$$




- $\Longrightarrow$ les VLMC combinent le pouvoir d'approximation des chaînes de Markov avec avec une grande souplesse. approach every stationary ergodic source.
- Elles ne sont pas plus compliquées à utiliser.

# Esitmateur de modèle BIC

- L'estimateur BIC pour $x \in A^n$ s'écrit :

$$\widehat{T_{BIC}} = \arg\min_{\mathcal{T}} \sum_{s \in \mathcal{T}} H(\mathcal{T}(x, s)) + \frac{|T|\,(|A| - 1)}{2} \log n$$

- Th: (Csiszár & Talata, Garivier) : $\widehat{T_{BIC}}$ est un estimateur *consistant*.
- Rq: il existe une procédure qui calcule $\widehat{T_{BIC}}$ en temps linéaire.

# Plan de l'exposé

- Théorie de l'information et MDL
- Chaînes de Markov d'ordre variable
- HMM à émission continue

# Chaînes de Markov cachées

- chaque état caché a une loi d'émission differente.
- on passe d'un état caché à l'autre par un processus markovien.
- à chaque étape, on émet un caractère.
- Ex:

# Inégalités de mélange

Si $A=\mathbb{N}$ ou $A=\mathbb{R}$, on ne peut appliquer la théorie de l'information classique, mais on peut montrer des inégalités *de type BIC*. En choisissant pour $q_n^k$ un *mélange* de toutes les lois de transition et d'émission possible, on obtient

- Cas Poisson :

$$0 \leqslant \sup_{\theta\in\Theta_k} \log \mathbb{P}_\theta(X_1^n) - \log q^k(X_1^n) \leqslant \frac{k^2}{2}\log n + \frac{k}{2\tau^2}|X|_{(n)}^2 + c_{kn} + d \tag{1}$$

- Cas gaussien :

$$0 \leqslant \sup_{\theta\in\Theta_k} \log f_\theta(X_1^n) - \log q^k(X_1^n) \leqslant \frac{k^2}{2}\log n + k\tau X_{(n)} + c_{kn} + e_{kn}. \tag{2}$$

# Chaînes de Markov cachées

- De façon analogue à BIC, on définit l'estimateur :

$$\widehat{k} = \arg\min_{k \geqslant 1} \left\{ - \sup_{\theta \in \Theta_k} \log g_\theta(X_1^n) + \mathrm{pen}(n, k) \right\}$$

- **Théorème :** Le vrai ordre est identifié à partir d'un certain rang dès que pour $n \geqslant 3, k \geqslant 1$,

$$\mathrm{pen}(n, k) = \sum_{\ell=1}^{k} \frac{\ell^2 + \alpha}{2} \log n + R_{kn},$$

# Commentaires

- Rq: on peut faire la même chose avec des états cachés i.i.d. (mélange), en remplaçant les facteurs $k^2$ par $2k-1$.
- Avantages : pas besoin de borne a priori sur l'ordre ni sur les paramètres des lois d'émission.
- Inconvénients : en pratique la vraisemblance n'est pas facile à maximiser.

# Micro-bibliography


- **Universal modeling and coding** - Rissanen, Langdon, IEEE-IT 1981
- **Universal coding, Information, Prediction, and Estimation** -Rissanen, IEEE-IT 1984
- **The Context-Tree Weighting Method : basic Properties** - Willems, Shtarkov, Tjalkens IEEE-IT 1995
- **Redundancy Rates for Renewal and Other Processes** - Csiszár, Shields - IEEE-IT 1996
- **Variable length Markov chains** - Bühlmann, Wyner, Abraham, Annals of Statistics 1999
- **Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL.** - Csiszár, Talata (Budapest), IEEE-IT 2004