# Utilisation de matrices de Hankel non bornées pour l'apprentissage spectral de langages stochastiques

Mattias Gybels*[1], François Denis†[1], et Amaury Habrard‡[2]

[1]LIF, Aix-Marseille Université, CNRS
[2]Laboratoire Hubert Curien, Université de Saint-Etienne, CNRS

1er juin 2013

## Résumé

Un problème de base en inférence grammaticale consiste à inférer un modèle probabiliste, par exemple sous la forme d'un automate pondéré, à partir d'un échantillon $S$ de chaînes tirées indépendamment selon une distribution cible $p$. Des avancées récentes - les méthodes spectrales - reformulent cette tâche comme un problème d'algèbre linéaire : le modèle inféré se calcule aisément à partir d'une décomposition en valeurs singulières tronquée d'une matrice $H$, appelée matrice de Hankel, qui résume l'information contenue dans l'échantillon et dont les lignes $U$ et les colonnes $V$ sont indexées par des chaînes. Les performances du modèle dépendent à la fois de la distance entre la matrice de Hankel réelle et sa version empirique calculée à partir de $S$ ainsi que du choix des ensembles indexant la matrice. Les approches existantes se basent sur des ensembles $U$ et $V$ de taille finie, généralement petite, et les bornes de concentration qui sont invoquées sur la différence entre les matrices de Hankel empirique et réelle dépendent de ces tailles. Nous proposons dans cet article une borne de concentration indépendante des tailles de $U$ et de $V$ qui laisse penser qu'il n'y a pas d'inconvénient majeur à ne pas borner a priori ces tailles. Nous fournissons des comptes-rendus d'expériences dans lesquelles nous comparons les résultats obtenus à partir de différentes versions de la matrices de Hankel empirique montrant l'intérêt d'utiliser des ensembles $U$ et $V$ non bornés.

**Mots-clef** : Apprentissage automatique, inférence grammaticale, méthodes spectrales.

*mattias.gybels@lif.univ-mrs.fr
†francoisdenis@lif.univ-mrs.fr
‡amaury.habrard@univ-st-etienne.fr

## 1 Introduction

L'inférence grammaticale probabiliste est un domaine de l'apprentissage automatique qui s'intéresse à l'apprentissage de modèles de langages probabilistes à partir d'un échantillon de chaînes ou d'arbres. Dans cet article, nous nous intéressons à l'apprentissage de distributions de probabilités sur des chaînes définies sur un alphabet fini de symboles $\Sigma$. Depuis quelques années, ce problème connaît un développement important grâce à l'arrivée de méthodes spectrales qui offrent l'avantage d'être à la fois efficaces et fondées théoriquement grâce à des garanties de consistance. Ces méthodes constituent maintenant un domaine très actif de recherche auquel plusieurs workshops ou tutoriels ont été dédiés dans les grandes conférences internationales (ICML'2012, NIPS'2012, ICML'2013). Les principales contributions récentes couvrent un spectre large d'approches pour des modèles de séquences [HKZ09, BDR09, SBG10, SBS+10, CC12], d'arbres [BHD10, PSX11], ou encore de transduction [BQC11], parfois combinés avec des approches d'optimisation convexe [BQC12, BM12].

Le succès des approches spectrales tient en particulier au fait qu'elles fournissent une solution très élégante à un problème de base de l'inférence grammaticale probabiliste : étant donné un échantillon $S$ i.i.d. de chaînes $w_1, \ldots, w_l \in \Sigma^*$ tirés selon une distribution inconnue $p$, trouver un modèle de $p$ dans la classe des séries rationnelles définies sur $\Sigma^*$. Les séries rationnelles sont des applications de $\Sigma^*$ dans $\mathbb{R}$ qui admettent une représentation sous la forme d'automates pondérés (dont les automates probabilistes ou les HMMs sont des cas particuliers) de la forme $\langle I, (M_x)_{x\in\Sigma}, T\rangle$ où $I$ est le vecteur de poids initiaux, $M_x$ la matrice de transition associée au symbole $x$ et $T$ un vecteur de poids termi-

naux, la dimension de ces éléments correspondant au nombre d'états de l'automate. Le problème de départ peut alors se formuler comme l'apprentissage d'un automate pondéré $A$ dont la série $r_A$ associée est une bonne approximation de $p$. Le nombre minimal d'états d'un automate calculant une série rationnelle donnée $r$ a une interprétation algébrique : c'est la dimension d'un sous-espace de l'espace vectoriel des séries rationnelles, contenant $r$ et stable par des opérateurs $T_x$ constants associés à chaque symbole de $\Sigma$.

La recherche de cet espace, à partir d'un échantillon fini, constitue le cœur des méthodes spectrales. Pour en donner une illustration dans le cas idéal, la méthode spectrale de base consiste à :

- former la matrice de Hankel $H_S$ associée à $S$, c'est-à-dire la matrice infinie dont les lignes et les colonnes sont indexées par les mots de $\Sigma^*$ et dont les éléments sont définis par $H_S[u, v] = p_S(uv)$, $p_S$ étant la distribution empirique définie par $S$ ;
- déterminer, via une décomposition en valeurs singulières (SVD) de $H_S$, une approximation $LDR^T$ de rang $d$ de $H_S$, pour une valeur de $d$ bien choisie (cette étape correspond à l'estimation du sous-espace cherché) ;
- et former l'automate pondéré $\langle I, (M_x)_{x\in\Sigma}, T\rangle$ où $I$ est la première ligne de $R$, $T = R^T p_S$ et $M_x = R^T T_x R$ où $T_x$ est une matrice constante qui ne dépend que de $x$ et $\Sigma^*$.

La méthode est légitimée par le résultat de consistance suivant : si la distribution cible $p$ est elle-même une série rationnelle de rang $d$ et si la matrice de Hankel $H$ est renseignée avec les valeurs exactes de $p(uv)$, alors $H$ est de rang $d$ et l'automate inféré calcule $p$.

En pratique, cet idéal se heurte à un sérieux problème : la taille de la matrice $H_S$ devient rapidement considérable et une décomposition en valeurs singulières de cette matrice devient prohibitive. Il faut donc se résoudre à considérer des sous-matrices $H_S^{U,V}$ de la matrice $H_S$, pour des sous-ensembles de lignes $U$ et de colonnes $V$ bien choisis. Mais dans ce cas, la méthode de base décrite ci-dessus n'est plus consistante et des variantes de cette méthode ont été mises au point de manière à conserver la consistance [HKZ09, Bai11, BQC12]. On peut montrer qu'on peut se limiter à des ensembles $U$ et $V$ de cardinal $d$, ce qui conduit en pratique à des matrices de petites taille. Mais seuls les mots de $S$ pouvant s'écrire sous la forme $uv$ où $u \in U$ et $v \in V$ seront retenus par l'algorithme d'apprentissage, entraînant une perte sévère d'information. Une manière de remédier à cela consiste à tenter d'inférer d'autres séries que la distribution cible $p$, par exemple : la série préfixe $u \mapsto p(u\Sigma^*)$ ou la série de facteurs $u \mapsto \sum_{v,w\in\Sigma^*} p(vuw)$ sont des séries rationnelles de même rang que $p$, qui permettent de mieux exploiter l'information contenue dans $S$, et à partir desquelles on peut facilement retrouver $p$. Mais quoiqu'il en soit, on se retrouve avec un compromis à faire entre le temps de calcul de l'algorithme et la quantité d'information laissée de côté. La définition de la matrice de Hankel empirique à traiter est donc une question critique pour l'applicabilité des méthodes spectrales et l'impact de ce choix a été peu étudié jusqu'à présent.

A notre connaissance, il n'existe pas de formule analytique reliant directement la précision de la série inférée $||p - r_A||$ et la distance $||H^{U,V} - H_S^{U,V}||$ entre les matrices de Hankel exacte et empirique. Cependant, nous pouvons borner la distance entre les $d$ vecteurs singuliers droits $R$ et $R_S$, de $H^{U,V}$ et de $H_S^{U,V}$, qui interviennent de manière essentielle dans la reconstruction de l'automate inféré $A$ :

$$||\sin(\text{angles}(R, R_S))||_2 \leq \frac{||H^{U,V} - H_S^{U,V}||_2}{\sigma_d}$$

où $\sigma_d$ est la plus petite valeur singulière non nulle de $H^{U,V}$. Nous pouvons en déduire que

$$||\sin(\text{angles}(R, R_S))||_F \leq \sqrt{d}\frac{||H^{U,V} - H_S^{U,V}||_2}{\sigma_d}.$$

Cette inégalité est difficile à analyser dans le cas général, et ne permet pas de dire quelle taille d'ensemble $U$ et $V$ on a intérêt à choisir car le numérateur et le dénominateur de la partie droite de l'inégalité croissent avec la longueur des chaînes dans $U$ et $V$.

Par exemple, considérons le cas très simple du langage rationnel de rang 1, défini sur un alphabet d'une lettre par $p(a^n) = (1 - \rho)\rho^n$. La matrice de Hankel de $p$ est définie par $H[i, j] = (1 - \rho)\rho^{i+j}$. Soit $H^{(N)}$ la sous-matrice de $H$ définie pour $0 \leq i, j \leq N$. Elle est symétrique et de rang 1. Sa plus petite valeur singulière $\sigma^{(N)}$ est égale à son unique valeur propre :

$$\sigma^{(N)} = \sqrt{\sum_{i=0}^{N} \rho^{2i}} = \sqrt{\frac{1 - \rho^{2N+2}}{1 - \rho^2}}$$

qui est une fonction croissante de $N$.

Une alternative consiste à prendre $U = V = \Sigma^*$. En effet, on peut montrer que l'espérance $\mathbb{E}(||H^{U,V} - H_S^{U,V}||_2)$ est bornée par 1.

Dans cet article, nous démontrons des inégalités de concentration pour $||H - H_S||_2$ qui suggèrent qu'en l'absence d'information sur des paramètres difficiles à estimer, comme la vitesse de convergence de la valeur singulière minimale de $H^{U,V}$ vers la valeur singulière

minimale de $H$, on peut avoir intérêt à travailler avec la matrice $H_S$ et utiliser des techniques, éventuellement randomisées, permettant de traiter des matrices volumineuses.

# 2 Préliminaires

## 2.1 Valeurs singulières, valeurs propres

Soit $M \in \mathbb{R}^{m\times n}$ une matrice. Les *valeurs singulières* de $M$ sont les racines carrées des valeurs propres de la matrice symétrique $M^T M$.

Si $M$ est symétrique, alors $M^T M = M^2$ et les valeurs singulières de $M$ sont les valeurs absolues de ses valeurs propres non nulles, et les vecteurs singuliers coïncident avec les vecteurs propres non nuls qui leur sont associés. En particulier,

$$\begin{aligned}\sigma_{max}(M) &= Sup(|\lambda_{max}(M)|, |\lambda_{min}(M)|)\\ &= Sup(\lambda_{max}(M), -\lambda_{min}(M)|)\end{aligned}$$

où $\lambda_{max}(M)$ (resp. $\lambda_{min}(M), \sigma_{max}(M)$) désigne la valeur propre maximale (resp. la valeur propre minimale, la valeur singulière maximale) de $M$.

Si $A, B, C$ sont des matrices symétriques, si $\mu_1 \geq \ldots \geq \mu_n$ (resp. $\nu_1 \geq \ldots \geq \nu_n$, resp. $\rho_1 \geq \ldots \geq \rho_n$) désignent les valeurs propres de $A$ (resp. $B$, resp. $C$) et si $A = B + C$, alors $\nu_i + \rho_n \leq \mu_i \leq \nu_i + \rho_1$ pour tout $1 \leq i \leq n$ (inégalités de Weyl).

## 2.2 Normes matricielles

On peut définir plusieurs normes sur les matrices $M \in \mathbb{R}^{m\times n}$ :

- les normes $||\cdot||_k$ induites par les normes correspondantes sur $\mathbb{R}^n$ et définies par $||M||_k = max_{x\neq 0}\frac{||Mx||_k}{||x||_k}$. On peut montrer que
  - $||M||_1 = Max_{1\leq j\leq n}\sum_{i=1}^{m}|M[i,j]|$,
  - $||M||_\infty = Max_{1\leq i\leq m}\sum_{j=1}^{n}|M[i,j]|$
  - $||M||_2 = \sigma_{max}(M)$ and $\sigma_{min}(M) = 1/||M^+||_2$ où $M^+$ désigne la pseudo-inverse de $M$.
- la norme de Frobenius $||\cdot||_F$ définie par

$$||M||_F = \sqrt{\sum_{i,j} M[i,j]^2} = \sqrt{\sum_{k=1}^{\min(m,n)} \sigma_k^2}$$

où les $\sigma_k$ sont les valeurs singulières de $M$.

On a les deux inégalités suivantes :

$$||M||_2 \leq ||M||_F \leq \sqrt{\text{rank}(M)}\cdot||M||_2 \quad (1)$$

$$||M||_2 \leq \sqrt{||M||_1||M||_\infty} \quad (2)$$

Ces normes peuvent être étendues, sous certaines conditions, aux matrices infinies et les inégalités précédentes restent vraies dès lors que les normes correspondantes sont définies.

## 2.3 Langages stochastiques rationnels

Cette section introduit la notion de langage stochastique rationnel qui admet une définition algébrique particulièrement adaptée aux approches spectrales que nous considérons ici.

Soit $\Sigma$ un alphabet fini et $\Sigma^*$ l'ensemble des mots de longueur finie définissables sur $\Sigma$. On note $|w|$ la longueur d'un mot $w$, $\Sigma^n \subset \Sigma^*$ l'ensemble des mots de longueur $n$ et $\epsilon \in \Sigma^*$ le mot vide. Etant donné un mot $w$, nous notons les ensembles des préfixes et facteurs de $w$ respectivement par $\text{Pref}(w) = \{u | w = uv \text{ avec } u, k \in \Sigma^*\}$ et $\text{Fact}(w) = \{f | w = ufv \text{ avec } u, v, f \in \Sigma^*\}$. Une *série* définie sur $\Sigma$ est une application $r : \Sigma^* \mapsto \mathbb{R}$. Une série $r$ est *convergente* si la suite de terme général $r(\Sigma^{\leq n})$, égal par définition à $\sum_{w\in\Sigma^{\leq n}} r(w)$, est convergente : on note alors cette limite par $r(\Sigma^*)$. Un *langage stochastique* $p$ est une distribution de probabilités définies sur $\Sigma^*$, c'est-à-dire une série ne prenant que des valeurs positives ou nulles et convergeant vers 1.

Soit $n \geq 1$ et $M$ un morphisme défini de $\Sigma^*$ vers $\mathcal{M}(n)$, l'ensemble des matrices carrées $n \times n$ à coefficients réels. Pour tout $u \in \Sigma^*$, on note $M(u)$ par $M_u$ et $\Sigma_{x\in\Sigma} M_x$ par $M_\Sigma$. Une série $r$ définie sur $\Sigma$ est *rationnelle* s'il existe un entier $n \geq 1$, $I, T \in \mathbb{R}^n$ et un morphisme $M : \Sigma^* \mapsto \mathcal{M}(n)$ tel que pour tout $u \in \Sigma^*$,

$$r(u) = I^T M_u T.$$

Le triplet $\langle I, M, T\rangle$ est une *représentation linéaire* de dimension $n$ de $r$. Le vecteur $I$ peut être interprété comme un vecteur de poids initiaux, $T$ comme un vecteur de poids terminaux et le morphisme $M$ comme un ensemble de paramètres matriciels associés aux lettres de l'alphabet. Un *langage stochastique rationnel* est donc un langage stochastique admettant une représentation linéaire.

Nous introduisons maintenant la notion de matrice de Hankel associée à une série $r$. Soient $U, V \subseteq \Sigma^*$. La *matrice de Hankel* $H_r^{U,V}$ est la matrice indexée par $U \times V$ et définie par $H_r^{U,V}[u, v] = r(uv)$. Si $U = V = \Sigma^*$, $H_r^{U,V}$ est simplement notée $H_r$. Sauf mention contraire, nous supposons toujours que $\epsilon \in U$ et que $U$ et $V$ sont ordonnés par longueur d'abord puis par ordre lexicographique.

On peut montrer qu'une série $r$ est rationnelle ssi le rang de la matrice $H_r$ est fini. Il est alors égal à la dimension minimale d'une représentation linéaire de $r$.

Soit $r$ une série rationnelle ne prenant que des valeurs positives ou nulles et convergente. On peut démontrer que la limite supérieure $\rho(r) = \overline{lim}_{n\to\infty} (r(\Sigma^n))^{1/n} < 1$. On peut montrer que si $\langle I, M, T\rangle$ est une représentation linéaire de dimension minimale $d$ de $r$, alors $\rho(r)$ est le rayon spectral de la matrice $M_\Sigma$. En particulier, la série $M_\Sigma^n$ est convergente et $r(\Sigma^*) = I^T(Id - M_\Sigma)^{-1}T$, où $Id$ désigne la matrice identité de dimension $d$.

D'autres séries rationnelles convergentes peuvent être naturellement associées à un langage stochastique rationnel $p$. En voici trois exemples que nous utiliserons par la suite.

- $\tilde{p}$ défini par $\tilde{p}(u) = p(\tilde{u})$ où $\tilde{u}$ est le miroir de $u$,
- $\overline{p}$ défini par $\overline{p}(u) = p(u\Sigma^*)$, la série associée aux *préfixes* du langage, et
- $\hat{p}$ défini par $\hat{p}(u) = \sum_{v,w\in\Sigma^*} p(vuw)$, la série associée aux *facteurs* du langage.

Si $\langle I, M, T\rangle$ est une représentation linéaire minimale de $p$, on a pour tout $u \in \Sigma^*$,

- $\tilde{p}(u) = I^T M_{\tilde{u}} T$ et $\tilde{p}(\Sigma^*) = 1$ ; $\tilde{p}$ est un langage stochastique rationnel
- $\overline{p}(u) = I^T M_u (Id - M_\Sigma)^{-1} T$ et $\overline{p}(\Sigma^*) = I^T (Id - M_\Sigma)^{-2} T$,
- $\hat{p}(u) = I^T (Id - M_\Sigma)^{-1} M_u (Id - M_\Sigma)^{-1} T$ et $\hat{p}(\Sigma^*) = I^T (Id - M_\Sigma)^{-3} T$.

La donnée d'une représentation linéaire de l'une de ces quatre variantes permet de reconstituer les autres.

## 2.4 Inférence

Nous présentons maintenant l'algorithme spectral pour l'apprentissage de langages stochastiques rationnels. Le principe de l'algorithme consiste à retrouver une représentation linéaire du langage cible à l'aide d'une décomposition en valeurs singulières de la matrice de Hankel. Pour définir cet algorithme, nous avons d'abord besoin de caractériser les vecteurs de poids initiaux et terminaux, ainsi que le morphisme $M$ en fonction de cette matrice de Hankel.

Définissons, pour tout $s \in \Sigma^*$, la matrice $T_s$, dont les lignes et les colonnes sont indexées par $\Sigma^*$ et définie par $T_s[u, v] = 1$ si $v = us$ et 0 sinon. On montre facilement que l'application $s \mapsto T_s$ est un morphisme. En effet, $T_{s_1} T_{s_2}[u, v] = \sum_{w\in\Sigma^*} T_{s_1}[u, w] T_{s_2}[w, v] = 1$ ssi $v = us_1s_2$ et 0 sinon.

Si $X$ est une matrice dont les lignes sont indexées par $\Sigma^*$. On a $T_s X[u, v] = \sum_w T_s[u, w] X[w, v] = X[us, v]$ : les lignes de $T_s X$ sont incluses dans l'ensemble des lignes de $X$.

Si $E$ est le vecteur dont les colonnes sont indexées par $\Sigma^*$ et dont toutes les coordonnées sont nulles sauf la première, égale à 1 : alors, $E^T T_s$ est égale à la première ligne de $T_s$, dont toutes les coordonnées sont nulles sauf celle indexée par $s$ qui vaut 1 (nous rappelons que la première ligne de $T_s$ est indexée par le mot vide).

Soit $r$ une série rationnelle de dimension $d$, et $U \subset \Sigma^*$ tel que la matrice $H_r^{U\times\Sigma^*}$ (notée $H$ dans ce qui suit) soit de rang $d$. Soit $H = LDR^T$ une décomposition en valeur singulière réduite. $R$ est une matrice de dimension $\infty \times d$ dont les colonnes forment un ensemble de vecteurs orthonormés, les vecteurs singuliers droits de $H$ : $R^T R = Id$ et $RR^T H^T = H^T$ ($RR^T$ est la projection orthogonale sur l'espace engendré par les lignes de $H$).

On en déduit aisément, par récurrence sur $n$ que pour tout mot $u = x_1 \dots x_n$,

$$(R^T T_{x_1} R) \circ \dots \circ (R^T T_{x_n} R) R^T H^T = R^T T_u H^T.$$

En effet, l'égalité est évidemment vraie pour $n = 0$ puisque $T_\epsilon = Id$. Et l'on a $R^T T_x R R^T T_u H^T = R^T T_x T_u H^T = R^T T_{xu} H^T$ puisque les colonnes de $T_u H^T$ sont des lignes de $H$ et que $T$ est un morphisme.

En particulier, si $P$ est la première ligne de $H$, on a

$$E^T R (R^T T_{x_1} R) \circ \dots \circ (R^T T_{x_n} R) R^T P^T = E^T T_u P^T = r(u).$$

Autrement dit, $\langle R^T E, (R^T T_x R)_{x\in\Sigma}, R^T P^T\rangle$ est une représentation linéaire de $r$ de dimension $d$. Il faut remarquer que $r$ n'intervient que dans les vecteurs singuliers droits $R$ et dans le vecteur $P$.

On en déduit immédiatement un algorithme d'apprentissage d'un langage stochastique rationnel $p$ : à partir d'un échantillon $S$ i.i.d. de $p$,

- choisir un ensemble $U$ et former la matrice de Hankel $H_S^{U\times\Sigma^*}$,
- choisir un rang $d$ et effectuer une SVD réduite et tronquée au rang $d$ de $H_S^{U\times\Sigma^*}$,
- former la représentation linéaire $\langle R_S^T E, (R_S^T T_x R_S)_{x\in\Sigma}, R^T P_S^T\rangle$ à partir des vecteurs singuliers droits $R_S$ et de la distribution empirique $P_S$.

Cet algorithme est clairement consistant si l'on ne borne pas le nombre de colonne de la matrice de Hankel (et si $U$ et $d$ le sont, relativement à la cible). En revanche, il ne l'est plus si on limite les colonnes : des variantes, légèrement plus complexes, permettent de conserver la consistance en limitant les colonnes à un sous ensemble $V$ de $\Sigma^*$ [HKZ09, Bai11, BQC12].

Même si la représentation linéaire inférée à partir de $S$ dépend très simplement des données, à notre connaissance, il n'existe pas de formules analytiques reliant directement la distance entre la distribution cible $p$ et la distribution inférée en fonction de la distance entre $p$ et la distribution empirique $p_S$. Le mieux dont on dispose est décrit par l'inégalité suivante, provenant de

la théorie des perturbations des matrices [Ste90] : soit $\theta$ l'angle principal (voir [MBI92] par exemple) entre les espaces engendrés par les vecteurs singuliers droits de $R$ et de $R_S$, et soit $R_S^{\perp}$ un ensemble de $d$ vecteurs orthonormés et orthogonaux à ceux de $R_S$,

$$|sin(\theta)| = ||R_S^{\perp} R||_2 \leq \frac{||H_S^{U\times V} - H_r^{U\times V}||_2}{\sigma_d}$$

où $\sigma_d$ est la valeur singulière minimale de $H_r^{U\times V}$. Une formule analogue, utilisant la norme de Frobenius à la place de la norme spectrale, peut être aussi obtenue.

Cette formule ne donne pas d'indication nette sur l'impact ou l'intérêt de borner ou non l'ensemble $V$. En effet, les inégalités de Weyl peuvent être utilisées pour montrer qu'à la fois le numérateur et le dénominateur de la partie droite de l'inégalité croissent avec $V$. Dans la plupart des travaux publiés, les auteurs travaillent sur des variantes des matrices de Hankel permettant d'utiliser au mieux les données disponibles, en se basant par exemple sur les séries préfixes $\overline{p}$ ou facteur $\hat{p}$, au prix d'une variance plus importante, mais permettant de conserver des tailles de matrices assez faibles. Nous souhaitons explorer la possibilité alternative de travailler avec la matrice de Hankel de base, qui conduit à une représentation linéaire inférée très simple, en bornant le moins possible le nombre de colonnes retenues.

# 3 Bornes de concentration pour les matrices de Hankel

Nous nous intéressons au comportement (en terme de consistance) de l'algorithme spectral appliqué à une matrice de Hankel de base non bornée. Avant d'étudier ce cas en détail, nous mentionnons d'abord un résultat connu pour le cas d'une matrice de Hankel bornée. Puis nous donnons ensuite quelques outils mathématiques qui nous seront nécessaires pour dériver notre résultat.

## 3.1 Cas d'une matrice de Hankel bornée

Soient $X_1,\ldots,X_N \in \mathbb{R}^{d_1\times d_2}$ des matrices aléatoires i.i.d. avec $||X_i||_2 \leq M$ presque sûrement. Soient $d = min\{d_1, d_2\}$ et $D{=}max\{d_1, d_2\}$, on peut montrer que

$$||\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X]||_2 \leq \frac{6M}{\sqrt{N}}\left(\sqrt{\log d} + \sqrt{\log\frac{1}{\delta}}\right)$$

avec une probabilité supérieur à $1-\delta$ (voir par exemple [Kak10] pour une démonstration).

Ce résultat peut être utilisé pour fournir des bornes de concentration de la matrice de Hankel $H_S^{U,V}$ d'un langage stochastique $p$ lorsque $U$ et $V$ sont bornés. Soit $\xi$ une variable aléatoire prenant ses valeurs dans $\Sigma^*$ et de loi $p$. On peut associer à $\xi$ les matrices aléatoires

- $X$, définie par $X[u,v] = 1_{\xi=uv}$.
- $X_p$, définie par $X_p[u,v] = 1_{uv\in \mathrm{Pref}(\xi)}$
- $X_f$, définie par $X_f[u,v] = \#\{(x,y)|xuvy = w\}$.

Cela conduit dans chaque cas à une borne de la forme

$$||H_S - H||_2 \leq \frac{6M}{\sqrt{N}}\left(\sqrt{\log d} + \sqrt{\log\frac{1}{\delta}}\right) \qquad (3)$$

où $M = 1$ dans le premier cas et $M = \sqrt{D}$ dans le cas préfixe, valeur qui est atteinte en prenant $U = \{\epsilon\}$ et un alphabet $\Sigma$ à une seule lettre. Dans le cas facteur, $M$ n'est pas bornée si aucune hypothèse n'est faite sur la longueur maximale des mots tirés. Si ceux-ci ont une longueur $\leq D$, on a $M \leq (D+1)\sqrt{D}$.

Ce résultat ne peut pas être immédiatement généralisé au cas où l'une des deux dimensions de la matrice n'est pas bornée. Nous utilisons des résultats récents ([Tro12, HKZ11]) pour obtenir cette généralisation.

## 3.2 Borne de Bernstein pour les matrices aléatoires

Nous donnons maintenant un résultat pour le cas de matrices aléatoires de dimension non bornée.

Soient $\xi_1,\ldots,\xi_n$ des variables aléatoires, et pour chaque $i = 1,\ldots,n$, soit $X_i = X_i(\xi_1,\ldots,\xi_i)$ une matrice **symétrique** fonction de $\xi_1,\ldots,\xi_i$. La notation $\mathbb{E}_i[\cdot]$ est un raccourci pour $\mathbb{E}[\cdot|\xi_1,\ldots,\xi_{i-1}]$.

**Théorème 1.** *(Matrix Bernstein Bound)[HKZ11]. S'il existe $b > 0, \sigma > 0$, et $k > 0$ tels que pour tout $i = 1,\ldots,n$,*

$$\mathbb{E}_i[X_i] = 0,\ \lambda_{max}(X_i) \leq b,$$

$$\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i(X_i^2)\right) \leq \sigma^2,$$

$$\mathbb{E}\left[tr\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i(X_i^2)\right)\right] \leq \sigma^2 k$$

*presque sûrement, alors pour tout $t > 0$,*

$$Pr\left[\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) > \sqrt{\frac{2\sigma^2 t}{n}} + \frac{bt}{3n}\right] \leq \frac{k\cdot t}{e^t - t - 1}.$$

Le théorème précédent est valide pour des matrices aléatoires symétriques, mais cette hypothèse n'est pas toujours vérifiée de manière directe comme dans le cas des matrices de Hankel. Il est cependant possible d'étendre ce résultat de manière générale à n'importe quelle matrice réelle grâce au principe de dilatation.

Soit $Z$ une matrice aléatoire. La *dilatation* ([Tro12]) de $Z$ est la matrice aléatoire symétrique $X$ définie par $X = \begin{bmatrix} 0 & Z \\ Z^T & 0 \end{bmatrix}$. On a alors

$$X^2 = \begin{bmatrix} ZZ^T & 0 \\ 0 & Z^TZ \end{bmatrix} \text{ et } \lambda_{max}(X) = ||X||_2 = ||Z||_2.$$

En effet, on peut facilement vérifier que si $(u, v)$ est un vecteur propre de $X$ associé à la valeur propre $\lambda$, alors $(-u, v)$ est aussi un vecteur propre de $X$ lié à la valeur propre $-\lambda$. Donc, $\lambda_{max}(X) = \sigma_{max}(X) = ||X||_2$.

Si les variables aléatoires $\xi_i$ sont i.i.d., et si chaque matrice $Z_i$ ne dépend que de $\xi_i$, les matrices $X_1, \dots, X_n$ sont aussi i.i.d. et l'on obtient le corollaire suivant :

**Corollaire 1.** *Soient $\xi_1, \dots, \xi_n$ des variables aléatoires i.i.d., soient $Z_1 = Z(\xi_1), \dots, Z_n = Z(\xi_n)$ des matrices i.i.d. et pour tout $i = 1, \dots, n$, soit $X_i$ la dilatation de $Z_i$. S'il existe $b > 0, \sigma > 0$, et $k > 0$ tels que pour tout $i = 1, \dots, n$,*

$$\mathbb{E}[X_1] = 0,\ ||X_1||_2 \leq b,$$
$$||\mathbb{E}(X_1^2)||_2 \leq \sigma^2,\ tr(\mathbb{E}(X_1^2)) \leq \sigma^2 k$$

*presque sûrement, alors pour tout $t > 0$,*

$$Pr\left[||\frac{1}{n}\sum_{i=1}^{n} X_i||_2 > \sqrt{\frac{2\sigma^2 t}{n}} + \frac{bt}{3n}\right] \leq k \cdot t(e^t - t - 1)^{-1}.$$

## 3.3 Borne pour les matrices de Hankel

Pour obtenir notre borne, l'idée est d'appliquer le corollaire 1 précédent en le spécialisant à la différence entre la matrice de Hankel empirique et son espérance, en utilisant le principe de dilatation. Nous allons donc chercher les bonnes valeurs de $b$, $\sigma^2$ et $k$ nous permettant d'appliquer le corollaire.

Soit $p$ un langage stochastique rationnel défini sur $\Sigma^*$, soit $U, V \subseteq \Sigma^*$ et soit $H^{U,V}$ la matrice de Hankel associée à $p$. Soit $\overline{S}_p = \sum_{u \in \Sigma^*} p(u\Sigma^*) = \overline{p}(\Sigma^*)$. Si $\langle I, M, T\rangle$ est une représentation linéaire minimale de la série $p$, on a $\overline{S}_p = I^t(I - M_\Sigma)^{-2}T$.

Soit $\xi$ une variable aléatoire prenant ses valeurs dans $\Sigma^*$ et distribuée selon $p$. Soit $Z = \phi(\xi) \in \mathbb{R}^{|U|\times|V|}$ la matrice aléatoire définie par $Z[u, v] = \mathbf{1}_{\xi=uv} - p(uv)$.

Soient $\xi_1, \dots, \xi_n$ des copies indépendantes de $\xi$, soit $Z_i = \phi(\xi_i)$ et soit $X_i$ la dilatation de $Z_i$ pour $i = 1, \dots, n$.

**Théorème 2.** *Soit $S$ un échantillon de $N$ mots tirés indépendamment selon $p$. Pour tout $t > 0$,*

$$Pr\left[||H_S^{U,V} - H^{U,V}||_2 > \sqrt{\frac{2\overline{S}_p t}{N}} + \frac{2t}{3N}\right] \leq 2t(e^t - t - 1)^{-1}.$$

*Démonstration.* On doit vérifier les quatre conditions du corollaire 1.

Pour tout $(u, v) \in U \times V$, $\mathbf{1}_{\xi=uv}$ suit une loi de Bernoulli de paramètre $p(uv)$. Donc, $\mathbb{E}(Z) = 0$ et $\mathbb{E}[X_1] = 0$.

Pour tout $u \in U$, on a $\sum_{v\in V} |Z[u, v]| \leq \sum_{v\in V}[p(uv) + \mathbf{1}_{\xi=uv}] \leq 2$. De même, pour tout $v \in V$, $\sum_{u\in U} |Z[u, v]| \leq 2$. On en déduit que

$$||Z||_1 \leq 2\ ,\ \ ||Z||_\infty \leq 2 \text{ et}$$
$$||X_1||_2 = ||Z||_2 \leq \sqrt{||Z||_\infty ||Z||_1} \leq 2.$$

Soit $w$ une réalisation de $\xi$, pour tout $(u, u') \in U^2$ :

$$ZZ^T[u, u'] = \sum_{v\in V} Z[u, v]Z^T[v, u'] = \sum_{v\in V} Z[u, v]Z[u', v],$$

$$Z[u, v]Z[u', v] = (1_{w=uv} - p(uv))(1_{w=u'v} - p(u'v))$$
$$= p(uv)p(u'v) + \begin{cases} 1 - 2p(uv) & \text{si } u = u' \& uv = w \\ -p(u'v) & \text{si } u \neq u' \& uv = w \\ -p(uv) & \text{si } u \neq u' \& u'v = w \\ 0 & \text{sinon.} \end{cases}$$

Considérons d'abord les éléments diagonaux :

$$ZZ^T[u, u] = \sum_{v\in V}[p(uv)^2 + 1_{uv=w}(1 - 2p(w))] \text{ et}$$

$$\begin{aligned}\mathbb{E}[ZZ^T[u, u]] &= \sum_{w\in\Sigma^*} p(w)ZZ^T[u, u] \\ &= \sum_{v\in V} p(uv)^2 + \sum_{v\in V} p(uv)(1 - 2p(uv)) \\ &= p(uV) - \sum_{v\in V} p(uv)^2 \leq p(uV) \leq p(u\Sigma^*).\end{aligned}$$

Supposons maintenant que $u \neq u'$ :

$$ZZ^T[u, u'] = \sum_{v\in V}[p(uv)p(u'v) - 1_{uv=w}p(u'v) - 1_{u'v=w}p(uv))] \text{ et}$$

$$\mathbb{E}[ZZ^T[u, u']] = \sum_{w\in\Sigma^*} p(w)ZZ^T[u, u'] = -\sum_{v\in V} p(uv)p(u'v),$$

$|\mathbb{E}[ZZ^T[u,u']]| = \sum_{v\in V} p(uv)p(u'v) \leq min(p(uV), p(u'V))$

et, pour tout $u \in U$,

$$\sum_{u'\in U} |\mathbb{E}[ZZ^T[u,u']]| \leq p(uV) + \sum_{u'\in U\setminus\{u\}} p(u'V)$$
$$= p(U\cdot V) \leq \overline{S}_p$$

De même, pour tout $u' \in U$,

$$\sum_{u\in U} |\mathbb{E}[ZZ^T[u,u']]| \leq p(U\cdot V) \leq \overline{S}_p.$$

On en déduit que $||\mathbb{E}[ZZ^T||_* \leq p(U\cdot V) \leq \overline{S}_p$ pour $||\cdot||_* = ||\cdot||_1, ||\cdot||_\infty$ or $||\cdot||_2$.

Soit $\tilde{p}$ le langage stochastique miroir de $p$ défini par $\tilde{p}(u) = p(\tilde{u})$, soit $\tilde{\xi}$ une variable aléatoire distribuée selon $\tilde{p}$, et soit la matrice aléatoire définie par $\tilde{Z}[\tilde{u},\tilde{v}] = \mathbf{1}_{\tilde{\xi}=\tilde{u}\tilde{v}} - \tilde{p}(\tilde{u}\tilde{v})$. On peut vérifier que $\tilde{Z}[\tilde{u},\tilde{v}] = Z^T[u,v]$ et que $\overline{S}_{\tilde{p}} = \overline{S}_p$.

On a donc $||\mathbb{E}[Z^TZ||_* \leq p(V\cdot U) \leq \overline{S}_p$ pour $||\cdot||_* = ||\cdot||_1, ||\cdot||_\infty$ ou $||\cdot||_2$. Finalement, $||\mathbb{E}(X_1^2)||_2 \leq \overline{S}_p$ et $tr(\mathbb{E}(X_1^2)) \leq 2\overline{S}_p$.

On peut donc prendre $b = 2$, $\sigma^2 = \overline{S}_p$ et $k = 2$ pour remplir les conditions du corollaire 1.

□

Le nombre de lignes et de colonnes n'interviennent pas dans les bornes du théorème 2 ce qui laisse penser qu'il est envisageable de ne pas limiter la taille de la matrice de Hankel utilisée pour construire l'hypothèse.

Le théorème 1 ne permet pas d'obtenir des bornes comparables pour les matrices de Hankel de $\overline{p}$ et $\hat{p}$. En effet, les matrices aléatoires correspondantes n'ont pas une norme spectrale bornée. Il est possible d'utiliser une technique de troncature pour obtenir des bornes mais les résultats ne sont pas homogènes et difficilement interprétables. Une meilleure solution consiste sans doute à utiliser des inégalités de Bernstein pour variables non bornées et sous-exponentielles ([Tro12]) : en effet, les variantes préfixes et facteurs des matrices de Hankel font intervenir de manière essentielle la concentration des longueurs des chaînes observées autour de leur longueur moyenne, qui est une variable non bornée et sous-exponentielle. Cette extension fait partie des perspectives de cette première étude.

# 4 Expériences

Dans cette section, nous commençons par présenter brièvement la base de référence PAutomaC que nous avons utilisée. Nous comparons ensuite les valeurs de concentration prédites par les formules habituelles (équation 3) et par celle que nous avons démontrée (théorème 2). Nous comparons ensuite les résultats obtenus par l'algorithme spectral de base pour les matrices de Hankel "standard", "préfixes" et "facteurs".

## 4.1 Les données PAutomaC

La compétition PAutomaC (Probabilistic Automata learning Competition [1]), proposée en marge de la conférence ICGI en 2012, s'intéresse au problème de l'apprentissage de distributions de probabilités sur des chaînes modélisées par des automates finis. Nous considérons la tâche basée sur des données artificielles qui se présente sous la forme d'un ensemble de 48 problèmes à résoudre. Chaque problème étant modélisé par un modèle cible auquel on associe un échantillon d'apprentissage tiré selon le modèle et un ensemble de test. L'évaluation consiste à approximer au mieux les probabilités des chaînes de l'ensemble de test, la qualité de l'approximation étant évaluée par le critère de perplexité défini comme suit :

$$2^{-\left(\sum_{w\in TestSet} p_T(w)*\log(p_C(w))\right)}$$

où $p_T$ est la distribution du modèle cible et $p_C$ la distribution d'un modèle candidat que l'on cherche à évaluer, cf [VEdlH12] pour plus d'informations.

Pour notre étude, nous avons retenu 11 problèmes pour lesquels le rapport entre les dimensions de la matrice de Hankel et sa parcimonie permettait aux algorithmes disponibles sous `NumPy` et `SciPy` de traiter les données sans optimisation. Les caractéristiques de chacun des problèmes sont résumées dans la table 1.

## 4.2 Instanciation des bornes

Le tableau 2 indique les bornes de $||H - H_S||_2$, calculées pour les 11 problèmes que nous avons retenus, à la confiance de 95%, prédites par le théorème 2 (a) et par l'équation 3 (b). L'équation 3 requière une dimension $d$ égale à la longueur de la plus longue chaîne retenue dans $U$. Nous avons choisi de retenir pour $U$ l'ensemble préfixe de 200 mots qui maximise $p_S(U)$. On peut constater que la borne du théorème 2, quoique ne dépendant pas de la dimension de la matrice, est très significativement plus précise que celle de l'équation 3 dans un cas d'utilisation réel. Des résultats analogues ont été obtenus pour tous les problèmes de PAutomaC.

La table 3 indique les bornes de $||\overline{H} - \overline{H}_S||_2$ calculées à la confiance de 95% à partir de l'équation 3

1. `http://ai.cs.umbc.edu/icgi2012/challenge/Pautomac/`

| **Numéro du problème** | **3** | **4** | **7** | **15** | **25** | **29** | **31** | **38** | **39** | **40** | **42** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Taille alphabet** | 4 | 4 | 13 | 14 | 10 | 6 | 5 | 10 | 14 | 14 | 9 |
| **Taille moyenne mots** | 7.219 | 5.259 | 5.523 | 12.461 | 9.723 | 5.287 | 6.001 | 7.177 | 7.736 | 8.716 | 6.350 |
| **Taille maximum mots** | 67 | 55 | 36 | 110 | 90 | 59 | 59 | 84 | 106 | 106 | 70 |
| **Taille échantillon $S$** | 20000 | 100000 | 20000 | 20000 | 20000 | 20000 | 20000 | 20000 | 20000 | 20000 | 20000 |
| **Type de modèle cible** | PFA | PFA | DPFA | PFA | HMM | PFA | PFA | HMM | PFA | DPFA | DPFA |
| **Nb d'états de la cible** | 25 | 12 | 12 | 26 | 40 | 36 | 12 | 14 | 6 | 65 | 6 |
| **Taille $H_S^{U,V}$ classique** | 1.9g | 0.5g | 0.17g | 27g | 13g | 0.4g | 1.4g | 8g | 7.7g | 15g | 3.4g |
| **Parcimonie** | .0053% | .0185% | .0212% | .0009% | .0015% | .0116% | .0061% | .0018% | .0019% | .0011% | .0033% |
| **Taille $H_S^{U,V}$ préfixes** | 2.5g | 1.8g | 0.7g | 291g | 99g | 2.4g | 7.6g | 60g | 75g | 165g | 25g |
| **Parcimonie** | .0058% | .0191% | .0208% | .0001% | .0016% | .0122% | .0066% | .0019% | .0020% | .0012% | .0035% |
| **Taille $H_S^{U,V}$ facteurs** | 73g | 6.4g | 3g | 3363g | 797g | 15.7g | 44g | 460g | 761g | 1925g | 202g |
| **Parcimonie** | .0058% | .0197% | .0199% | .0001% | .0016% | .0115% | .0069% | .0020% | .0020% | .0012% | .0036% |

TABLE 1 – Caractéristiques des 10 problèmes PAutomaC utilisés dans l'étude. Les différents types de modèles cibles correspondent aux automates finis probabilistes non déterministes (PFA), déterministes (DPFA) et aux modèles de Markov cachés (HMM). La taille des matrices $H_S^{U,V}$ correspond à une estimation exprimée en milliards (g) et la parcimonie associée correspond au pourcentage d'entrées non nulles dans la matrice.

| Number | 3 | 4 | 7 | 15 | 25 | 29 | 31 | 38 | 39 | 40 | 42 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\overline{S}_p$ | 8.235 | 6.253 | 6.519 | 13.398 | 10.654 | 6.347 | 6.967 | 8.091 | 8.815 | 9.744 | 7.391 |
| $\|\|H - H_S\|\|_2$ (a) | 0.069 | 0.027 | 0.061 | 0.087 | 0.078 | 0.061 | 0.063 | 0.068 | 0.071 | 0.074 | 0.065 |
| $\|\|H - H_S\|\|_2$ (b) | 0.171 | 0.077 | 0.171 | 0.171 | 0.171 | 0.171 | 0.171 | 0.171 | 0.171 | 0.171 | 0.171 |

TABLE 2 – Bornes de $||H - H_S||_2$ à la confiance de 95% prédites par le théorème 2 (a) et par l'équation 3 (b).

où $U$ est un ensemble préfixe de 200 mots ($d = 200$) et où le nombre de colonnes $D$ est fixé à 200 (cas c) ou déterminé par $U$ et $S$ (cas d). Sans surprise, ces bornes sont non seulement beaucoup plus mauvaises que celles qui ont été calculées pour $||H - H_S||_2$ mais elles dépendent aussi fortement du nombre de colonnes sélectionnées. Cela pose plusieurs questions. Existe t-il une borne pour la norme spectrale de ces matrices qui ne dépendent pas de leur nombre de colonnes et de lignes ? Autrement dit, existe t-il un analogue du théorème 2 pour ces matrices ? Et sinon, selon quelle heuristique choisir les colonnes de cette matrice ? Il semble plus robuste d'utiliser la matrice $H$ classique sans borner le nombre de colonnes.

### 4.3 Résultats : perplexité

Nous avons testé la méthode spectrale de base, telle qu'elle est décrite dans la section 2.4, en construisant des matrices de Hankel, standard, préfixes ou facteurs, en limitant les lignes à des ensembles préfixes d'au plus 200 mots et sans limitation du nombre de colonnes. La figure 1 compare les meilleurs résultats que nous avons obtenus pour chaque matrice, estimés par un calcul de perplexité sur l'échantillon test fourni dans le benchmark, avec les résultats de la cible et d'un algorithme d'apprentissage par trigram, qui constitue une méthode plancher de référence dans le domaine. On peut constater que l'algorithme utilisant la matrice de Hankel standard est pratiquement toujours la meilleure des trois. Sans aucune optimisation, ces méthodes basiques auraient été classées parmi les 4 meilleurs résultats du challenge Pautomac dans 5 jeu de données sur 11.

Outre le fait que la méthode basée sur la matrice de Hankel standard l'emporte souvent sur les deux autres, elle est aussi beaucoup plus robuste en ce sens qu'augmenter le nombre de lignes de la matrice de Hankel ne constitue jamais un handicap et que le choix du rang $d$ n'influe pas brutalement sur les résultats. La figure 2 illustre ces comportements sur l'un des problèmes que nous avons étudié. Bien que la méthode préfixe produise une solution de bonne qualité, cette solution est confinée dans une vallée qu'il pourra être difficile de découvrir. Ce phénomène est assez généralement observé. Il est peut-être dû à une instabilité causée par une mauvaise approximation de la matrice de Hankel.

## 5 Conclusion

Le cœur des méthodes spectrales en inférence grammaticale probabiliste consiste en une décomposition en valeurs singulières d'une matrice de Hankel résumant les données d'apprentissage, soit sous une forme proche

| Numéro | 3 | 4 | 7 | 15 | 25 | 29 | 31 | 38 | 39 | 40 | 42 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $M$ | 3.16 | 4.36 | 3 | 2.65 | 2.24 | 3.61 | 2.83 | 2.24 | 2.83 | 2.24 | 3 |
| $\|\|\overline{H} - \overline{H}_S\|\|_2$ (c) | 0.43 | 0.28 | 0.4 | 0.34 | 0.28 | 0.51 | 0.38 | 0.28 | 0.38 | 0.28 | 0.4 |
| $M$ | 8.25 | 7.48 | 6.08 | 10.54 | 9.54 | 7.75 | 7.75 | 9.22 | 10.34 | 10.34 | 8.43 |
| $\|\|\overline{H} - \overline{H}_S\|\|_2$ (d) | 1.12 | 0.49 | 0.82 | 1.37 | 1.18 | 1.09 | 1.03 | 1.14 | 1.37 | 1.28 | 1.13 |

TABLE 3 – Bornes de $||\overline{H} - \overline{H}_S||_2$ pour la version préfixes à la confiance de 95% prédites par l'équation 3 avec limitation du nombre colonnes à 200 (c), ou sans limitation (d).

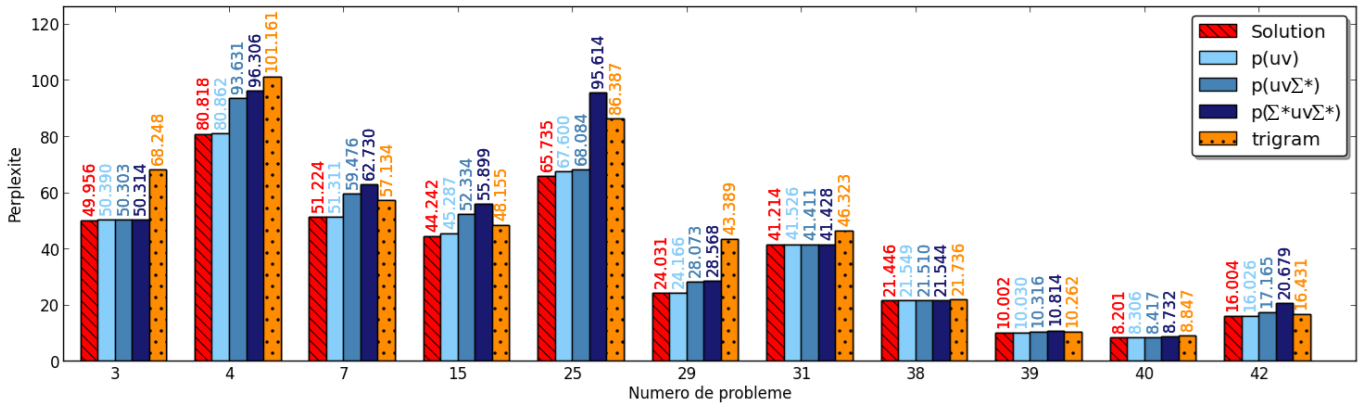


FIGURE 1 – Comparaison des méthodes spectrales de base sur les 11 problèmes sélectionnés.

de la distribution empirique - et dans ce cas, les plus grands exemples de l'échantillon ne peuvent être pris en compte que si la taille de la matrice le permet -, soit sous des formes plus sophistiquées permettant de prendre en compte tous les exemples, même dans des matrices de taille limitée.

La distance entre la matrice réelle, construite à partir de la distribution cible, et la matrice construite à partir de l'échantillon est critique : d'elle dépend la qualité du modèle inféré [HKZ09, Bai11]. Or il se trouve que les bornes de concentration disponibles ne sont pas très précises. Elles semblent indiquer que la précision du modèle inféré dépend négativement de la taille des matrices construites. Ce n'est pas le cas. Nous avons démontré une borne de concentration sur les matrices de Hankel standard qui ne dépend pas de la taille de ces matrices et qui se révèle par ailleurs plus serrée que les bornes usuelles sur le benchmark PAutomaC - qui est maintenant une référence du domaine. Les expériences menées sur un ensemble de problèmes extraits de ce benchmark confirment que les résultats s'affinent - ou ne subissent aucune dégradation - lorsque la matrice de Hankel servant à construire le modèle croît. Il n'est pas clair que cela reste vrai lorsqu'on utilise des formes plus complexes de matrices de Hankel, par exemple basées sur les préfixes ou les facteurs des observations. Les bornes invoquées usuellement se détériorent rapidement avec les tailles des matrices, nous ne disposons pas de bornes indépendantes de ces tailles et nous ne savons même pas s'il en existe. C'est une des perspective de la présente étude : tenter de trouver des bornes de concentration pour les formes "préfixes" et "facteurs" des matrices de Hankel, qui ne dépendent pas de leur dimension. Par ailleurs, les expériences que nous avons menées laissent penser que les algorithmes d'apprentissage utilisant ces variantes des matrices de Hankel sont moins robustes que ceux qui utilisent la forme la plus simple. Cela nous incite à développer des algorithmes d'apprentissage capable de traiter des très grandes matrices de Hankel simples au moyen de techniques de SVD randomisées sur des matrices très creuses.

## Références


[Bai11] R. Bailly. *Méthodes spectrales pour l'inférence grammaticale probabiliste de langages stochastiques rationnels*. PhD thesis, Aix-Marseille Université, 2011.

[BDR09] R. Bailly, F. Denis, and L. Ralaivola. Grammatical inference as a principal component analysis problem. In *Proceedings of ICML*, page 5, 2009.

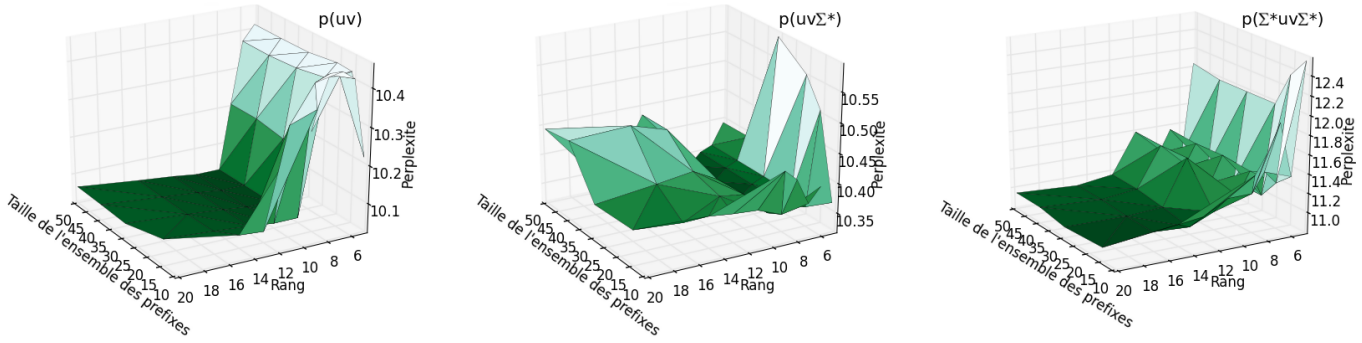


FIGURE 2 – Détail des résultats des 3 méthodes spectrales de base sur le problème 39.


[BHD10] R. Bailly, A. Habrard, and F. Denis. A spectral approach for probabilistic grammatical inference on trees. In *Proceedings of ALT*, pages 74–88, 2010.

[BM12] B. Balle and M. Mohri. Spectral learning of general weighted automata via constrained matrix completion. In *Proceedings of NIPS*, pages 2168–2176, 2012.

[BQC11] B. Balle, A. Quattoni, and X. Carreras. A spectral learning algorithm for finite state transducers. In *Proceedings of ECML/PKDD (1)*, pages 156–171, 2011.

[BQC12] B. Balle, A. Quattoni, and X. Carreras. Local loss optimization in operator models : A new insight into spectral learning. In *Proceedings of ICML*, 2012.

[CC12] S.B. Cohen and M. Collins. Tensor decomposition for fast parsing with latent-variable pcfgs. In *Proceedings of NIPS*, pages 2528–2536, 2012.

[HKZ09] D. Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proceedings of COLT*, 2009.

[HKZ11] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices. *ArXiv e-prints*, 2011.

[Kak10] S. Kakade. Multivariate analysis, dimensionality reduction, and spectral methods. Lecture Notes (Matrix Concentration Derivations), 2010.

[MBI92] J. Miao and A. Ben-Israel. On principal angles between subspaces in $\mathbb{R}^n$. *Linear Algebra Appl.*, 171 :81–98, 1992.

[PSX11] A.P. Parikh, L. Song, and E.P. Xing. A spectral algorithm for latent tree graphical models. In *Proceedings of ICML*, pages 1065–1072, 2011.

[SBG10] S. Siddiqi, B. Boots, and G.J. Gordon. Reduced-rank hidden Markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-2010)*, 2010.

[SBS+10] L. Song, B. Boots, S.M. Siddiqi, G.J. Gordon, and A.J. Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of ICML*, pages 991–998, 2010.

[Ste90] G. W. Stewart. Perturbation theory for the singular value decomposition. In *SVD and Signal Processing II : Algorithms, Analysis and Applications*, pages 99–109. Elsevier, 1990.

[Tro12] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4) :389–434, 2012.

[VEdlH12] S. Verwer, R. Eyraud, and C. de la Higuera. Results of the pautomac probabilistic automaton learning competition. *Journal of Machine Learning Research - Proceedings Track*, 21 :243–248, 2012.