Quadratic Weighted Automata: Spectral Algorithm and Likelihood Maximization [*]

Raphael BAILLY RAPHAEL.BAILLY@LIF.UNIV-MRS.FR
Aix-Marseille Université,UMR CNRS 6166, LIF, QARMA

Editor: Chun-Nan Hsu and Wee Sun Lee

Abstract

In this paper, we address the problem of non-parametric density estimation on a set of strings Σ^*. We introduce a probabilistic model – called quadratic weighted automaton, or QWA – and we present some methods which can be used in a density estimation task. A spectral analysis method leads to an effective regularization and a consistent estimate of the parameters. We provide a set of theoretical results on the convergence of this method. Experiments show that the combination of this method with likelihood maximization may be an interesting alternative to the well-known Baum-Welch algorithm.

Keywords: grammatical inference, non-parametric density estimation.

1. Introduction

The framework of this paper is non-parametric density estimation of an unknown distribution over a set of strings built from a finite alphabet Σ. This problem is usually solved by minimizing the Kullback-Leibler divergence between a parameterized distribution p_θ and the target p. The most used models for this task are Hidden Markov Models (HMMs) or equivalently Probabilistic Automata (PAs). From a training sample S, this problem classically boils down to optimizing a functional depending on the parameters $\boldsymbol{\theta}$ of the model

$$\boldsymbol{\theta}^* = \arg\min_{\theta} kl(p_S || p_{\boldsymbol{\theta}})$$

where p_S is the empirical distribution built from S, $p_{\boldsymbol{\theta}}$ is the distribution parameterized by $\boldsymbol{\theta}$, and $kl(p_S||p_{\boldsymbol{\theta}})$ is the Kullback-Leibler divergence of $p_{\boldsymbol{\theta}}$ with respect to p_S. Solving this problem is known to be NP-hard for a given HMM structure.

The models being used in this task (HMMs, weighted automata...) generally encompass the class of *Probabilistic Deterministic Finite Automata (PDFAs)*. As any finite-support distribution p can be modeled by a PDFA, in particular, every empirical distribution p_S built from a finite sample S, the minimum of the functional is reached, and one has $p_{\boldsymbol{\theta}^*} = p_S$. Thus, one needs a regularization before the minimization step. Classically, this regularization is achieved by giving a bound on the number of states – or the number of parameters – of the model.

We introduce in this paper a class of probabilistic models, the *quadratic weighted automata* (or QWAs), a subclass of weighted automata (WAs), which are a generalization of probabilistic

[*] This work was partially supported by the ANR LAMPADA ANR-09-EMER-007 project and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

automata. Consistent methods have been proposed for the estimation of WAs parameters, such as DEES (Denis et al. (2006)) or Spectral methods (Bailly and Denis (2011),Hsu et al. (2009b)). However, a major drawback of using WAs as probabilistic models is that the nonnegativity of a series computed by a WA is undecidable. In other words, the series computed by a WA produced by these algorithms may return negative values for some strings (a problem known as NPP - negative probability problem). As there exists no syntactical property to bound the absolute sum of such models, this prevents the existence of methods of likelihood maximization, thus the use of WAs in a density estimation task.

Attempts have been made to solve this problem, for instance with NOOMs (norm observable operator models Zhao and Jaeger (2010)). It is possible to ensure that such models compute a distribution, and one can approach a local maximum of the likelihood of a sample using a gradient ascent algorithm, but this method suffers from the same drawbacks as the Baum-Welch algorithm: one can only reach a local maximum of the likelihood.

We show in this paper that QWAs can be used to obtain a probabilistic model computing an actual distribution – avoiding the NPP – for which it is possible to perform likelihood maximization, and for which the (consistent) spectral method of regularization and parameter estimation known for WAs apply. An additional benefit is given by the possibility of combining these methods: the estimated parameters can be used as a starting point for the likelihood maximization process (instead of randomly chosen parameters). This advantage is illustrated by the experiments described at the end of this paper.

We introduce in section 2 some preliminary definitions and properties used in the rest of the paper. In section 3, we address the expressiveness of QWAs, that is the relations between the different classes of distributions modeled by QWAs and other models. We present the spectral algorithm in section 4. In section 5, we address some statistical properties and inequalities deduced from the theoretical variance of p_S. We provide in section 6 some results about consistency. The ℓ_1 convergence of the series is treated in section 7. In section 8, we shortly describe how to perform a gradient ascent on the log-likelihood of the training sample, with an $O(|S|)$ computational cost for each iteration. Some experimental results are exposed in section 9. We conclude in Section 10, and discuss about some outlooks of this work.

2. Preliminaries

2.1. Tools

Let Σ^* be the set of strings on the finite alphabet Σ. The empty string is denoted by ε, and the length of a string u is denoted by $|u|$. For any integer k, we denote by Σ^k the set $\{u \in \Sigma^* \mid |u| = k\}$ and by $\Sigma^{\leq k}$ the set $\{u \in \Sigma^* \mid |u| \leq k\}$.

Given an alphabet Σ, one considers the set $\mathbb{R}^{\Sigma^*}$ of all the mappings from Σ^* into $\mathbb{R}$ (called series). This set is an $\mathbb{R}$-vector space. For any series r and any string $u \in \Sigma^*$, we denote by $\dot{u}r$ the series defined by $\dot{u}r(w) = r(uw)$. The *residual space* of r is the vector space spanned by $\{\dot{u}r\}_{u\in\Sigma^*}$. Its elements are called *residuals* of r.

A *weighted automaton (WA)* A with d states is defined by an initial vector $A.\boldsymbol{I} \in \mathbb{R}^d$, a terminal vector $A.\boldsymbol{T} \in \mathbb{R}^d$, and a set of $d \times d$ real-valued matrices $A.M_x$, one for each symbol $x \in \Sigma$. The value of the series r_A for a string $w = w_1 \dots w_n$ is defined by

$$r_A(w_1 \dots w_n) = A.\boldsymbol{I}^T \cdot A.M_{w_1} \dots A.M_{w_n} \cdot A.\boldsymbol{T}$$

For any WA A, one denotes by $A.M$ the matrix $A.M = \sum_{x \in \Sigma} A.M_x$.

A *probabilistic automaton (PA)* A is a WA where parameters are non-negative, $\sum_{1 \le i \le d} A.\boldsymbol{I} = 1$, and $\forall 1 \le i \le d, \sum_{1 \le j \le d} A.M_{ij} + A.\boldsymbol{T}_i = 1$. HMMs and PAs model the same distributions.

A *probabilistic deterministic automaton (PDA, or PDFA)* A is a PA verifying that there exists at most one i s.t. $A.\boldsymbol{I} \neq 0$, and $\forall x \in \Sigma, \forall 1 \le i \le d$ there exists at most one j such that $(A.M_x)_{ij} \neq 0$.

A series r is *rational* if it satisfies one of the two following equivalent conditions:

- the dimension of the residual space of r is finite
- r can be computed by a weighted automaton

The *rank* of a rational series r (resp. a WA A) is the dimension of its residual space (resp. the residual space of series r_A computed by A). If r has rank d, there exists a d-state WA computing r. If a WA A has d states, its rank is $\le d$. A is said to be *minimal* if its rank equals its number of states. One denotes by $|A|$ the WA obtained by taking the absolute values of all parameters of A.

One defines $r_A(\Sigma^n) = \sum_{|w|=n} r_A(w)$. One has $r_A(\Sigma^n) = A.\boldsymbol{I}^T \cdot A.M^n \cdot A.\boldsymbol{T}$. If the sum $\sum_{n \in \mathbb{N}} r_A(\Sigma^n)$ is convergent, its limit is denoted by $r_A(\Sigma^*)$. It can be efficiently computed in polynomial time using linear algebra properties.

Let r be a rational series. The *spectral radius* of r is denoted by $\rho(r)$. It is defined by $\rho(r) = \inf_\rho(\exists C s.t. \forall k, r(\Sigma^{>k}) < C\rho^k)$. The *spectral radius* of a WA A, denoted by $\rho(A)$, is defines by $\rho(A) = \rho(r_A)$. One has $r(\Sigma^*) < \infty \Leftrightarrow \rho(r) < 1$. The spectral radius of a rational series is computable in polynomial time.

2.2. Sum and Product

Given two rational series r_A (d states) and r_B (d' states), the sum $r_A + r_B$ and the product $r_A.r_B$ are also rational series. The sum is computed by the WA $A + B$ with $d + d'$ states defined by:

$$(A+B).\boldsymbol{I} = \begin{pmatrix} A.\boldsymbol{I} \\ B.\boldsymbol{I} \end{pmatrix}, (A+B).\boldsymbol{T} = \begin{pmatrix} A.\boldsymbol{T} \\ B.\boldsymbol{T} \end{pmatrix}, (A+B).M_x = \begin{pmatrix} A.M_x & 0 \\ 0 & B.M_x \end{pmatrix}$$

The product is computed by the WA $A \otimes B$ with dd' states defined by:

$$(A \otimes B).\boldsymbol{I} = A.\boldsymbol{I} \otimes B.\boldsymbol{I}, (A \otimes B).\boldsymbol{T} = A.\boldsymbol{T} \otimes B.\boldsymbol{T}, (A \otimes B).M_x = A.M_x \otimes B.M_x$$

Where $\otimes$ denotes the Kronecker product between two matrices.

One considers the Hilbert space $\mathbb{R}^{\Sigma^*}$ composed of the rational series r such that $\sum_{w \in \Sigma^*} r(w)^2 < \infty$ equipped with the inner product $< \cdot, \cdot >$ defined by $< r, s >= \sum_{w \in \Sigma^*} r(w)s(w)$.

Definition 1 *A* Quadratic Weighted Automaton *is a WA of the form $A \otimes A$, where A is a WA: it computes the series r_A^2.*

Let $\boldsymbol{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, one denotes $diag(\boldsymbol{v})$ the $n \times n$ matrix defined by $diag(\boldsymbol{v})_{ii} = \boldsymbol{v}_i$, $diag(\boldsymbol{v})_{ij} = 0$ for $i \neq j$.

Let us recall some properties of matrix norms. First, an *induced* norm $\|\|_p$ for matrices is a norm related to the corresponding vector norm: $\|M\|_p = \max_{\|\boldsymbol{v}\|_p=1} \|M\boldsymbol{v}\|_p$. Any induced norm is consistent (i.e. sub-multiplicative). One has the following properties: $\|M\|_\infty = \|M^T\|_1 = \max_i \sum_j |M_{ij}|$, $\|M\|_2 = (\rho(M^T M))^{1/2}$ where $\rho(M^T M)$ is the spectral radius of $M^T M$. The *Frobenius norm*, denoted $\|\|_F$, is given by $\|M\|_F = (\sum_{i,j} M_{ij}^2)^{1/2}$. Given an $n \times m$ matrix M of rank r, one has the inequalities:

$$\|M\|_2 \leq \|M\|_F \leq \sqrt{r}\|M\|_2, \frac{1}{\sqrt{m}}\|M\|_1 \leq \|M\|_2 \leq \sqrt{n}\|M\|_1$$

One has, for any matrix norm $\|\|$, the property: $\rho(M) = \lim_{n\to\infty} \|M^n\|^{1/n}$ (called Gelfand's formula). In particular, for any consistent norm $\|\|$, one has $\rho(M) \leq \|M\|$.

2.3. Linear representation of a rational series

Let r be a rational series of rank d. Let E be the residual space of r, and $B = \{w_1, \dots w_d\}$ a basis of E in the space $\mathbb{R}^{\Sigma^*}$. The vector space E is stable by the operator $\dot{x}$ for $x \in \Sigma^*$: one has $\dot{x}(\dot{u}r) = \dot{\overline{ux}}r$. The WA defined by

- $I =$ coordinates of r in the basis B
- $M_x^T =$ matrix of $\dot{x}$ in the basis B
- $T = (w_1(\epsilon), \dots, w_d(\epsilon))$

computes the series r. This WA is called the *linear representation* of r in the basis B.

Let $V = \{v_1, \dots\}$ be a set of strings (suffixes). Let W be the matrix defined by $W_{ij} = w_j(v_i)$. One supposes that V is such that W has rank d. Let $t \in E$, and let t_B its representation in the basis B. The vector $t_V = Wt_B$ is the vector $(t(v_1), \dots)$ corresponding to the valuation of t for the set V. Let W^+ be a pseudo-inverse of W, i.e. satisfying $W^+W = I_d$ where I_d is the identity matrix of rank d. Then W^+t_V represents t in the basis B. In particular, $I = W^+r_V$.

Let W_x be the matrix defined by $W_{xij} = w_j(xv_i)$. By linearity of the operator $\dot{x}$, the vector W_xt_B is the vector $(\dot{x}t(v_1), \dots)$. Thus, one has $M_x^T = W^+W_x$.

Let U be a set of strings (prefixes) such that $\{\dot{u}r | u \in U\}$ spans E. Let r_U be the vector $(r(u_1), \dots)$. For each $1 \leq i \leq d$, let $w_i^\circ \in \mathbb{R}^U$ be such that $w_i = \sum_u w_i^\circ(u)\dot{u}r$. One then has $w_i(\epsilon) = w_i^{\circ T} r_U$. Let W° be the matrix defined by $W_{ij}^\circ = w_j^\circ(u_i)$. One then has $T = W^{\circ T} r_U$. Let X be the matrix defined by $X_{ij} = r(u_iv_j)$. X is called Hankel matrix of r for the sets (U, V). Let X_x be the matrix defined by $X_{ij} = r(u_ixv_j)$. One then has $W = X^TW^\circ$. One has $\dot{x}w_i = \sum_{u\in U}(w_i^\circ)_u\dot{ux}r$ and thus $W_x = X_x^TW^\circ$. To sum up:

$$I = W^+r_V, M_x^T = W^+W_x = W^+X_x^TW^\circ, T = W^{\circ T}r_U$$

is a linear representation of r in the basis $B = \{w_1, \dots w_d\}$.

2.4. Singular value decomposition

Let X be an $m \times n$ matrix, let U and V be two orthonormal matrices, and let Γ be an $m \times n$ diagonal matrix such that $X = U\Gamma V^T$. The columns of U (resp V) are called *left singular vectors* (resp. *right singular vectors*). The diagonal entries of Γ are called *singular values* of X. One has the property:

- left singular vectors are eigenvectors of XX^T
- right singular vectors are eigenvectors of X^TX
- singular values are square roots of eigenvalues of X^TX or XX^T

Let d be the rank of X. Then, there exists only d non-zero singular values. Let D be the $d \times d$ diagonal matrix with non-zeros eigenvalues of X^TX ordered by decreasing magnitude. Let W be the matrix built from the d columns of V corresponding to non-zero eigenvalues. Let W^* be the matrix build from the d columns of U corresponding to non-zero eigenvalues. One then has $X = W^* D^{1/2} W^T$. As W and W^* are orthonormal, one can write:

$$W = X^T W^* D^{-1/2}$$

2.5. Spectral Algorithm

Let us recall here some properties of the spectral algorithm. These methods are introduced in Hsu et al. (2009a) and Bailly et al. (2009).

Proposition 2 *Let r be a rational series of rank d. Let $U = \{u_1, \dots\}$ be a set of prefixes, $V\{v_1, \dots\}$ a set of suffixes, and let X be the matrix defined by $X_{ij} = r(u_i v_j)$. Let X_x be the matrix defined by $(X_x)_{ij} = r(u_i x v_j)$. One supposes that the rank of X is maximal, that is $rank(X) = d$. Let $\boldsymbol{r}_u = (r(u_1), \dots)$ and $\boldsymbol{r}_v = (r(v_1), \dots)$. Let $D^{1/2}$ be the $d \times d$ diagonal matrix composed of the singular values of X ordered by decreasing magnitude. Let W be the matrix of the d first right singular vectors, and W^* be the matrix of the d first left singular vectors. Then the WA defined by:*

$$I = W^T \boldsymbol{r}_v, M_x^T = W^T X_x^T W^* D^{-1/2}, T = \left(W^* D^{-1/2}\right)^T \boldsymbol{r}_u$$

computes r.

Proof Straightforward from the former remarks: one uses $W^\circ = W^* D^{-1/2}$ and $W^+ = W^T$. ■

Given a unknown distribution p computed by a WA, and a sample S i.i.d. with respect to p, one considers the empirical distribution p_S. The first part of the spectral algorithm consists in computing the number of significant non-zero singular values, providing a dimension d. One then builds the matrices X_x, $X_{S,x}$, W_S, W_S^*, D_S and the vectors $\boldsymbol{p_S}_v$ and $\boldsymbol{p_S}_v$, and computes an estimate of the target defined by:

$$I = W_S^T \boldsymbol{p_S}_v, M_x^T = W_S^T X_{S,x}^T W_S^* D_S^{-1/2}, T = \left(W_S^* D_S^{-1/2}\right)^T \boldsymbol{p_S}_u$$

3. QWA Expressiveness

Proposition 3 *Let p_{PDFA} (resp. p_{QWA}) denote the distribution modelled by a PDFA (resp. a QWA). Then $p_{PDFA} \subsetneq p_{QWA}$.*

Proof One first shows that $p_{PDFA} \subset p_{QWA}$. Let A be a PDFA computing a positive series: for any string $w \in \Sigma^*$, there exists at most one non-zero path, thus $|A|$ computes the same series as A. The automaton B obtained by taking the square root of all parameter of $|A|$ satisfies $r_B^2 = r_A$.

There exists a convergent rational series r which is not computable by a PDFA. Then r^2 is not computable by a PDFA: by the former argument, r would be computable by a PDFA. ■

One also has the properties $p_{QWA} \nsubseteq p_{HMM}$ and $p_{HMM} \nsubseteq p_{QWA}$, but this will not be addressed in this paper.

4. Algorithm Principle

Let p be a probability distribution, S be a sample i.i.d. with respect to p, p_S the empirical distribution. Let $q = p^{1/2}$ and $q_S = p_S^{1/2}$. The main idea is to apply the spectral algorithm to q_S.

One starts from a finite submatrix X (built from a set of prefixes U and a set of suffixes V) of the Hankel matrix of q, defined by $X_{ij} = q(u_i v_j) = \dot{u}_i q(v_j)$. One supposes that U and V are such that X has the same rank as q. Thus, an i.i.d. sample S with respect to q^2 provides an estimate of X, denoted $X_S = (p_S(u_i v_j))^{1/2}$.

The first goal is to estimate the rank d of X: one performs a SVD on the matrix X_S, and a statistical test on the obtained singular values to select a correct dimension d.

```
Algorithm 1: Quadratic Spectral Algorithm
Data: A sample $S = \{s_i, 1 \leq i \leq |S|\}$ i.i.d. according to a distribution $q^2$, a dimension $d$, an
      alphabet $\Sigma$, a set of prefixes $U$, a set of suffixes $V$.
Result: A Weighted Automaton $A$ computing $q$
begin
    $X_{i,j} \leftarrow \sqrt{p_S(u_i v_j)}$
    for each $x \in \Sigma$ do
        $X_{x,i,j} \leftarrow \sqrt{p_S(u_i x v_j)}$
    end
    $M = X^T X$
    $(\lambda_i, w_i, w_i^*) \leftarrow$ eigenvalues of $M$, and corresponding eigenvectors and dual eigenvectors
    $W = [w_1 \dots w_d]$, $W^* = [w_1^* \dots w_d^*]$, $D = diag(\lambda_i)_{1 \leq i \leq d}$
    $\boldsymbol{q}_u = (q_S(u_1), \dots, q_S(u_n))$, $\boldsymbol{q}_v = (q_S(v_1), \dots, q_S(v_m))$, $\boldsymbol{I} = W^T \boldsymbol{q}_v$, $\boldsymbol{T} = (\boldsymbol{q}_u^T W^* D^{-1/2})^T$
    for each $x \in \Sigma$ do
        $(M_x) \leftarrow (W^T X_x^T W^* D^{-1/2})^T$
    end
    return $A = \langle \Sigma, \{M_x\}_{x \in \Sigma}, \boldsymbol{I}, \boldsymbol{T} \rangle$
end
```

In a second part, we wish to estimate the model parameters from the singular vectors. Once the dimension is fixed, one builds the matrices of singular vectors (left and right) W_S, and W_S^*, and the eigenvalues D_S. The matrix of the operator $\dot{x}$ in the basis $\boldsymbol{w}_{S,1}, \dots, \boldsymbol{w}_{S,d}$ is close to $W_S^T X_{S,x}^T W_S^* D_S^{-1/2}$.

Once the WA A_S is computed by the algorithm 1, one considers the QWA $Q_S = A_S \otimes A_S$, which computes q_S^2. If q_S^2 is convergent, one can normalize it in order to obtain a probability distribution. In next sections, we address the convergence of q_S^2 towards q^2.

4.1. Example

To illustrate the method, let us consider the distribution p computed by the following WA:

$$\boldsymbol{I} = \begin{pmatrix} 9/67 \\ 18/67 \\ 9/67 \end{pmatrix} M_a = \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 0.64 \end{pmatrix} \boldsymbol{T} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Let us consider a sample S of size 1000, i.i.d with respect to p (see table 1). The empirical distribution p_S is:

	ϵ	a	a^2	a^3	a^4	a^5	a^6	a^7	a^8
p	0.5373	0.2270	0.1064	0.0545	0.0299	0.0173	0.0104	0.0064	0.0040
p_S	0.553	**0.223**	**0.106**	0.056	0.022	0.022	0.011	0.004	0.002
p_{A_2}	**0.5374**	0.2167	0.1030	**0.0544**	**0.0306**	**0.0177**	**0.0105**	**0.0062**	**0.0037**
r_{A_1}	0.5530	0.2226	0.1083	0.0553	0.0303	0.0200	0.0189	0.0259	0.0434
$r^2_{A_2}$	0.5451	0.2198	0.1045	0.0552	0.0310	0.0180	0.0106	0.0063	0.0038

Table 1: Values of p_S, p_{A_2}, r_{A_1} and $r^2_{A_2}$.

First, one applies the regular spectral algorithm to find a WA A_1. One will consider the sets $U = \{\epsilon, a, aa\}$ of prefixes and $V = \{\epsilon, a, aa\}$ of suffixes. The WA A_1 computed from p_S gives:

$$I = \begin{pmatrix} -2.439648e-04 \\ -8.318615e-03 \\ -6.055615e-01 \end{pmatrix} T = \begin{pmatrix} -0.039385 \\ -0.418262 \\ -0.907472 \end{pmatrix} M_a = \begin{pmatrix} 1.841546 & -0.112868 & -0.427651 \\ -0.035153 & 0.14176 & -0.919531 \\ -0.003970 & -0.027406 & 0.429745 \end{pmatrix}$$

Because of the negative entry in M_b, it is not possible to ensure that the series computed by A_1 is positive.

From the sample S, one can build the mapping $q_S = p_S^{1/2}$. One considers the following sets $U = \{\epsilon, a\}$ of prefixes and $U = \{\epsilon, a\}$ of suffixes. The same algorithm applied to q_S provides an estimate of $q = p^{1/2}$. One obtains the WA A_2:

$$I = \begin{pmatrix} 0.00991992 \\ -0.88085277 \end{pmatrix} T = \begin{pmatrix} 0.54554234 \\ -0.83808327 \end{pmatrix} M_a = \begin{pmatrix} 0.49736109 & 1.32375247 \\ 0.02290183 & 0.66584823 \end{pmatrix}$$

The sum of $r^2_{A_2}$ equals $r^2_{A_2}(\Sigma^*) = 1.01443774166$, one can normalize the series to obtain a probability distribution. The table 1 describes the values computed by the series r_{A_1}, $r^2_{A_2}$ and its normalization p_{A_2}.

5. Concentration Inequalities

5.1. Hankel Matrix

Let p be a probability distribution over Σ^*. Let $U = \{u_1, \dots\}$ and $V = \{v_1, \dots\}$ be two finite sets of strings. Let us consider the Hankel matrix X of $q = p^{1/2}$ defined by $X_{ij} = q(u_i v_j)$. Let S be a sample of size N i.i.d. according to p. X_S is defined as the empirical Hankel matrix built from $q_S = p_S^{1/2}$. In this section, we will bound the difference between$X$ and X_S.

Proposition 4 *Let X be the Hankel matrix of $q = p^{1/2}$ restricted to finite sets U and V of prefixes and suffixes. Let $m = max(|U|, |V|)$. Let X_S the empirical estimator of X from a sample S of size N i.i.d. with respect to p. Then, with probability at least $1 - \delta$ ($\delta > 0$):*

$$\Delta_X = \|X - X_S\|_F \leq \frac{m + \sqrt{m \log(\frac{1}{\delta})}}{\sqrt{N}}$$

First, one can bound the variance of X_S as an estimate of X:

Lemma 5 *Let p be a probability distribution over Σ^*. Let p_S its empirical estimate from an i.i.d sample S of size N with respect to p. Let $q = p^{1/2}$ and $q_S = p_S^{1/2}$. One has*

$$\sum_{u \in U, v \in V} \mathbb{E}(q_S(uv) - q(uv))^2 \leq |U||V| \frac{1}{N} \leq \frac{m^2}{N}.$$

Proof One has

$$\mathbb{E}(q_S(w) - q(w))^2 = \mathbb{E}(\frac{q_S^2(w) - q^2(w)}{q_S(w) + q(w)})^2 \leq \mathbb{E}(\frac{q_S^2(w) - q^2(w)}{q(w)})^2 = \frac{q^2(w)(1 - q^2(w))}{Nq^2(w)} \leq \frac{1}{N}$$

■

Proof [of proposition 4] This proof uses a construction similar to the proof of Proposition 19 in Hsu et al. (2009a). ■

Let $q_U = (q(u_1), \dots)$ and $q_V = (q(v_1), \dots)$. One has:

Proposition 6 $\|q_U - q_{S\,U}\|_F = O(\frac{\sqrt{m \log(\frac{1}{\delta})}}{\sqrt{N}})$ *and* $\|q_V - q_{S\,V}\|_F = O(\frac{\sqrt{m \log(\frac{1}{\delta})}}{\sqrt{N}})$.

Proof One uses the same arguments than before, with $|U| = 1$ or $|V| = 1$. ■

5.2. Singular Values

Let us first recall a known result: given an matrix A of rank d, and estimate of A denoted A_S, one can rewrite A_S as the sum $A + E$ where E models the error. One has the following result from Stewart and Sun (1990).

Proposition 7 *(Thm 4.11 in Stewart and Sun (1990)). Let* $A \in \mathbb{R}^{m \times n}$ *with* $m \geq n$*, and let* $A_S = A + E$*. If the singular values of* A *and* A_S *are* $(\lambda_1 \geq \ldots \geq \lambda_n)$ *and* $(\lambda_{S,1} \geq \ldots \geq \lambda_{S,n})$ *respectively, then*

$$|\lambda_{S,i} - \lambda_i| \leq \|E\|_2, i = 1, \dots, n.$$

One then has:

Proposition 8 *Let* X *be the Hankel matrix of* $q = p^{1/2}$ *restricted to finite sets* U *and* V *of prefixes and suffixes. Let* $m = max(|U|, |V|)$*. Let* X_S *the empirical estimator of* X *from a sample* S *of size* N *i.i.d. with respect to* p*. Let* $\lambda_1 \geq \cdots \geq \lambda_d$ *be the singular values of* X*, and* $\lambda_{S1} \geq \cdots \geq \lambda_{Sd}$ *the corresponding singular values of* X_S*. Let* $D^{1/2} = diag(\lambda_1, \dots, \lambda_d)$ *and* $D_S^{1/2} = diag(\lambda_{S1}, \dots, \lambda_{Sd})$*. Then, with probability at least* $1 - \delta$ $(\delta > 0)$ *one has:*

$$\forall 1 \leq \imath \leq d, |\lambda_i - \lambda_{Si}| \leq \frac{m + \sqrt{m \log(\frac{1}{\delta})}}{\sqrt{N}}, \|D^{1/2} - D_S^{1/2}\|_F \leq d^{1/2} \frac{m + \sqrt{m \log(\frac{1}{\delta})}}{\sqrt{N}}$$

Proof One applies the fact that $\|\|_2 \leq \|\|_F$ to proposition 4. ■

5.3. Singular Vectors

One wishes to bound the norm of the singular vectors $\|W_k\|$. To do so, one uses a result about eigenvectors and PCA from Zwald and Blanchard (2006). Let A be a symmetric positive Hilbert-Schmidt operator with positive eigenvalues $\lambda_1 > \cdots > \lambda_d > 0$. $\delta_k = \frac{1}{2}(\lambda_k - \lambda_{k+1})$. Let B be a symmetric positive Hilbert-Schmidt operator such that $\|B\|_F < \delta_k/2$. Let W_k (resp. $W_{S,k}$) be the matrix of the r first eigenvectors of A (resp. $A + B$). One has:

Lemma 9 *(Theorem 3 in Zwald and Blanchard (2006))* $\|W_k - W_{S,k}\|_F \leq \frac{2\|B\|_F}{\delta_k}$

From this, one can deduce:

Proposition 10 *Let X be the rank d Hankel matrix of $q = p^{1/2}$ restricted to finite sets U and V of prefixes and suffixes. Let $m = max(|U|, |V|)$.Let X_S the empirical estimator of X from a sample S of size N i.i.d. with respect to p. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the singular values of X. Let W and W_S be the matrices defined by $W_{ij} = w_j(v_i)$ and $W_{Sij} = w_{Sj}(v_i)$, where the w_j (resp. w_{Sj}) are the right singular vectors of X (resp. X_S). Let W^* and W^*_S be the matrices defined the same way with the left singular vectors of X (resp. X_S). Then, with probability at least $1 - \delta$ ($\delta > 0$) one has:*

$$\|W - W_S\|_F = O(\frac{m^{3/2}\sqrt{\log(\frac{1}{\delta})}}{\lambda_d\sqrt{N}}), \|W^* - W^*_S\|_F = O(\frac{m^{3/2}\sqrt{\log(\frac{1}{\delta})}}{\lambda_d\sqrt{N}})$$

Proof One applies the lemma 9. One has $\delta_d = \lambda_d/2$ because X has rank d. On has that $\|X^T X - X_S^T X_S\|_F = O(\|X\|_F \Delta_X)$, and $\|X\|_F \leq \sqrt{m}$. By symmetry, one has the result for W^*. ■

6. Consistency

In this section and in the next section, one supposes that the distribution p is such that $q = p^{1/2}$ is rational of rank d. Let S be sample i.i.d with respect to p, $|S| = N$. Let U and V two finite sets of strings such that the Hankel matrix of q has rank d. Let $m = \max(|U|, |V|)$. Let X be the Hankel matrix of q, and X_S be the Hankel matrix of $q_S = p_S^{1/2}$.

6.1. Rank Estimation

Theorem 11 *Let Λ be the set of singular values of X_S. Let Λ_μ be the subset of singular values of X_S greater than μ. For a given confidence parameter δ, let $d' = |\Lambda_\mu|$ for $\mu = \frac{m+\sqrt{m\log(\frac{1}{\delta})}}{\sqrt{N}}$. With probability greater than $1 - \delta$, one has $d \geq d'$.*

Proof Straightforward from Proposition 4 and Proposition 7: with probability greater than $1 - \delta$, the singular values in Λ_s match non-zeros singular values from the target Hankel matrix X. ■

Theorem 12 *Let λ_d the smallest non-zero singular value of X. Let Λ be the set of singular values of X_S. Let Λ_μ be the subset of singular values of X_S greater than μ. For a given confidence parameter*

δ, let $d' = |\Lambda_\mu|$ for $\mu = \frac{m+\sqrt{m\log(\frac{1}{\delta})}}{\sqrt{N}}$. Let us suppose that $N > \frac{4}{\lambda_d^2}\left(m + \sqrt{m\log(\frac{1}{\delta})}\right)^2$. Then, with probability greater than $1-\delta$, one has $d = d'$.

Proof The condition $N > \frac{4}{\lambda_d^2}(m + \sqrt{m\log(\frac{1}{\delta})})^2$ implies that $\mu < \frac{\lambda_d}{2}$, thus the corresponding singular value $\lambda_{S,d}$ from X_S satisfies $\lambda_{S,d} > 2\mu - \|X - X_S\|_2$. This quantity is greater than μ with probability at least $1-\delta$. ■

6.2. Parameters Estimation

One supposes here that the correct rank d has been found. Let $< I, (M_x)_{x\in\Sigma}, T >$ a linear representation of q in the basis of the right singular vectors $B = \{w_1, \dots w_d\}$ of X. Let I_S,$(M_{Sx})_{x\in\Sigma}$, T_S the linear representation outputted by the algorithm 1.

Proposition 13 *One has the following properties: $\|I - I_S\|_F = O(\frac{m^{3/2}\sqrt{d\log(\frac{1}{\delta})}}{\lambda_d\sqrt{N}})$, $\|T - T_S\|_F = O(\frac{m^{3/2}\sqrt{d\log(\frac{1}{\delta})}}{\lambda_d^2\sqrt{N}})$, $\|M_x - M_{xS}\|_F = O(\frac{m^2 d\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2\sqrt{N}})$.*

Proof One uses the former inequalities. For the first inequality, one uses the bounds $\|W\|_F \leq \sqrt{d}$ and $\|q_V\|_F \leq 1$. For the second and the third inequalities, one uses also $\|D^{-1/2} - D_S^{-1/2}\|_F = O(\frac{1}{\lambda_d^2}\|D^{1/2} - D_S^{1/2}\|_F)$ and $\|D^{-1/2}\|_F \leq \frac{\sqrt{d}}{\lambda_d}$. ■

Proposition 14 *Let $I^{\otimes 2} = I \otimes I$, $T^{\otimes 2} = T \otimes T$, $M_x^{\otimes 2} = M_x \otimes M_x$. One has:*

$$\|I^{\otimes 2} - I_S^{\otimes 2}\|_F = O(\|I\|_F\|I - I_S\|_F), \|T^{\otimes 2} - T_S^{\otimes 2}\|_F = O(\|T\|_F\|T - T_S\|_F)$$

$$\|M_x^{\otimes 2} - M_x^{\otimes 2}{}_S\|_F = O(\|M_x\|_F\|M_x - M_{xS}\|_F)$$

Proof One uses the properties $(M + \Delta_M) \otimes (M + \Delta_M) = M \otimes M + \Delta_M \otimes M + M \otimes \Delta_M + \Delta_M \otimes \Delta_M$, and $\|A \otimes B\|_F = \|A\|_F\|B\|_F$. ■

7. Convergence

7.1. Convergence of the series

We want to show that given a confidence parameter δ, there exists a sample size from which the WA provided by the algorithm 1. computes a convergent series.

Proposition 15 *Let $A =< I, (M_x)_{x\in\Sigma}, T >$ be the linear representation of q in a residual basis $B = \{w_1, \dots, w_d\}$. Let $A^{\otimes 2} = A \otimes A =< I^{\otimes 2}, (M_x^{\otimes 2})x \in \Sigma, T^{\otimes 2} >$. Let $M^\star = \sum_{x\in\Sigma} M_x^{\otimes 2}$. Then $\rho(M^\star) < 1$.*

Proof Let $A_J =< J, (M_x^{\otimes 2})x \in \Sigma, T^{\otimes 2} >$ with $J = (\alpha_{ij})_{1 \leq i,j \leq d}$. A_J computes the series $r_{A_J}(u) = \sum_{i,j} \alpha_{ij} w_i(u) w_j(u)$. As $\forall i, w_i^2$ is convergent, one has $\forall i, j, w_i w_j$ is convergent – by Schwartz's inequality. Suppose that J is a left eigenvector of $M^\star$, with eigenvalue λ. The series computed with $M^{\star T} J$ as initial vector is $u \mapsto r_{A_J}(\Sigma u) = \lambda r_{A_J}(u)$. As r_{A_J} is convergent, $\lambda < 1$. ■

It is clear that, by local differentiability of spectral radius, for a WA A_S outputted by the algorithm 1, one has $|\rho(M_S^\star) - \rho(M^\star)| = O(\Delta_X)$. Let $\rho_\otimes = \frac{\rho(M^\star)+1}{2}$. From now one supposes that the sample size is large enough to ensure that $\rho(M_S^\star) < \rho_\otimes$.

7.2. Pointwise Convergence

Proposition 16 *Let A_S be the WA outputted by the algorithm 1, computing the series r_{A_S}. Let $w \in \Sigma^k$, let δ be a confidence parameter. One has, with probability δ, $|r_{A_S}^2(w) - q^2(w)| = O(\frac{k|\Sigma|m^2 d\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2 \sqrt{N}})$*

Proof Let us recall that $\rho(A \otimes A) = \rho(A)^2$, and that an eigenvector v_λ of A corresponds to an eigenvector $v_{\lambda^2} = v_\lambda \otimes v_\lambda$ of $A \otimes A$. Let $\mathcal{M} = \{M_x\}_{x \in \Sigma}$. One defines the *generalized spectral radius* $\rho(\mathcal{M}) = \limsup_{k \to \infty} (\rho_k(\mathcal{M}))^{1/k}$ where $\rho_k(\mathcal{M}) = \sup\{\rho(M_{x_1} \dots M_{x_k})\}_{M_{x_i} \in \mathcal{M}}$. One also defines $\hat{\rho}_k(\mathcal{M}) = \sup\{\|M_{x_1} \dots M_{x_k}\|_F\}_{M_{x_i} \in \mathcal{M}}$ and the *joint spectral radius* $\hat{\rho}(\mathcal{M}) = \limsup_{k \to \infty} (\hat{\rho}_k(\mathcal{M}))^{1/k}$ (see Theys (2005),Blondel et al. (2008)).

Let us suppose that there exists $x_1 \dots x_k$ such that $\rho(M_{x_1} \dots M_{x_k}) = \mu^{k/2} > \rho_\otimes^{k/2}$. Let v_μ a corresponding eigenvector, and u be any d-dimensional vector such that $v_\mu^T u = C' > 0$. The WA $< v_\mu^{\otimes 2}, M_x^{\otimes 2}, u^{\otimes 2} >$ computes a positive series r for which there exists C such that $r(\Sigma^{nk}) < C\rho_\otimes^{nk}$. It also satisfies $r((x_1 \dots x_k)^n) = C'^2 \mu^{nk}$ which is contradictory. Thus, $\rho(\mathcal{M}) \leq \rho_\otimes < 1$. By the property $\rho(\mathcal{M}) = \hat{\rho}(\mathcal{M})$, one has that $\|I^{\otimes 2} M_{x_1}^{\otimes 2} \dots M_{x_k}^{\otimes 2}\|_F$ and $\|M_{x_1}^{\otimes 2} \dots M_{x_k}^{\otimes 2} T^{\otimes 2}\|_F$ tend to 0 as $k \to \infty$, thus are bounded by a real number L.
One has $\|I^{\otimes 2} M_{x_k}^{\otimes 2} \dots M_{x_{j-1}}^{\otimes 2} (M_{x_j}^{\otimes 2} - {M_{x_j}}_S^{\otimes 2}) M_{x_{j+1}}^{\otimes 2} \dots M_{x_k}^{\otimes 2} T^{\otimes 2}\|_F = O(L^2 \|M_{x_j}^{\otimes 2} - {M_{x_j}}_S^{\otimes 2}\|_F)$.
Finally, $|r_{A_S}^2(x_1 \dots x_k) - q^2(x_1 \dots x_k)| = O(\frac{k|\Sigma|m^2 d\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2 \sqrt{N}})$. ■

7.3. ℓ_1 Convergence

From the proposition 16 one has the property $\sum_{w<k} |r_{A_S}^2(w) - q^2(w)| = O(\frac{kl^2|\Sigma|^{k+1} m^2 d\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2 \sqrt{N}})$. The convergence of the tail comes from the exponential decreasing of a convergent rational series.

Proposition 17 *Let q be a rational series, $p = q^2$ a distribution, S a sample i.i.d. with respect to p. Let A_S be the WA outputted by the algorithm 1, computing the series r_{A_S}. Let δ be a confidence parameter. Let $\epsilon > 0$. There exists C such that*

$$N > C\epsilon^{-2-2\frac{\log(|\Sigma|)}{\log(\rho)}} (\frac{\log(1/\epsilon)}{\log(\rho)}) \frac{m^4 d^2 \log(1/\delta)}{\lambda_d^2}$$

implies, with probability $1 - \delta$, $\sum_w |r_{A_S}^2(w) - q^2(w)| < \epsilon$

Proof Let $\rho = \frac{\rho(M^\star)+1}{2}$. Let us suppose N large enough to ensure that there exists K such that $\sum_{|w|\geq k} |r^2_{A_S}(w)| \leq K\rho^k$ and $\sum_{|w|\geq k} |q^2(w)| \leq K\rho^k$.

For a given ϵ, let $k = \frac{\log(4K/\epsilon)}{\rho}$: one the has $\sum_{w\geq k} |r^2_{A_S}(w)| + \sum_{w\geq k} |q^2(w)| \leq \frac{\epsilon}{2}$ and

$$\begin{aligned}\textstyle\sum_{|w|<k} |q^2(w) - r^2_S(w)| \quad &< K|\Sigma|^{\frac{\log(4K/\epsilon)}{\log(\rho)}} \frac{\log(4K/\epsilon)}{\log(\rho)} m^2 d \frac{\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2\sqrt{N}} \\ &< K(4K/\epsilon)^{\frac{\log(|\Sigma|)}{\log(\rho)}} \frac{\log(4K/\epsilon)}{\log(\rho)} m^2 d \frac{\sqrt{\log(\frac{1}{\delta})}}{\lambda_d^2\sqrt{N}}\end{aligned}$$

One can choose a convenient C such that the conclusion holds. ■

The previous bound indicates a convergence rate of $O(\epsilon^{-2-2\frac{\log(|\Sigma|)}{\log(\rho)}})$. Under some certain assumptions – not too restrictive – one can obtain a tighter bound: if the rank of q^2 greater than $\frac{d(d+1)}{2}$, then one has the following result:

Proposition 18 *Let $q > 0$ be a rational series such that q^2 is a probability distribution of rank $\geq \frac{d(d+1)}{2}$, S an i.i.d sample with respect to q^2. Let A_S be the WA outputted by the algorithm* 1*, computing the series r_{A_S}. There exists C such that, for any $\epsilon > 0$, and for any confidence parameter $0 < \delta < 1$, the condition*

$$N > C\frac{\log^4(1/\epsilon)}{\epsilon^2}\frac{m^4d^2\log(1/\delta)}{\lambda_d^2}$$

implies that $\sum_{w\in\Sigma^*} |r^2_{A_S}(x) - q^2(x)| \leq \epsilon$ with probability $1-\delta$.

Proof [sketch] It follows the proof in Hsu et al. (2009b). In the original proof, one of the key property used is the fact that $\sum_{x\in\Sigma} |M_x|$ is pseudo-stochastic, i.e. $\| \sum_{x\in\Sigma} |M_x| \|_\infty \leq 1$, coming directly from the definition of an HMM. In the case of WAs, this is not true in general, but one can prove an analogous property for absolutely convergent rational series. ■

8. Likelihood Maximization

We show here how to compute the gradient of the log-likelihood function. The computing of the Hessian uses the same techniques. Let $A_{\boldsymbol{\theta}}$ be a QWA, with parameters $\boldsymbol{\theta}$, and let S be a given sample supposed to be i.i.d with respect to an unknown distribution q^2. One supposes that the series $r^2_{A_{\boldsymbol{\theta}}}$ is convergent, but unnormalized, thus the probability of a string w is given by $\frac{r^2_{A_{\boldsymbol{\theta}}}(w)}{r^2_{A_{\boldsymbol{\theta}}}(\Sigma^*)}$. Without any *a priori* distribution on the parameters $\boldsymbol{\theta}$, a good candidate for $\boldsymbol{\theta}$ is $\arg\max_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}))$, with

$$L(\boldsymbol{\theta}) = \sum_{w\in S}[\log(r^2_{A_{\boldsymbol{\theta}}}(w)) - \log(r^2_{A_{\boldsymbol{\theta}}}(\Sigma^*))]$$

In the general case, the mappings attached to each hidden state are not probability distributions, and the resulting probability cannot be reduced to a convex combination of those mappings. Thus, the convergence of an EM-type algorithm is not guaranteed.

We provide a set of algorithms which can be used to compute the gradient of $L(\boldsymbol{\theta})$. From there, one can perform a local likelihood maximization. The rank, and the parameters provided by the spectral algorithm can be used as a starting point for this maximization. For any string $w \in \Sigma^*$, and any WA parameters $\boldsymbol{\theta}$, let us denote $q_w(\boldsymbol{\theta})$ the value of the series computed with $\boldsymbol{\theta}$ for the string w. Let us denote $q_{\Sigma^*}(\boldsymbol{\theta})$ the sum of the series computed with parameters $\boldsymbol{\theta}$.

8.1. Computation of $\nabla_{r_w}(\theta)$ and $\nabla_{r_{\Sigma^*}}(\theta)$

```
Algorithm 2: $\nabla_{r_w}(\boldsymbol{\theta})$ Gradient Algorithm
Data: A WA with parameters $\theta = (\boldsymbol{I}, (M_x)_{x\in\Sigma}, \boldsymbol{T}$, a string $w$
Result: A WA with parameters $\nabla_{r_w}(\theta)$
begin
    Let $\boldsymbol{f}_0 = \boldsymbol{I}$, let $\boldsymbol{b}_{|w|+1} = \boldsymbol{T}$
    for $i$ from $1$ to $|w|$ do
        let $\boldsymbol{f}_i = M_{w_i}^T \cdot \boldsymbol{f}_{i-1}$ ,$\boldsymbol{b}_{|w|-i+1} = M_{w_{|w|-i+1}} \cdot \boldsymbol{b}_{|w|-i+2}$
    end
    Let $\nabla_{r_w}(\theta).\boldsymbol{I} = \boldsymbol{f}_{|w|}$, $\nabla_{r_w}(\theta).\boldsymbol{T} = \boldsymbol{b}_1$, $\nabla_{r_w}(\theta).M_a = 0_{d\times d}$
    for $i$ from $1$ to $|w|$ do
        let $\nabla_{r_w}(\theta).M_{w_i} + = \boldsymbol{f}_{i-1}\boldsymbol{b}_{i+1}^T$
    end
    return $\nabla_{r_w}(\theta)$
end
```

```
Algorithm 3: $\nabla_{r_{\Sigma^*}}(\boldsymbol{\theta})$ Gradient Algorithm
Data: A WA with parameters $\theta = (\boldsymbol{I}, (M_x)_{x\in\Sigma}, \boldsymbol{T}$
Result: A WA with parameters $\nabla_{r_{\Sigma^*}}(\theta)$
begin
    Let $M_\Sigma = (I - \sum_{x\in\Sigma} M_x)^{-1}$, et $\nabla_{r_{\Sigma^*}}(\theta).\boldsymbol{I} = M_\Sigma \boldsymbol{T}$, $\nabla_{r_{\Sigma^*}}(\theta).\boldsymbol{T} = M_\Sigma^T \boldsymbol{I}$
    for $x \in \Sigma$ do
        Let $\nabla_{r_{\Sigma^*}}(\theta).M_x = M_\Sigma^T \boldsymbol{I}\boldsymbol{T}^T M_\Sigma^T$
    end
    return $\nabla_{r_{\Sigma^*}}(\theta)$
end
```

Let $M_\Sigma = (Id - \sum M_x)^{-1}$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_I, \boldsymbol{\theta}_{M_{x_1}}, \ldots \boldsymbol{\theta}_T)$. The ways to compute $\nabla_{r_w}(\boldsymbol{\theta})$ and $\nabla_{r_{\Sigma^*}}(\boldsymbol{\theta})$ are detailed in algorithms 2 and 3. They mainly use the following properties:

$$\begin{array}{ll} r_{\Sigma^*}(\boldsymbol{\theta}) & = \boldsymbol{I}^T(Id - \sum M_x)^{-1}\boldsymbol{T} = \boldsymbol{I}^T M_\Sigma \boldsymbol{T} \\ r_{\Sigma^*}(\boldsymbol{\theta} + d\boldsymbol{\theta}_I) & = r_{\Sigma^*}(\boldsymbol{\theta}) + d\boldsymbol{\theta}_I^T M_\Sigma \boldsymbol{T} \\ r_{\Sigma^*}(\boldsymbol{\theta} + d\boldsymbol{\theta}_I) & = r_{\Sigma^*}(\boldsymbol{\theta}) + \boldsymbol{I}^T M_\Sigma d\boldsymbol{\theta}_I \\ r_{\Sigma^*}(\boldsymbol{\theta} + d\boldsymbol{\theta}_{M_x}) & \sim r_{\Sigma^*}(\boldsymbol{\theta}) + \boldsymbol{I}^T M_\Sigma \cdot d\boldsymbol{\theta}_{M_x} \cdot M_\Sigma \boldsymbol{T} \end{array}$$

Let us denote ∇_I the gradient restricted to $\boldsymbol{I}$, ∇_T the gradient restricted to $\boldsymbol{T}$ etc. One can check that :

$$\nabla_{I r_{\Sigma^*}(\boldsymbol{\theta})} = M_\Sigma \boldsymbol{T}, \nabla_{T r_{\Sigma^*}(\boldsymbol{\theta})} = M_\Sigma^T \boldsymbol{I}, \nabla_{M_x r_{\Sigma^*}(\boldsymbol{\theta})} = M_\Sigma^T \boldsymbol{I}\boldsymbol{T}^T M_\Sigma^T$$

8.2. Computation of $\nabla_{r^2_{\Sigma^*}}(\theta)$

Let r be a rational series computed with parameters $\boldsymbol{\theta}$. Let $\boldsymbol{\theta} \otimes \boldsymbol{\theta} = (\boldsymbol{\theta}_I \otimes \boldsymbol{\theta}_I, \boldsymbol{\theta}_{M_x} \otimes \boldsymbol{\theta}_{M_x}, \ldots \boldsymbol{\theta}_T \otimes \boldsymbol{\theta}_T)$ the parameters computing $s = r^2$.

One has $(\boldsymbol{\theta} + d\boldsymbol{\theta}) \otimes (\boldsymbol{\theta} + d\boldsymbol{\theta}) = \boldsymbol{\theta} \otimes \boldsymbol{\theta} + d\boldsymbol{\theta} \otimes \boldsymbol{\theta} + \boldsymbol{\theta} \otimes d\boldsymbol{\theta} + d\boldsymbol{\theta} \otimes d\boldsymbol{\theta}$. One then has:

$$\begin{array}{ll} s_{\Sigma^*}((\boldsymbol{\theta} + d\boldsymbol{\theta}) \otimes (\boldsymbol{\theta} + d\boldsymbol{\theta})) = s_{\Sigma^*}(\boldsymbol{\theta} \otimes \boldsymbol{\theta}) & \\ +\nabla_{s_{\Sigma^*}}(\boldsymbol{\theta} \otimes \boldsymbol{\theta})^T (d\boldsymbol{\theta} \otimes \boldsymbol{\theta} + \boldsymbol{\theta} \otimes d\boldsymbol{\theta}) & (1) \\ +\ldots & \end{array}$$

The rows (1) corresponds to the first order term of the Taylor developement of $r^2_{\Sigma^*}(\boldsymbol{\theta})$. Let $\boldsymbol{v}$ be a vector of dimension n^2, x and y of dimension n. One will denote $\overline{\boldsymbol{v}}$ the matrix $\overline{\boldsymbol{v}}_{ij} = \boldsymbol{v}_{n(i-1)+j}$. One can check that $\boldsymbol{v}^T(x \otimes y) = x^T \overline{\boldsymbol{v}} y$. One then has:

$$\nabla_{r^2_{\Sigma^*}(\boldsymbol{\theta})} = \quad \boldsymbol{\theta}^T [\overline{\nabla_{s_{\Sigma^*}}(\boldsymbol{\theta} \otimes \boldsymbol{\theta})} + \overline{\nabla_{s_{\Sigma^*}}(\boldsymbol{\theta} \otimes \boldsymbol{\theta})}^T]$$

One can check that the complexity computation of the gradient algorithm is in $O(|w|)$ – thus the complete step is linear in the size of S.

8.3. Computation of $\nabla_L(\theta)$

One then uses the property $\nabla_{\log(r^2_w(\theta))} = 2\frac{\nabla_{r_w(\boldsymbol{\theta})}}{r_w(\boldsymbol{\theta})}$ to obtain

$$L(\theta) = \sum_{w \in S} \left(\nabla_{\log(r^2_w(\theta))} - \nabla_{\log(r^2_{\Sigma^*}(\boldsymbol{\theta}))} \right) = \sum_{w \in S} \left(2\frac{\nabla_{r_w(\boldsymbol{\theta})}}{r_w(\boldsymbol{\theta})} - \frac{\nabla_{r^2_{\Sigma^*}(\boldsymbol{\theta})}}{r^2_{\Sigma^*}(\boldsymbol{\theta})} \right)$$

9. Experiments

The spectral algorithm provides a consistent estimate of the target parameters, but this estimate is not designed to perform well from a maximum likelihood point of view (i.e. minimizing the $\|\|\|_{KL}$ towards the target). To overcome this, one performs a single tep of gradient ascent of the likelihood after the spectral algorithm. In these experiments, the studied target distribution p is modeled by the following PA, and is not computable by any QWA – the goal here is also to study how the spectral algorithm behaves when the target is not computable by a QWA:

$$\boldsymbol{I} = \begin{pmatrix} 0.3 \\ 0.3 \end{pmatrix} M_0 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.25 \end{pmatrix} M_1 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{T} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

We compare the performances between the Baum-Welch algorithm, the spectral algorithm with 1 likelihood gradient ascent step, the spectral algorithm with complete likelihood maximization process, and the QWA maximum likelihood alone (with random starting parameters).

The Spectral algorithm is performed with sets of prefixes and suffixes sets both equals to $\{\epsilon, 0, 1, 00, 01, 10, 11\}$. We consider 2-state QWAs. The Baum-Welch and the QWA maximum likelihood methods are run for 400 iterations. We compare the performances for several sizes of training sample: 2000, 5000, 10000 and 20000 sequences i.i.d. with respect to p. For each sample size, 100 experiments have been carried out.

Our experiments show that, from a reasonable sample size (5000 examples), the combination spectral algorithm +1 step likelihood gradient ascent performs better than Baum-Welch in density estimation task (for the $\|\|\|_{KL}$ towards the target). Moreover, the computational cost of this combination is far lower than a Baum-Welch run.

sample	2000	5000	10000	20000
dimension 2	0%	11%	97%	100%
spectral +1 step LGA	47%	63%	64%	64%
Baum-Welch	32%	22%	17%	12%
spectral +complete LGA	10%	4%	8%	12%
random QWA LGA	11%	11%	11%	12%
spectral +1 step LGA (vs. Baum-Welch)	47%	65%	71%	72%

Figure 1: Compared performances ($\|\|_{KL}$ towards the target) of spectral algorithm + 1 step likelihood gradient ascent (LGA), Baum-Welch algorithm, spectral algorithm with complete likelihood maximization, and QWA likelihood maximization from random parameters. The first row represents the cases where the 2 first states are considered statistically significant. The last row focuses on the two methods: the spectral+ 1 step ML and Baum-Welch (percentage of cases where the firsone provides better results than the second one).

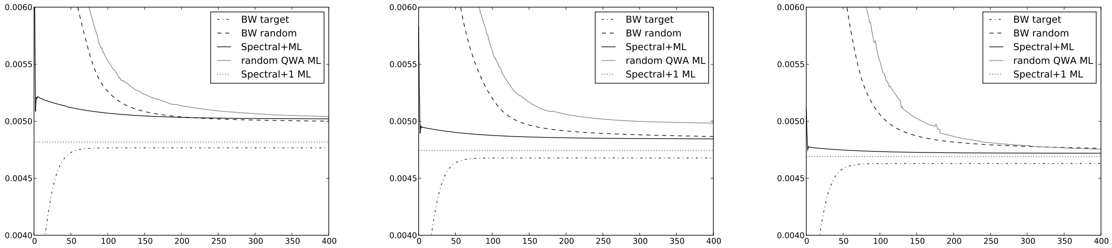

Figure 2: KL-divergence towards the target w.r.t. the number of iterations, for sample sizes of 5000, 10000 and 20000 examples. As a baseline, Baum-Wech has been trained on target parameters (BW target). For readability, Spectral+ML and random QWA ML are smoothed.

10. Conclusion and Further work

The set of methods employed in this paper offers an alternative to the Baum-Welch algorithm for density estimation: it provides both structure identification and consistent parameter estimate, with a low computational cost, and good performances.

To continue this work, one could try to apply this method to the field ofdistributions on trees, with Weighted Tree Automata. Another outlook would be to study how these methods can be used with distributions on substrings (i.e. parts of infinite strings). One could also try, as in Song et al. (2010), to adapt these methods to continuous distributions, using the method of Hilbert space embedding of distributions.

References

Raphaël Bailly and François Denis. Absolute convergence of rational series is semi-decidable. *Inf. Comput.*, 209(3):280–295, 2011.

Raphaël Bailly, François Denis, and Liva Ralaivola. Grammatical inference as a principal component analysis problem. In *Proceedings of the 26th International Conference on Machine Learning*, pages 33–40, Montréal, Canada, June 2009. Omnipress.

Vincent D. Blondel, Michael Karow, Vladimir Protassov, and Fabian R. Wirth, editors. *Special Issue on the Joint Spectral Radius: Theory, Methods and Applications*, volume 428 of *Linear Algebra and its Applications*, pages 2259–2404. Elsevier, 2008.

François Denis, Yann Esposito, and Amaury Habrard. Learning rational stochastic languages. In Gábor Lugosi and Hans-Ulrich Simon, editors, *19th Conference on Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 274–288. Springer, 2006. ISBN 3-540-35294-5.

D. Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models - long version. Technical report, Arxiv archive, 2009a. http://arxiv.org/abs/0811.4413.

Daniel Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proceedings of COLT'2009*. Springer, 2009b.

Le Song, Byron Boots, Sajid M. Siddiqi, Geoffrey J. Gordon, and Alexander J. Smola. Hilbert space embeddings of hidden markov models. In Johannes Fürnkranz and Thorsten Joachims, editors, *ICML*, pages 991–998. Omnipress, 2010.

G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Jacques Theys. *Joint Spectral Radius : theory and approximations*. PhD thesis, UCL - Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 2005.

Ming-Jie Zhao and Herbert Jaeger. Norm observable operator models. *Neural Computation*, 22(7): 1927–1959, 2010.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Proceedings of NIPS'05*, 2006.