Spectral Learning for (probabilistic) Grammatical Inference

Amaury Habrard

amaury.habrard@univ-st-etienne.fr
LaHC, CNRS, University Jean Monnet of Saint-Etienne

ESF Workshop 2014
Combining Learning and Symbolic Analysis for Software Documentation and Mastering Change

joint work with F. Denis, M. Gybels
Outline

1. Context
2. Spectral method
3. Other approaches
Outline

1. Context
2. Spectral method
3. Other approaches
The problem: regression over strings

Context: learn real functions over strings $\Sigma^* \rightarrow \mathbb{R}$

- $\Sigma$: a finite alphabet,
- $\Sigma^*$: strings built over $\Sigma$,
- $S$: training sample of strings from $\Sigma^*$ drawn i.i.d. from an unknown distribution $p$
- Focus on probability distributions but can be extended to general regression problems requiring real labels

Problem: Learning density functions

Given a finite sample $S$ drawn i.i.d. from $p$, find a good estimate of $p$ in a fixed class of models

Amaury Habrard
Spectral Learning for (probabilistic) Grammatical Inference
Automata-based models

- Series \( r : \Sigma^* \rightarrow \mathbb{R} \)
- Natural class: Probabilistic Automata/Hidden Markov Models

Syntactic characterization
Learning: state merging (PDFA)-Expectation Maximization (EM)

\[
\begin{align*}
r(abb) &= \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4} \times \frac{1}{4} + \frac{1}{6} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}
\end{align*}
\]
Weighted automata

- Series $r : \Sigma^* \rightarrow \mathbb{R}$
- Rational series: series that can be computed by a finite weighted automaton

No restriction on the parameters!
Rational Series - Linear representation

$n$-dimensional linear representation

\[ \langle I, (M_x)_{x \in \Sigma}, T \rangle, \ n \geq 1 \]

- \( I, T \in \mathbb{R}^n \), (init/terminal weights)
- \( M_x \in \mathbb{R}^{n \times n} \) (transition weights)

\[ I^t = [1 \ 0], \ T^t = [0 \ 1/4] \]
\[ M_a = \begin{bmatrix} 1/2 & 1/6 \\ 0 & 1/4 \end{bmatrix}, \ M_b = \begin{bmatrix} 0 & 1/3 \\ 1/4 & 1/4 \end{bmatrix} \]

\[ r(abb) = I^t \cdot M_a \cdot M_b \cdot M_b \cdot T \]

\[ \text{rank}(r) : \text{minimal rank} \text{ of operators } M_x \text{ for realizing the series} \]
\[ = \text{minimal number of states} \text{ of the equivalent WA} \]

\[ a : 1/2 \quad a : 1/6 \quad b : 1/3 \]
\[ 1 \rightarrow q_0 \quad b : 1/4 \quad q_1 \rightarrow 1/4 \]
Why WFA

Well studied class [Droste et al,09], lots of algorithms [Mohri,09]

Many possible usage

- probability distributions (this talk)
- Binary classifiers $\text{sign}(r(u) + \theta)$
- Real predictors
- Sequence predictor

Applications

Natural language processing, Computational biology, Speech Recognition, Image processing, System testing, ...
Why for probability distributions

Class of Rational Probability Distributions (RPD)

Probability distributions that can be represented by a rational series/weighted automaton

Pros
- **expressive** class (strictly more expressive than HMM and PA)
- algebraic models → good **characterization for learning**
- **sparsity**

Cons
- general series do not generate density functions - no **syntactic characterization** of RPD
- **undecidable** whether a rational series generate a probability distribution
Example [Denis et al., ’06–’08]

Stochastic language: sparsity

For any $\alpha \in \mathbb{R}$ and $0 < \rho < 1$, there exists $\lambda > 0$ such that the automaton defines a stochastic language, which can be computed by a PA iff $\alpha / \pi \in \mathbb{Q}$. 
Outline

1. Context
2. Spectral method
3. Other approaches
The Hankel matrix

**Definition:** Hankel matrix $H_r$ of a series $r : \Sigma^* \rightarrow \mathbb{R}$

- rows indexed by **prefixes** and columns by **suffixes**
- $H[u, v] = r(uv)$
- Structured by **redundancies**: $w = u_1v_1 = u_2v_2$  
  $\Rightarrow H_r[u_1, v_1] = H_r[u_2, v_2]$
- bi-infinite

\[
H_r = \begin{pmatrix}
\epsilon & a & b & aa & \cdots \\
\epsilon & r(\epsilon) & r(a) & r(b) & r(aa) & \cdots \\
a & r(a) & r(aa) & r(ab) & r(aaa) & \cdots \\
b & r(b) & r(ba) & r(bb) & r(baa) & \cdots \\
aa & r(aa) & r(aaa) & r(aab) & r(aaaa) & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\]

→ a way to **represent strings by real numbers**
**Fundamental result**

**Theorem (Carlyle and Paz ’71, Fliess ’74)**

A series $r$ is rational iff the rank of $H_r$ is finite and $\text{rank}(r) = \text{rank}(H_r)$

**One can recover $r$ from $H_r$**

1. Assume $\text{rank}(H_r) = k$
2. **SVD**: $H_r = LDR^t$ with $L, R \in \mathbb{R}^{\infty \times k}$ and $D \in \mathbb{R}^{k \times k}$ diagonal

\[
\begin{align*}
H_{\text{infinite}} & = L_{\text{infinity x k}} \ D_{k \times k} \ R^t_{k \times \text{infinity}} \\
\end{align*}
\]
Fundamental result

**Theorem (Carlyle and Paz ’71, Fliess ’74)**

A series $r$ is rational iff the **rank** of $H_r$ is **finite** and $\text{rank}(r) = \text{rank}(H_r)$

**One can recover $r$ from $H_r$**

- Assume $\text{rank}(H_r) = k$
- **SVD:** $H_r = LDR^t$ with $L, R \in \mathbb{R}^{\infty \times k}$ and $D \in \mathbb{R}^{k \times k}$ diagonal
- $\langle R^t E, (R^t T_x R)_{x \in \Sigma}, R^t P \rangle$ is a **linear representation** of $r$ s.t.
  - $E = (1, 0, 0, \ldots, 0, \ldots)^t$
  - $P = (r(\epsilon), r(a), r(b), \ldots)^t$
  - $T_x$ constant translation matrix: $T_x[u, v] = \delta_{v=ux}$
- **$k$-dimensional** linear representation
In practice: empirical distribution

- Consider a learning sample $S$, and $p_S$ empirical distribution
- Fix a set of prefixes/suffixes to index the empirical Hankel matrix

$S = \{a, a, b, b, ba, aa\}$

$$H_S = \begin{pmatrix}
\epsilon & a & b & aa \\
\epsilon & 0.0 & 2/6 & 2/6 & 1/6 \\
a & 2/6 & 1/6 & 0.0 & 0.0 \\
b & 2/6 & 1/6 & 0.0 & 0.0 \\
aa & 1/6 & 0.0 & 0.0 & 0.0
\end{pmatrix}$$
Prefix and factor series

- Small prefix/suffix sets may lead to a **loss of information**
- **Better** use of the **sample information** with other series
  - \( \bar{p}(u) = p(u\Sigma^*) = \sum_{w \in \Sigma^*} p(uw) \) - **prefix**-based
  - \( \hat{p}(u) = \sum_{w_1, w_2 \in \Sigma^*} p(w_1uw_2) \) - **factor**-based

\( \bar{p}(u) \) and \( \hat{p}(u) \) are **rational** can be used easily to **find** the parameters of \( p \)

\[
\begin{align*}
\bar{H}_S &= \begin{pmatrix}
\epsilon & a & b & aa \\
\epsilon & 1 & 3/6 & 3/6 & 1/6 \\
a & 3/6 & 1/6 & 0.0 & 0.0 \\
b & 3/6 & 1/6 & 0.0 & 0.0 \\
aa & 1/6 & 0.0 & 0.0 & 0.0
\end{pmatrix} \\
\hat{H}_S &= \begin{pmatrix}
\epsilon & a & b & aa \\
\epsilon & 14/6 & 7/6 & 3/6 & 1/6 \\
a & 7/6 & 1/6 & 0.0 & 0.0 \\
b & 3/6 & 1/6 & 0.0 & 0.0 \\
aa & 1/6 & 0.0 & 0.0 & 0.0
\end{pmatrix}
\end{align*}
\]
Very sparse matrices

<table>
<thead>
<tr>
<th></th>
<th>$H_S$</th>
<th>$\bar{H}_S$</th>
<th>$\hat{H}_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of the full matrix:</td>
<td>$55,000 \times 60,000$</td>
<td>$55,000 \times 450,000$</td>
<td>$450,000 \times 450,000$</td>
</tr>
<tr>
<td>Elements in the full matrix:</td>
<td>111,677</td>
<td>891,902</td>
<td>7,272,311</td>
</tr>
<tr>
<td>Sparsity of the full matrix:</td>
<td>99.9967 %</td>
<td>99.9965 %</td>
<td>99.9964 %</td>
</tr>
<tr>
<td>Elements in the 500x500 matrix:</td>
<td>4,173</td>
<td>12,591</td>
<td>29,847</td>
</tr>
<tr>
<td>Sparsity of the 500x500 matrix:</td>
<td>99.3308 %</td>
<td>94.9636 %</td>
<td>88.0612 %</td>
</tr>
</tbody>
</table>
Spectral learning of rational distribution

**Basic Algorithm**

**Input:** $S$ sample of strings, $k$ estimated rank, $H_S \in \mathbb{R}^{m \times n}$

1. Build $H_S$ from the **empirical distribution** $p_S$
2. Apply a rank-$k$ **truncated SVD** of $H_S = LDR^T$

**Output** $\langle R^t E, (R^t T_x R)_{x \in \Sigma}, R^t p_S \rangle$

Many different approaches [Hsu et al., COLT’09], [Bailly et al., ICML’09], [Balle et al., ICML’12], ...
Properties

[Hsu et al., COLT’09; Bailly et al., ICML’09, 11; Balle, 2013]

- **Consistency** (in the realizable case)
  - for parameters $T, I, M_X$, in $O(\sqrt{\frac{k^4 \log(1/\delta)}{\sigma_k^2 |S|}})$
  - for the rank when $|S| > \left(\frac{k}{\sigma_k^2} \left(1 + \sqrt{\log(1/\delta)}\right)^2\right)$

- **Probably Approximatively Correct (PAC) bounds**
  $\exists c > 0$ s.t. $\forall \epsilon > 0, \delta > 0$, $|S| > c \log(1/\delta) \log^2(1/\epsilon)/\epsilon^2$
  implies $\sum_{w \in \Sigma^*} |r_S(w) - p(w)| < \epsilon$, with proba $1 - \delta$

- **Kernels**, feature-based representations
The quality of information in the Hankel matrix is crucial

- Previous study seemed to indicate that the concentration is **dimension dependent** \((d)\), for \(H \in \mathbb{R}^{d \times d}\)
  \[
  \|H_S - H\| \leq \frac{6C}{\sqrt{|S|}}(\sqrt{\log d} + \sqrt{\log(1/\delta)})
  \]

- New bound [Denis et al., ICML’14] **independent of** \(d\), with proba \(1 - 2a(e^a - a - 1)^{-1}\):
  \[
  \|H_S - H\| < \sqrt{\frac{2alt(I_d - M_S)^{-2}T}{|S|}} + \frac{2a}{3|S|}
  \]
  can be optimized in finite dimension

- For **prefix** and **factor** series: use **smoothed** statistics
  \[
  \bar{p}(u) = \sum_{w \in \Sigma^*} \eta^{|w|} p_S(uw) \quad (0 < \eta < 1),
  \]
  \[
  \hat{p}(u) = \sum_{w_1, w_2 \in \Sigma^*} \eta^{|w_1|+|w_2|} p_S(w_1uw_2)
  \]

⇒ Limit the matrix size w.r.t. **computational resources**
Results

\( H \in \mathbb{R}^{m \times n}, \text{rank} = k, \text{sample } S \)

- **Fast** \( O(m \times n \times k) \) (sparse structures, randomized SVD) linear in \( |S| \) - significantly faster than EM

- **Parameterization** is done for free

- Provide **good models** for various tasks [Balle et al., ICML 2014, Gybels et al. ICGI’14] (POS Tagging, PAutomac challenge [Verwer et al., ICGI’12])

  but in average not as good as EM if EM is allowed to explore a large space of parameters

- The same principle can be used for other models (PCFG [Cohen et al., ACL’12-NAACL’13], tree models [Bailly et al., ALT’10; Parikhet al., ICML’11], transducers [Balle et al., ECML’11; Bailly et al., NIPS’13], . . .)
Some issues

- Do maximize the **likelihood** of the sample (preliminary results [Gybels et al., ICGI’14])
- Need to deal with **negative values**
- Some **tensor decompositions** allows one to restrict to HMM ($|\Sigma| > \#\text{states}$) [Anankumar et al., 2012]
- The **truncated SVD decomposition** implies an inadequate approximation
Outline

1. Context
2. Spectral method
3. Other approaches
Structured low rank approximation (SLRA) problem

- Let $\mathbf{p}_S = \langle p_S(w_1), \ldots, p_S(w_{n_p}) \rangle$
- $H_S = S(\mathbf{p}_S) = \sum_i S_i p_S(w_i)$ with $S_i \in \mathbb{R}^{m \times n}$ s.t. $S_i[u, v] = 1$ if $w_i = uv$, 0 otherwise.
- Example, $S = \{\epsilon, a, a\}$
  $$H_S = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \times p_S(\epsilon) + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \times p_S(a) = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 0 \end{bmatrix}$$

SLRA

$$\minimize_{\hat{\mathbf{p}}_S \in \mathbb{R}^{n_p}} \| \hat{\mathbf{p}}_S - \mathbf{p}_S \| \text{ subject to } rank(S(\hat{\mathbf{p}}_S)) \leq k.$$ 

⇒ solves the original problem but highly non convex, very efficient on small problems but intractable for $\sim 100 \times 100$ matrices [Markovsky, 2012] - ERC project
Matrix completion

Convex optimisation

\[ H' = \arg\min_{H \in \mathcal{H}} \ell(H, H_S) + \lambda \|H\| \]

- loss \( \ell \): **agreement** between \( H \) and \( H_S \)
- \( \| \cdot \| \): controls the **complexity/rank** of \( H \)
  - avoids ill-posedness matrix completion problem
  - \( \rightarrow \) simpler models

- \( \mathcal{H} \) space of Hankel matrices - **convex constraints**

- **Generalization bounds** (uniform stability and matrix completion) [Balle Mohri, NIPS 2012]

\[
S = \{(a, 1), (b, 3), (aa, 5), (ba, 2), \ldots\}
\]

\[
\begin{pmatrix}
\varepsilon & a & b & aa \\
\varepsilon & 0 & 1 & 3 & 5 \\
a & 1 & 5 & 0.3 & 3.8 \\
b & 3 & 2 & 0.2 & 0.7 \\
aa & 5 & 3.8 & 0 & 2.8 \\
\end{pmatrix}
\]
Perspectives

- **transfer learning/domain adaptation**

<table>
<thead>
<tr>
<th>Electronics</th>
<th>Video games</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ (1) Compact; easy to operate; very good picture quality; looks sharp!</td>
<td>(2) A very good game! It is action packed and full of excitement. I am very much hooked on this game.</td>
</tr>
<tr>
<td>✔️ (3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and sharp.</td>
<td>(4) Very realistic shooting action and good plots. We played this and were hooked.</td>
</tr>
<tr>
<td>❌ (5) It is also quite blurry in very dark settings. I will never buy HP again.</td>
<td>(6) It is so boring. I am extremely unhappy and will probably never buy UbiSoft again.</td>
</tr>
</tbody>
</table>

- Adapting the model to **changing data**

  Lifelong learning (ERC C. Lampert)
Conclusion

- **Efficient** and **simple** approach
- **Theoretical** guarantees
- Can be extended to other problems
- Tools and methods coming from **statistical learning** and **optimization**
- Lots of interesting perspectives, theoretical and practical