Learning Automata with Hankel Matrices

Borja Balle

[Disclaimer: Work done before joining Amazon]
Brief History of Automata Learning

- [1967] Gold: Regular languages are learnable in the limit
- [1987] Angluin: Regular languages are learnable from queries
- [1993] Pitt & Warmuth: PAC-learning DFA is NP-hard
- [1994] Kearns & Valiant: Cryptographic hardness
- [90’s, 00’s] Clark, Denis, de la Higuera, Oncina, others: Combinatorial methods meet statistics and linear algebra
Talk Outline

• Exact Learning
  – Hankel Trick for Deterministic Automata
  – Angluin’s L* Algorithm

• PAC Learning
  – Hankel Trick for Weighted Automata
  – Spectral Learning Algorithm

• Statistical Learning
  – Hankel Matrix Completion
The Hankel Matrix

\[
H \in \mathbb{R}^{\Sigma^* \times \Sigma^*}
\]

\[
p \cdot s = p' \cdot s' \Rightarrow H(p, s) = H(p', s')
\]

\[
f : \Sigma^* \rightarrow \mathbb{R}
\]

\[
H_f(p, s) = f(p \cdot s)
\]
Theorem (Myhill-Nerode ‘58)
The number of distinct rows of a binary Hankel matrix $H$ equals the minimal number of states of a DFA recognizing the language of $H$. 
From Hankel Matrices to DFA

\[
\begin{bmatrix}
\epsilon & a & b & aa & ab & ba & bb & \cdots \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & \cdots \\
a & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
b & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
aa & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
ab & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
ba & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
bb & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
\vdots & & & & & & & \\
aba & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
abb & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
\vdots & & & & & & & \\
\end{bmatrix}
\]

Diagram:

- States: \(\epsilon\), \(a\), \(ab\)
- Transitions:
  - \(\epsilon\) to \(a\)
  - \(a\) to \(ab\)
  - \(ab\) to \(\epsilon\)

\(\delta(\epsilon, a) = a\)
\(\delta(a, b) = ab\)
\(\delta(ab, b) = \epsilon\)
Closed and Consistent Finite Hankel Matrices

The DFA synthesis algorithm requires:
• Sets of prefixes $P$ and suffixes $S$
• Hankel block over $P' = P \cup P\Sigma$ and $S$
• Closed: $\text{rows}(P\Sigma) \subseteq \text{rows}(P)$
• Consistent: $\text{row}(p) = \text{row}(p') \Rightarrow \text{row}(p \cdot a) = \text{row}(p' \cdot a)$
Learning from Membership and Equivalence Queries

• Setup:
  – Two players, Teacher and Learner
  – Concept class $C$ of function from $X$ to $Y$ (known to Teacher and Learner)

• Protocol:
  – Teacher secretly chooses concept $c$ from $C$
  – Learner’s goal is to discover the secret concept $c$
  – Teacher answers two types of queries asked by Learner
    • Membership queries: what is the value of $c(x)$ for some $x$ picked by the Learner?
    • Equivalence queries: is $c$ equal to hypothesis $h$ from $C$ picked by the Learner?
      – If not, return counter-example $x$ where $h(x)$ and $c(x)$ differ

Angluin's $L^*$ Algorithm

1) Initialize $P = \{ \varepsilon \}$ and $S = \{ \varepsilon \}$
2) Maintain the Hankel block $H$ for $P' = P \cup P\Sigma$ and $S$ using membership queries
3) Repeat:
   - While $H$ is not closed and consistent:
     - If $H$ is not consistent add a distinguishing suffix to $S$
     - If $H$ is not closed add a new prefix from $P\Sigma$ to $P$
   - Construct a DFA $A$ from $H$ and ask an equivalence query
     - If yes, terminate
     - Otherwise, add all prefixes of counter-example $x$ to $P$

**Complexity**

- $O(n)$ EQs and $O(|\Sigma|^2 \cdot n \cdot L)$ MQs

Weighted Finite Automata (WFA)

Graphical Representation

- State $q_1$ with transitions $a, 1.2$ and $b, 2$.
- State $q_2$ with transitions $a, -2$, $b, 0$, $a, 3.2$, $b, 5$, $a, -1$, $b, -2$.

Algebraic Representation

- $A = \langle \alpha, \beta, \{A_a\}_{a \in \Sigma} \rangle$

$$\alpha = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} \quad A_a = \begin{bmatrix} 1.2 & -1 \\ -2 & 3.2 \end{bmatrix} \quad \beta = \begin{bmatrix} 1.2 \\ 0 \end{bmatrix} \quad A_b = \begin{bmatrix} 2 & -2 \\ 0 & 5 \end{bmatrix}$$

Functional Representation

$$A(x_1 \cdots x_t) = \alpha^\top A_{x_1} \cdots A_{x_t} \beta$$
Hankel Matrices and WFA

**Theorem (Fliess ’74)**
The rank of a *real* Hankel matrix $H$ equals the minimal number of states of a WFA recognizing the weighted language of $H$

$$A(p_1 \cdots p_t s_1 \cdots s_{t'}) = \alpha^T A_{p_1} \cdots A_{p_t} A_{s_1} \cdots A_{s_{t'}} \beta$$
From Hankel Matrices to WFA

\[ H_a(p, s) = A(pas) \]

\[ A(p_1 \cdots p_t a s_1 \cdots s_{t'}) = \alpha^\top A_{p_1} \cdots A_{p_t} A_a A_{s_1} \cdots A_{s_{t'}} \beta \]

\[
\begin{bmatrix}
    \ldots & \vdots & \vdots \\
    \vdots & \ddots & \vdots \\
    \vdots & \vdots & \ddots \\
    \vdots & \vdots & \vdots \\
    \vdots & \vdots & \vdots \\
    \vdots & \vdots & \vdots \\
\end{bmatrix}
\begin{bmatrix}
    \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot \\
\end{bmatrix}
\begin{bmatrix}
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
\end{bmatrix}
\]
WFA Reconstruction via Singular Value Decomposition

Input: Hankel $H'$ over $P' = P \cup P\Sigma$ and $S$, number of states $n$

1) Extract from $H'$ the matrix $H$ over $P$ and $S$
2) Compute the rank $n$ SVD $H = U D V^T$
3) For each symbol $a$:
   - Extract from $H'$ the matrix $H_a$ over $P$ and $S$
   - Compute $A_a = D^{-1}U^T H_a V$

Robustness Property $\|H' - \hat{H}'\| \leq \varepsilon \Rightarrow \|A_a - \hat{A}_a\| \leq O(\varepsilon)$

Probably Approximately Correct (PAC) Learning

• Fix a class $D$ of distributions over $X$
• Collect $m$ i.i.d. samples $Z = (x_1, ..., x_m)$ from some unknown distribution $d$ from $D$
• An algorithm that receives $Z$ and outputs a hypothesis $h$ is a PAC-learner for the class $D$ if:
  — Whenever $m > \text{poly}(|d|, 1/\varepsilon, \log 1/\delta)$, with probability at least $1 - \delta$ the hypothesis satisfies $\text{distance}(d,h) < \varepsilon$
• The algorithm is an efficient PAC-learner if it runs in poly-time

Estimating Hankel Matrices from Samples

**Sample**

\[
\begin{cases}
\text{aa, b, bab, a,} \\
\text{bbab, abb, babba, abbb,} \\
\text{ab, a, aabba, baa,} \\
\text{abbab, baba, bb, a}
\end{cases}
\]

**Concentration Bound**

\[\|H - \hat{H}\| \leq O\left(\frac{1}{\sqrt{m}}\right)\]

**Empirical Hankel Matrix**

\[
\begin{bmatrix}
\epsilon & a & b & aa & ab & \ldots \\
0 & 3 & 1 & 1 & 1 & 1 \\
3 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{bmatrix}
\]
Spectral PAC Learning of Stochastic WFA

• Algorithm:
  1. Estimate empirical Hankel matrix
  2. Use spectral WFA reconstruction

• Efficient PAC-learning:
  – **Running time**: linear in $m$, polynomial in $n$ and size of Hankel matrix
  – **Accuracy measure**: $L_1$ distance on all strings of length at most $L$
  – **Sample complexity**: $L^2 |\Sigma| n^{1/2} / \sigma^2 \varepsilon^2$
  – **Proof**: robustness + concentration + telescopic $L_1$ bound

---


Statistical Learning in the Non-realizable Setting

• Fix an unknown distribution \( d \) over \( X \times Y \) (inputs, outputs)
• Collect \( m \) i.i.d. samples \( Z = ((x_1, y_1), \ldots, (x_m, y_m)) \) from \( d \)
• Fix a hypothesis class \( F \) of functions from \( X \) to \( Y \)
• Find a hypothesis \( h \) from \( F \) that has good accuracy on \( Z \)

Empirical Risk Minimization

\[
\min_{h \in F} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i)
\]

• In such a way that it has good accuracy on future \( (x, y) \) from \( d \)

\[
\mathbb{E}_{(x, y) \sim d}[\ell(h(x), y)] \leq \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) + \text{complexity}(Z, F)
\]
Learning WFA via Hankel Matrix Completion

\[
\min_h P \left( \begin{array}{c}
1 \\
m \\
\end{array} \right) \leq \left( \begin{array}{c}
`p \\
h \\
p \\
x \\
1 \\
i \\
\end{array} \right) \leq \left( \begin{array}{c}
`p \\
h \\
p \\
x \\
1 \\
i \\
\end{array} \right),
\]

\[
\left( \begin{array}{c}
a \\
b \\
aa \\
ab \\
ba \\
bb \\
\end{array} \right) \leq \left( \begin{array}{c}
? \\
? \\
2 \\
1 \\
? \\
0 \\
\end{array} \right)
\]

Generalization Bounds for Learning WFA

• The generalization power of WFA can be controlled by:
  – Bounding the norm of the weights
  – Bounding the norm of the language (in a Banach space)
  – Bounding the norm of the Hankel matrix

\[
\mathbb{E}_{(x,y) \sim d}[\ell(A(x), y)] \leq \frac{1}{m} \sum_{i=1}^{m} \ell(A(x_i), y_i) + \tilde{O} \left( \frac{\|H_{A}\|^{*}}{m} + \frac{1}{\sqrt{m}} \right)
\]
Some Practical Applications

- **L* algorithm**: learn DFA of network protocol implementations and compare against specification to find bugs
  

- **Spectral algorithm**: use as initial point of gradient-based methods, increases speed and accuracy
  

- **Hankel completion**: sample-efficient sequence-to-sequence models outperforming CRFs in small alphabets
  
Want to Learn More?

• EMNLP’14 tutorial (slides, video, code)
  – Variations on spectral algorithm
  – Extensions to weighted tree automata
  – https://borjaballe.github.io/emnlp14-tutorial/

• Survey papers

• Implementations: Sp2Learn, LibLearn, libalf
Thanks!

Xavier Carreras  
Mehryar Mohri  
Prakash Panangaden

Joelle Pineau  
Doina Precup  
Ariadna Quattoni

- Guillaume Rabusseau
- Franco M. Luque
- Pierre-Luc Bacon
- Pascale Gourdeau
- Odalric-Ambrym Maillard
- Will Hamilton
- Lucas Langer
- Shay Cohen
- Amir Globerson
Learning Automata with Hankel Matrices

Borja Balle

[Disclaimer: Work done before joining Amazon]