Learning Automata with Hankel Matrices

Borja Balle

research cambridge

[*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
- [2009] Hsu-Kakade-Zhang & Bailly-Denis-Ralaivola: Spectral learning

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

$$H \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$$

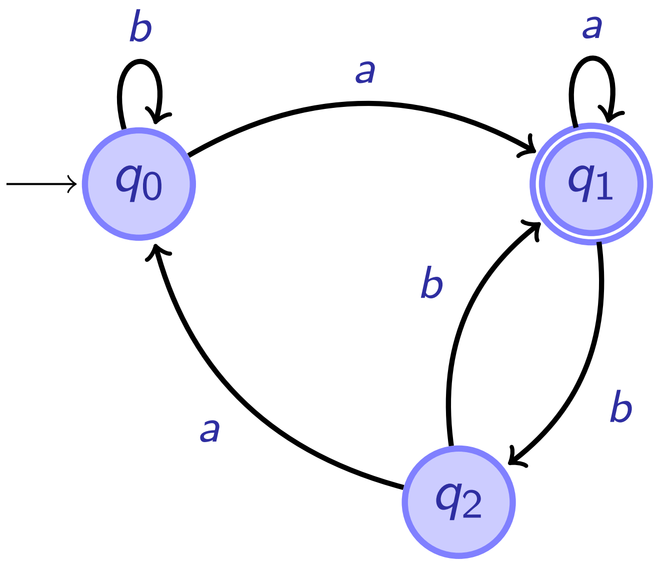
$$p \cdot s = p' \cdot s' \Rightarrow H(p, s) = H(p', s')$$

$$f : \Sigma^\star \to \mathbb{R}$$

$$H_f(p, s) = f(p \cdot s)$$

$$
\begin{array}{c}
 \\ \epsilon \\ a \\ b \\ aa \\ ab \\ ba \\ bb \\ \vdots \\ p \\ \vdots
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc}
\epsilon & a & b & aa & ab & ba & bb & \cdots & s & \cdots
\end{array} \\
\left[
\begin{array}{cccccccccc}
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & & \vdots & \\
 & & & & & & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & H(p, s) & \\
 & & & & & & & & &
\end{array}
\right]
\end{array}
$$

Hankel Matrices and DFA

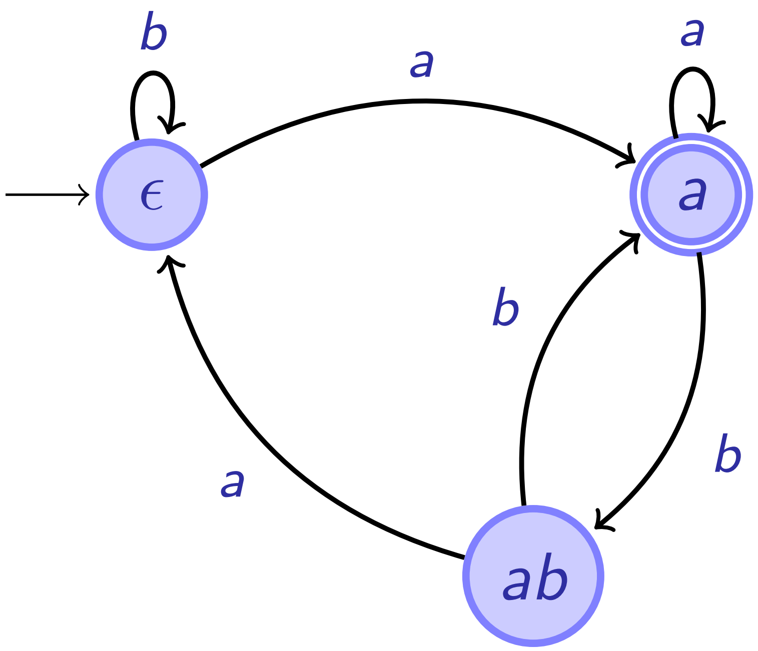


	ϵ	a	b	aa	ab	ba	bb	$\cdots$
ϵ	0	1	0	1	0	1	0	
a	1	1	0	1	0	0	1	
b	0	1	0	1	0	1	0	
aa	1	1	0	1	0	0	1	
ab	0	0	1	1	0	1	0	
ba	1	1	0	1	0	0	1	
bb	0	1	0	1	0	1	0	
$\vdots$								

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

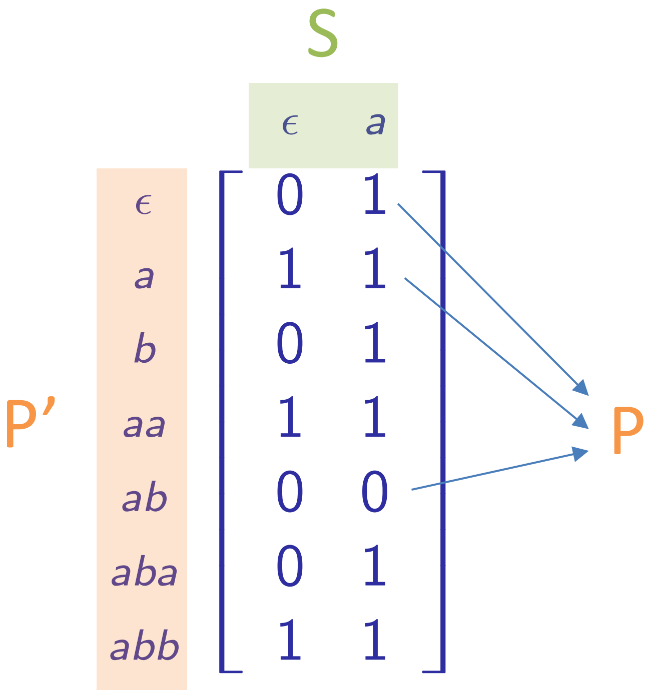
	ϵ	a	b	aa	ab	ba	bb	$\cdots$
ϵ	0	1	0	1	0	1	0	
a	1	1	0	1	0	0	1	
b	0	1	0	1	0	1	0	
aa	1	1	0	1	0	0	1	
ab	0	0	1	1	0	1	0	
ba	1	1	0	1	0	0	1	
bb	0	1	0	1	0	1	0	
$\vdots$								
aba	0	1	0	1	0	1	0	
abb	1	1	0	1	0	0	1	
$\vdots$								



Closed and Consistent Finite Hankel Matrices

The DFA synthesis algorithm requires:

- Sets of prefixes P and suffixes S
- Hankel block over P’ = P ∪ PΣ and S
- **Closed**: rows(PΣ) ⊆ rows(P)
- **Consistent**: row(p) = row(p’) ⇒ row(p·a) = row(p’·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, D. (1988). Queries and concept learning.

Angluin's L* Algorithm

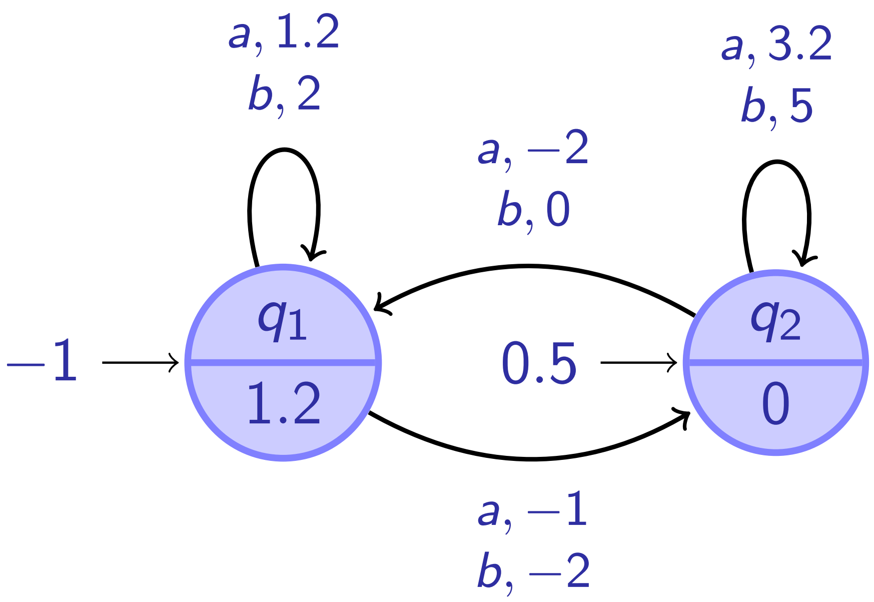
1) Initialize $P = \{\varepsilon\}$ and $S = \{\varepsilon\}$
2) Maintain the Hankel block H for $P' = P \cup P\Sigma$ and S using *membership queries*
3) Repeat:
   - While H is not closed and consistent:
     - If H is not consistent add a distinguishing suffix to S
     - If H is not closed add a 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 $O(|\Sigma|\, n^2\, L)$ MQs

Angluin, D. (1987). Learning regular sets from queries and counterexamples.

Weighted Finite Automata (WFA)

Graphical Representation



Algebraic Representation

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

$$\alpha = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} \qquad A_a = \begin{bmatrix} 1.2 & -1 \\ -2 & 3.2 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 1.2 \\ 0 \end{bmatrix} \qquad A_b = \begin{bmatrix} 2 & -2 \\ 0 & 5 \end{bmatrix}$$

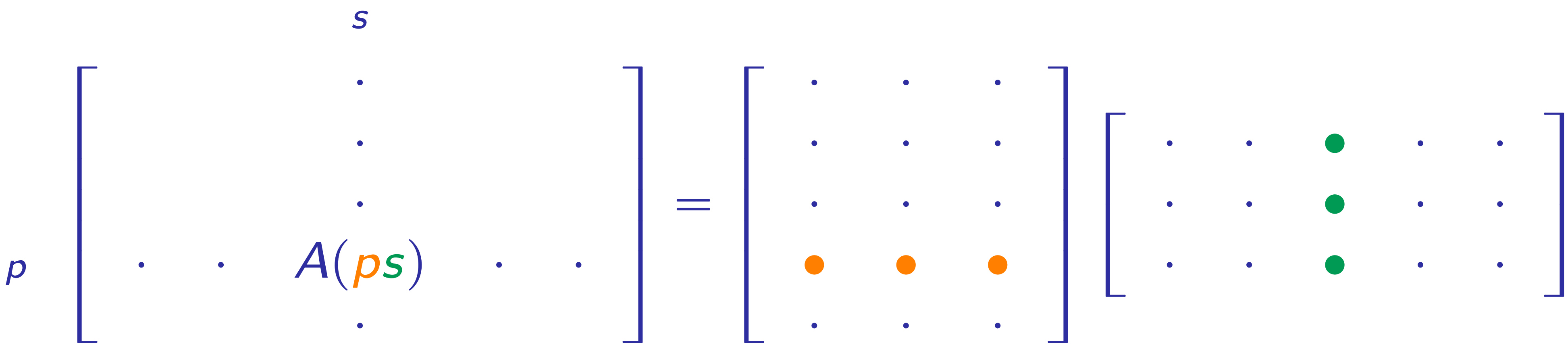
Functional Representation

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

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

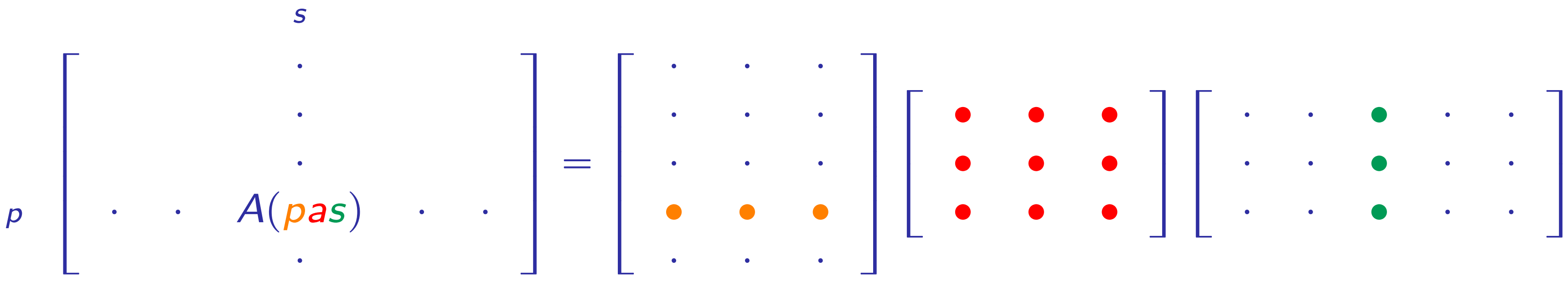
$$A(p_1 \cdots p_t s_1 \cdots s_{t'}) = \alpha^\top A_{p_1} \cdots A_{p_t} A_{s_1} \cdots A_{s_{t'}} \beta$$



From Hankel Matrices to WFA

$$H_a(p, s) = A(pas)$$

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



$$H = P\ S \qquad H_a = P\ A_a\ S \qquad A_a = P^+\ H_a\ S^+$$

WFA Reconstruction via Singular Value Decomposition

Input: Hankel H' over P' = P ∪ PΣ 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 $\|H' - \hat{H}'\| \leqslant \varepsilon \Rightarrow \|A_a - \hat{A}_a\| \leqslant O(\varepsilon)$

Balle, B., Carreras, X., Luque, F. M., & Quattoni, A. (2014). Spectral learning of weighted automata.

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 $m > poly(|d|, 1/\varepsilon, \log 1/\delta)$, with probability at least $1 - \delta$ the hypothesis satisfies $distance(d,h) < \varepsilon$
- The algorithm is an *efficient* PAC-learner if it runs in poly-time

Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R. E., & Sellie, L. (1994). On the learnability of discrete distributions.

Valiant, L. G. (1984). A theory of the learnable.

Estimating Hankel Matrices from Samples

Sample

$$\left\{ \begin{array}{c} aa,\ b,\ bab,\ a, \\ bbab,\ abb,\ babba,\ abbb, \\ ab,\ a,\ aabba,\ baa, \\ abbab,\ baba,\ bb,\ a \end{array} \right\}$$

Empirical Hankel Matrix

$$\begin{array}{c} \\ \epsilon \\ a \\ b \\ aa \\ ab \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \epsilon & a & b & aa & ab & \cdots \end{array} \\ \left[\begin{array}{cccccc} \frac{0}{16} & \frac{3}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \\ \frac{3}{16} & \frac{1}{16} & \frac{1}{16} & \frac{0}{16} & \frac{0}{16} & \\ \frac{1}{16} & \frac{0}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \\ \frac{1}{16} & \frac{0}{16} & \frac{0}{16} & \frac{0}{16} & \frac{0}{16} & \\ \frac{1}{16} & \frac{0}{16} & \frac{1}{16} & \frac{0}{16} & \frac{0}{16} & \\ & & & & & \end{array} \right] \end{array}$$

Concentration Bound

$$\|H - \hat{H}\| \leqslant O\left(\frac{1}{\sqrt{m}}\right)$$

research cambridge

Denis, F., Gybels, M., & Habrard, A. (2014, January). Dimension-free concentration bounds on hankel matrices for spectral learning.

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

Bailly, R., Denis, F., & Ralaivola, L. (2009). Grammatical inference as a principal component analysis problem.

Hsu, D., Kakade, S. M., & Zhang, T. (2009). A spectral algorithm for learning hidden markov models.

Statistical Learning in the Non-realizable Setting

- Fix an unknown distribution d over X x 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

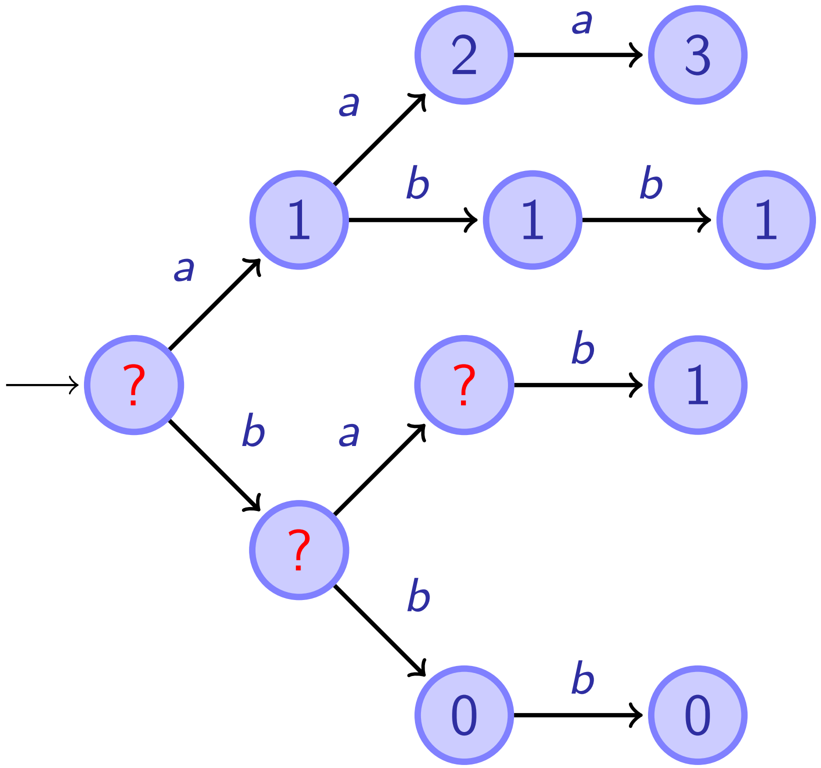
$$\min_{h \in F} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i)$$

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

$$\mathbb{E}_{(x,y)\sim d}[\ell(h(x), y)] \leqslant \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) + \text{complexity}(Z, F)$$

Learning WFA via Hankel Matrix Completion

$$\left\{ \begin{array}{cc} (bab,1) & (bbb,0) \\ (aaa,3) & (a,1) \\ (ab,1) & (aa,2) \\ (aba,2) & (bb,0) \end{array} \right\}$$



	ϵ	a	b
a	1	2	1
b	?	?	0
aa	2	3	?
ab	1	2	?
ba	?	?	1
bb	0	?	0

Balle, B., & Mohri, M. (2012). Spectral learning of general weighted automata via constrained matrix completion.

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

$$\mathbb{E}_{(x,y)\sim d}[\ell(A(x),y)] \leqslant \frac{1}{m}\sum_{i=1}^{m}\ell(A(x_i),y_i) + \tilde{O}\left(\frac{\|H_A\|_\star}{m} + \frac{1}{\sqrt{m}}\right)$$

Balle, B., & Mohri, M. (2017). Generalization Bounds for Learning Weighted Automata

Some Practical Applications

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

De Ruiter, J., & Poll, E. (2015). Protocol State Fuzzing of TLS Implementations.

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

Jiang, N., Kulesza, A., & Singh, S. P. (2016). Improving Predictive State Representations via Gradient Descent.

- Hankel completion: sample-efficient sequence-to-sequence models outperforming CRFs in small alphabets

Quattoni, A., Balle, B., Carreras Pérez, X., & Globerson, A. (2014). Spectral regularization for max-margin sequence tagging.

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
  - B. Balle and M. Mohri (2015). Learning Weighted Automata
  - M. R. Thon and H. Jaeger (2015). Links between multiplicity automata, observable operator models and predictive state representations
  - F. Vaandrager (2017). Model Learning
- Implementations: Sp2Learn, LibLearn, libalf

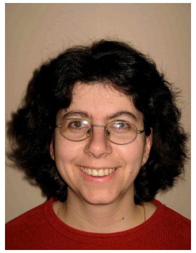
Thanks!

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

research cambridge

[*Disclaimer*: Work done before joining Amazon]