Learning Automata with Hankel Matrices

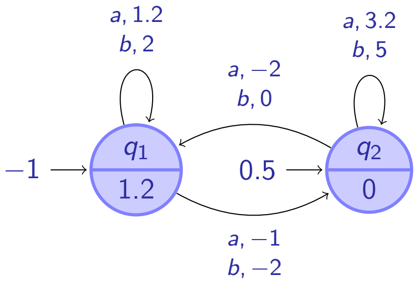
Borja Balle

Amazon Research Cambridge

Highlights — London, September 2017

Weighted Finite Automata (WFA) (over $\mathbb{R}$)

Graphical Representation



Algebraic Representation

$$A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}_{a \in \Sigma} \rangle$$

$$\boldsymbol{\alpha} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} \qquad \mathbf{A}_a = \begin{bmatrix} 1.2 & -1 \\ -2 & 3.2 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} 1.2 \\ 0 \end{bmatrix} \qquad \mathbf{A}_b = \begin{bmatrix} 2 & -2 \\ 0 & 5 \end{bmatrix}$$

Behavioral Representation

Each WFA A computes a function $A : \Sigma^\star \to \mathbb{R}$ given by $A(x_1 \cdots x_T) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_T} \boldsymbol{\beta}$

In This Talk...

- Describe a core algorithm common to many algorithms for learning weighted automata
- Explain the role this core plays in three learning problems in different setups
- Survey extensions to more complex models and some applications

Outline

1. From Hankel Matrices to Weighted Automata
2. From Data to Hankel Matrices
3. From Theory to Practice

Outline

1. From Hankel Matrices to Weighted Automata

2. From Data to Hankel Matrices

3. From Theory to Practice

Hankel Matrices and Fliess' Theorem

Given $f : \Sigma^\star \to \mathbb{R}$ define its Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ as

$$\mathbf{H}_f = \begin{array}{c} \\ \epsilon \\ a \\ b \\ \vdots \\ p \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} \epsilon & a & b & \cdots & s & \cdots \end{array} \\ \left[\begin{array}{cccccc} f(\epsilon) & f(a) & f(b) & & \vdots & \\ f(a) & f(aa) & f(ab) & & \vdots & \\ f(b) & f(ba) & f(bb) & & \vdots & \\ & & & & & \\ \cdots & \cdots & \cdots & & f(ps) & \\ & & & & & \end{array}\right] \end{array}$$

Theorem [Fli74]

1. The rank of $\mathbf{H}_f$ is finite if and only if f is computed by a WFA
2. The rank $\mathrm{rank}(f) = \mathrm{rank}(\mathbf{H}_f)$ equals the number of states of a minimal WFA computing f

The Structure of Hankel Matrices

$$A(p_1 \cdots p_T s_1 \cdots s_{T'}) = \alpha^\top \mathbf{A}_{p_1} \cdots \mathbf{A}_{p_T} \mathbf{A}_{s_1} \cdots \mathbf{A}_{s_{T'}} \beta$$

$$\mathbf{H} = \begin{array}{c c} & s \\ p & \left[\begin{array}{ccccc} & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ \cdot & \cdot & f(ps) & \cdot & \cdot \\ & & \cdot & & \end{array}\right] \end{array} = \left[\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot \end{array}\right] \left[\begin{array}{ccccc} \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \end{array}\right]$$

$$A(p_1 \cdots p_T a s_1 \cdots s_{T'}) = \alpha^\top \mathbf{A}_{p_1} \cdots \mathbf{A}_{p_T} \mathbf{A}_a \mathbf{A}_{s_1} \cdots \mathbf{A}_{s_{T'}} \beta$$

$$\mathbf{H}_a = \begin{array}{c c} & s \\ p & \left[\begin{array}{ccccc} & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ \cdot & \cdot & f(pas) & \cdot & \cdot \\ & & \cdot & & \end{array}\right] \end{array} = \left[\begin{array}{ccc} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot \end{array}\right] \left[\begin{array}{ccc} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{array}\right] \left[\begin{array}{ccccc} \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \end{array}\right]$$

Algebraically: Factorizing $\mathbf{H}$ lets us solve for $\mathbf{A}_a$

$$\mathbf{H} = \mathbf{P}\ \mathbf{S} \quad \Longrightarrow \quad \mathbf{H}_\sigma = \mathbf{P}\ \mathbf{A}_a\ \mathbf{S} \quad \Longrightarrow \quad \mathbf{A}_a = \mathbf{P}^+\ \mathbf{H}_a\ \mathbf{S}^+$$

SVD-based Reconstruction [HKZ09; Bal+14]

Inputs

- Desired number of states r
- Basis $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ with $\mathcal{P}, \mathcal{S} \subset \Sigma^\star$, $\epsilon \in \mathcal{P} \cap \mathcal{S}$
- Finite Hankel blocks indexed by prefixes and suffixes in $\mathcal{B}$:
  - $\mathbf{H}^{\mathcal{B}} \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}}$
  - $\mathbf{H}^{\mathcal{B}}_{\Sigma} = \{\mathbf{H}^{\mathcal{B}}_a \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}} : a \in \Sigma\}$

Algorithm: Spectral$(\mathbf{H}^{\mathcal{B}}, \mathbf{H}^{\mathcal{B}}_{\Sigma}, r)$

1. Compute the rank r SVD of $\mathbf{H}^{\mathcal{B}} \approx \mathbf{U}\mathbf{D}\mathbf{V}^\top$
2. Let $\mathbf{A}_a = \mathbf{D}^{-1}\mathbf{U}^\top \mathbf{H}_a \mathbf{V}$
3. Let $\boldsymbol{\alpha} = \mathbf{V}^\top \mathbf{H}^{\mathcal{B}}(\epsilon, -)$ and $\boldsymbol{\beta} = \mathbf{D}^{-1}\mathbf{U}^\top \mathbf{H}^{\mathcal{B}}(-, \epsilon)$
4. Return $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle$

Running time:

1. SVD takes $O(|\mathcal{P}||\mathcal{S}|r)$
2. Matrix multiplications take $O(|\Sigma||\mathcal{P}||\mathcal{S}|r)$

Properties of Spectral [HKZ09; Bal13; BM15a]

Consistency

- If $\mathcal{P}$ is prefix-closed, $\mathcal{S}$ is suffix-closed, and $r = \text{rank}(\mathbf{H}^{\mathcal{B}}) = \text{rank}([\mathbf{H}^{\mathcal{B}} | \mathbf{H}_{\Sigma}^{\mathcal{B}}])$
- Then $\forall p \in \mathcal{P}$, $\forall s \in \mathcal{S}$, $\forall a \in \Sigma$, the WFA $A = \text{Spectral}(\mathbf{H}^{\mathcal{B}}, \mathbf{H}_{\Sigma}^{\mathcal{B}}, r)$ satisfies $A(p \cdot s) = \mathbf{H}^{\mathcal{B}}(p, s)$ and $\tilde{A}(p \cdot a \cdot s) = \mathbf{H}_{a}^{\mathcal{B}}(p, s)$

Recovery

- If $\mathbf{H}^{\mathcal{B}}$ and $\mathbf{H}_{\Sigma}^{\mathcal{B}}$ are sub-blocks of $\mathbf{H}_f$ with $r = \text{rank}(f) = \text{rank}(\mathbf{H}^{\mathcal{B}})$
- Then the WFA $A = \text{Spectral}(\mathbf{H}^{\mathcal{B}}, \mathbf{H}_{\Sigma}^{\mathcal{B}}, r)$ satisfies $A \equiv f$

Robustness

- If $r = \text{rank}(\mathbf{H}^{\mathcal{B}}) = \text{rank}([\mathbf{H}^{\mathcal{B}} | \mathbf{H}_{\Sigma}^{\mathcal{B}}])$ and $\|\mathbf{H}^{\mathcal{B}} - \hat{\mathbf{H}}^{\mathcal{B}}\| \leqslant \varepsilon$ and $\|\mathbf{H}_{a}^{\mathcal{B}} - \hat{\mathbf{H}}_{a}^{\mathcal{B}}\| \leqslant \varepsilon$ for all $a \in \Sigma$
- Then $\langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle = \text{Spectral}(\mathbf{H}^{\mathcal{B}}, \mathbf{H}_{\Sigma}^{\mathcal{B}}, r)$ and $\langle \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \{\hat{\mathbf{A}}_a\} \rangle = \text{Spectral}(\hat{\mathbf{H}}^{\mathcal{B}}, \hat{\mathbf{H}}_{\Sigma}^{\mathcal{B}}, r)$ satisfy $\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|$, $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|$, $\|\mathbf{A}_a - \hat{\mathbf{A}}_a\| \leqslant \varepsilon$

Outline

1. From Hankel Matrices to Weighted Automata

2. From Data to Hankel Matrices

3. From Theory to Practice

Learning Models

1. Exact query learning: membership + equivalence queries [BV96; BBM06; BM15a]
2. Distributional PAC learning: samples from a stochastic WFA [HKZ09; BDR09; Bal+14]
3. Statistical learning: optimize output predictions wrt a loss function [BM12; BM15b]

Exact Learning of WFA with Queries

Setup:

- Unknown $f : \Sigma^\star \to \mathbb{R}$ with $\mathrm{rank}(f) = n$
- Membership oracle: $\mathrm{MQ}_f(x)$ returns $f(x)$ for any $x \in \Sigma^\star$
- Equivalence oracle: $\mathrm{EQ}_f(A)$ returns $\mathtt{true}$ if $f \equiv A$ and $(\mathtt{false}, z)$ if $f(z) \neq A(z)$

Algorithm:

1. Initialize $\mathcal{P} = \mathcal{S} = \{\epsilon\}$ and maintain $\mathcal{B} = (\mathcal{P}, \mathcal{S})$
2. Let $A = \mathrm{Spectral}(\mathbf{H}^{\mathcal{B}}, \mathbf{H}^{\mathcal{B}}_{\Sigma}, \mathrm{rank}(\mathbf{H}^{\mathcal{B}}))$
3. While $\mathrm{EQ}(A) = (\mathtt{false}, z)$
   3.1 Let $z = p \cdot a \cdot s$ with p the longest prefix of z in $\mathcal{P}$
   3.2 Let $\mathcal{S} = \mathcal{S} \cup \mathrm{suffixes}(s)$
   3.3 While $\exists p \in \mathcal{P}$ and $\exists a \in \Sigma$ such that $\mathbf{H}^{\mathcal{B}}_a(p, -) \notin \mathrm{rowspan}(\mathbf{H}^{\mathcal{B}})$, add $p \cdot a$ to $\mathcal{P}$
   3.4 Let $A = \mathrm{Spectral}(\mathbf{H}^{\mathcal{B}}, \mathbf{H}^{\mathcal{B}}_{\Sigma}, \mathrm{rank}(\mathbf{H}^{\mathcal{B}}))$

Analysis:

- At most $n + 1$ calls to EQ_f and $O(|\Sigma| n^2 L)$ calls to MQ_f, where $L = \max |z|$
- Can be improved to $O((|\Sigma| + \log L) n^2)$ calls to MQ_f; can reduce calls to EQ_f by increasing calls to MQ_f

PAC Learning Stochastic WFA

Setup:

- Unknown $f : \Sigma^\star \to \mathbb{R}$ with $\mathrm{rank}(f) = n$ defining probability distribution on $\Sigma^\star$
- Data: $x^{(1)}, \dots, x^{(m)}$ i.i.d. strings sampled from f
- Parameters: n and $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ such that $\mathrm{rank}(\mathbf{H}^{\mathcal{B}}) = n$ and $\epsilon \in \mathcal{P} \cap \mathcal{S}$

Algorithm:

1. Estimate Hankel matrices $\hat{\mathbf{H}}^{\mathcal{B}}$ and $\hat{\mathbf{H}}_a^{\mathcal{B}}$ for all $a \in \Sigma$ using empirical probabilities

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} 1[x^{(i)} = x]$$

2. Return $\hat{A} = \mathrm{Spectral}(\hat{\mathbf{H}}^{\mathcal{B}}, \hat{\mathbf{H}}_\Sigma^{\mathcal{B}}, n)$

Analysis:

- Running time is $O(|\mathcal{P} \cdot \mathcal{S}| m + |\Sigma||\mathcal{P}||\mathcal{S}| n)$
- With high probability $\sum_{|x| \leqslant L} |f(x) - \hat{A}(x)| = O\left(\frac{L^2 |\Sigma| \sqrt{n}}{\sigma_n(\mathbf{H}_f^{\mathcal{B}})^2 \sqrt{m}}\right)$

Statistical Learning of WFA

Setup:

- Unknown distribution $\mathcal{D}$ over $\Sigma^\star \times \mathbb{R}$
- Data: $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$ i.i.d. string-label pairs sampled from $\mathcal{D}$
- Parameters: n, convex loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, convex regularizer R, regularization parameter $\lambda > 0$, and $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ with $\epsilon \in \mathcal{P} \cap \mathcal{S}$

Algorithm:

1. Build $\mathcal{B}' = (\mathcal{P}', \mathcal{S})$ with $\mathcal{P}' = \mathcal{P} \cup \mathcal{P} \cdot \Sigma$
2. Find the Hankel matrix $\hat{\mathbf{H}}^{\mathcal{B}'}$ solving $\min_{\mathbf{H}} \frac{1}{m} \sum_{i=1}^m \ell(\mathbf{H}(x^{(i)}), y^{(i)}) + \lambda R(\mathbf{H})$
3. Return $\hat{A} = \mathrm{Spectral}(\hat{\mathbf{H}}^{\mathcal{B}}, \hat{\mathbf{H}}_{\Sigma}^{\mathcal{B}}, n)$, where $\hat{\mathbf{H}}^{\mathcal{B}}$ and $\hat{\mathbf{H}}_{\Sigma}^{\mathcal{B}}$ are submatrices of $\hat{\mathbf{H}}^{\mathcal{B}'}$

Analysis:

- Running time is polynomial in n, m, $|\Sigma|$, $|\mathcal{P}|$, and $|\mathcal{S}|$
- With high probability

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\hat{A}(x), y)] \leqslant \frac{1}{m} \sum_{i=1}^m \ell(\hat{A}(x^{(i)}), y^{(i)}) + O\left(\frac{1}{\sqrt{m}}\right)$$

Outline

1. From Hankel Matrices to Weighted Automata
2. From Data to Hankel Matrices
3. From Theory to Practice

Extensions

1. More complex models
   - Transducers and taggers [BQC11; Qua+14]
   - Grammars and tree automata [Luq+12; Bal+14; RBC16]
   - Reactive models [BBP15; LBP16; BM17a]
2. More realistic setups
   - Multiple related tasks [RBP17]
   - Timing data [BBP15; LBP16]
   - Single trajectory [BM17a]
   - Probabilistic models [BHP14]
3. Deeper theory
   - Convex relaxations [BQC12]
   - Generalization bounds [BM15b; BM17b]
   - Approximate minimisation [BPP15]
   - Bisimulation metrics [BGP17]

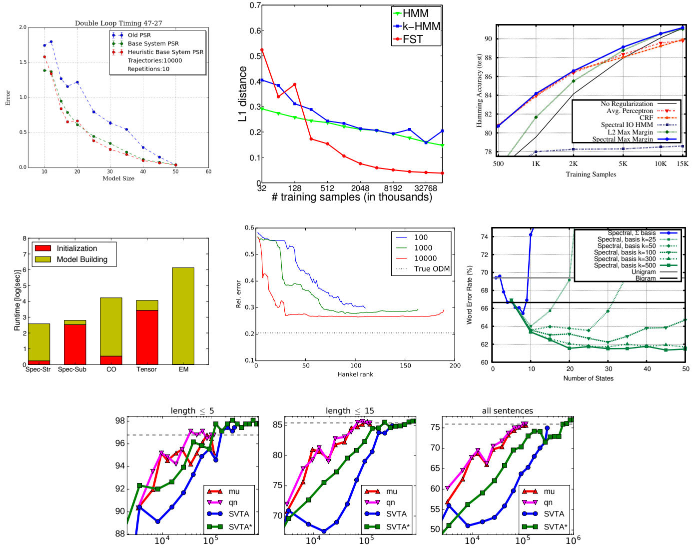
And It Works Too!

Spectral methods are competitive against traditional methods:

- Expectation maximization
- Conditional random fields
- Tensor decompositions

In a variety of problems:

- Sequence tagging
- Constituency and dependency parsing
- Timing and geometry learning
- POS-level language modelling



Open Problems and Current Trends

- Optimal selection of $\mathcal{P}$ and $\mathcal{S}$ from data
- Scalable convex optimization over sets of Hankel matrices
- Constraining the output WFA (eg. probabilistic automata)
- Relations between learning and approximate minimisation
- How much of this can be extended to WFA over semi-rings?
- Spectral methods for initializing non-convex gradient-based learning algorithms

Conclusion

Take home points

- A single building block based on SVD of Hankel matrices
- Implementation only requires linear algebra
- Analysis involves linear algebra, probability, convex optimization
- Can be made practical for a variety of models and applications

Want to know more?

- EMNLP'14 tutorial (with slides, video, and code)
  `https://borjaballe.github.io/emnlp14-tutorial/`
- Survey papers [BM15a; TJ15]
- Python toolkit Sp2Learn [Arr+16]
- Neighbouring literature: Predictive state representations (PSR) [LSS02] and Observable operator models (OOM) [Jae00]

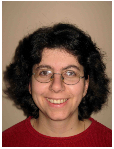
Thanks To All My Collaborators!

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

Bibliography I

[Arr+16] D. Arrivault, D. Benielli, F. Denis, and R. Eyraud. "Sp2Learn: A Toolbox for the Spectral Learning of Weighted Automata". In: *ICGI*. 2016.

[Bal+14] B. Balle, X. Carreras, F.M. Luque, and A. Quattoni. "Spectral learning of weighted automata: A forward-backward perspective". In: *Machine Learning* (2014).

[Bal13] B. Balle. "Learning Finite-State Machines: Algorithmic and Statistical Aspects". PhD thesis. Universitat Politècnica de Catalunya, 2013.

[BBM06] L. Bisht, N. H. Bshouty, and H. Mazzawi. "On Optimal Learning Algorithms for Multiplicity Automata". In: *COLT*. 2006.

[BBP15] P.-L. Bacon, B. Balle, and D. Precup. "Learning and Planning with Timing Information in Markov Decision Processes". In: *UAI*. 2015.

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

Bibliography II

[BGP17] B. Balle, P. Gourdeau, and P. Panangaden. "Bisimulation Metrics for Weighted Automata". In: *ICALP*. 2017.

[BHP14] B. Balle, W. L. Hamilton, and J. Pineau. "Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison". In: *ICML*. 2014.

[BM12] B. Balle and M. Mohri. "Spectral learning of general weighted automata via constrained matrix completion". In: *NIPS*. 2012.

[BM15a] B. Balle and M. Mohri. "Learning Weighted Automata (invited paper)". In: *CAI*. 2015.

[BM15b] B. Balle and M. Mohri. "On the Rademacher complexity of weighted automata". In: *ALT*. 2015.

[BM17a] B. Balle and O.-A. Maillard. "Spectral Learning from a Single Trajectory under Finite-State Policies". In: *ICML*. 2017.

Bibliography III

[BM17b] B. Balle and M. Mohri. "Generalization Bounds for Learning Weighted Automata". In: *Theor. Comput. Sci. (to appear)* (2017).

[BPP15] B. Balle, P. Panangaden, and D. Precup. "A Canonical Form for Weighted Automata and Applications to Approximate Minimization". In: *LICS*. 2015.

[BQC11] B. Balle, A. Quattoni, and X. Carreras. "A spectral learning algorithm for finite state transducers". In: *ECML-PKDD*. 2011.

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

[BV96] F. Bergadano and S. Varricchio. "Learning behaviors of automata from multiplicity and equivalence queries". In: *SIAM Journal on Computing* (1996).

[Fli74] M. Fliess. "Matrices de Hankel". In: *Journal de Mathématiques Pures et Appliquées* (1974).

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

Bibliography IV

[Jae00] H. Jaeger. "Observable operator models for discrete stochastic time series". In: *Neural Computation* (2000).

[LBP16] L. Langer, B. Balle, and D. Precup. "Learning Multi-Step Predictive State Representations". In: *IJCAI*. 2016.

[LSS02] M. Littman, R. S. Sutton, and S. Singh. "Predictive representations of state". In: *NIPS*. 2002.

[Luq+12] F.M. Luque, A. Quattoni, B. Balle, and X. Carreras. "Spectral learning in non-deterministic dependency parsing". In: *EACL*. 2012.

[Qua+14] A. Quattoni, B. Balle, X. Carreras, and A. Globerson. "Spectral Regularization for Max-Margin Sequence Tagging". In: *ICML*. 2014.

[RBC16] G. Rabusseau, B. Balle, and S. B. Cohen. "Low-Rank Approximation of Weighted Tree Automata". In: *AISTATS*. 2016.

[RBP17] G. Rabusseau, B. Balle, and J. Pineau. "Multitask Spectral Learning of Weighted Automata". In: *NIPS*. 2017.

Bibliography V

[TJ15] M. R. Thon and H. Jaeger. "Links between multiplicity automata, observable operator models and predictive state representations: a unified learning framework". In: *Journal of Machine Learning Research* (2015).

Learning Automata with Hankel Matrices

Borja Balle

Amazon Research Cambridge

Highlights — London, September 2017