Spectral Learning of General Weighted Automata via Constrained Matrix Completion

Borja Balle[1] Mehryar Mohri[2]

[1]Universitat Politècnica de Catalunya

[2]Courant Institute of Mathematical Sciences and Google Research

NIPS 2012, Lake Tahoe

The Problem – Regression over Strings

Data: i.i.d. sample S with strings $+$ real labels from distribution $\mathcal{D}$

$$S = (abbca, 3.4)\ (baa, 0.6)\ (ccaaaabba, -2.9)\ (abba, 1.1)\ \ldots$$

Goal: Learn a regressor $\hat{f} : \Sigma^\star \to \mathbb{R}$ with small generalization error

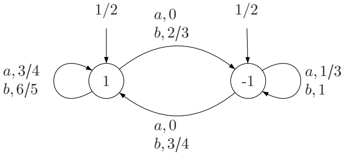
$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(\hat{f}(x), y)\right]$$

Examples:

- *Reward modeling* in reinforcement learning
- *Biological measurement* as a function of DNA/AA sequence
- Learn from *expert labeling* in natural language processing

Hypothesis Class: Weighted Finite Automata (WFA)

Graphical representation



Algebraic representation

$$\boldsymbol{\alpha}^\top = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix} \qquad \boldsymbol{\beta}^\top = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

$$\mathbf{A}_a = \begin{bmatrix} 3/4 & 0 \\ 0 & 1/3 \end{bmatrix} \qquad \mathbf{A}_b = \begin{bmatrix} 6/5 & 2/3 \\ 3/4 & 1 \end{bmatrix}$$

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

$$f(x_1 \cdots x_t) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\beta}$$

$(n^2|\Sigma| + 2n)$ parameters

Why WFA?

- Expressive well-studied class [DKV09]
- Rich family of algorithms [Mohri09]
  (weighted minimization, determinization, ϵ-removal)
- Widely used in applications [MPR08,AK09,BGC09,KM09]
  (speech recognition, image processing, OCR, system testing)

Overview

Our Result: A supervised learning algorithm for WFA that combines spectral learning and matrix completion

In the rest of the talk I will...

1. Recall the spectral method in a nutshell
2. Describe a family of learning algorithms
3. Give a generalization bound

Spectral Learning in a Nutshell

(Workshop on Friday!)

Key Ideas:

- Matrix of observables $\mathbf{M}$ contains sufficient information
- SVD decomposition of $\mathbf{M}$ is used to recover model
- Computation is noise tolerant

General Scheme:

$$\begin{tikzcd}
\mathbf{f} \arrow[r, mapsto, "\text{definition}"] \arrow[rrd, mapsto, bend left, "\text{sampling}"] & \mathbf{M} & \\
\wr \arrow[r, phantom, "\underset{\text{THEORY}}{\Longleftarrow}"] & \wr & \text{Sample} \arrow[dl, mapsto, "\text{summary}"] \\
\hat{\mathbf{f}} & \hat{\mathbf{M}} \arrow[l, mapsto, "\text{learning}"] &
\end{tikzcd}$$

Spectral Learning by...

Has been applied to many models:

- Sequential models
  [HKZ09,BDR09,SBG10,BSG10,BQC11,Bailly11,BQC12]
- Tree-like structures
  [BHD10,PSX11,ACHKSZ11,LQBC12,CSCFU12,DRCFU12]
- Other graphical models
  [SBSGS10,AFHKL12,AHK12,PSITX12]

In the particular case of WFA $\mathbf{M}$ is a well-known matrix...

The Hankel Matrix

Definition: Hankel matrix $\mathbf{H}_f$ of a function $f : \Sigma^\star \to \mathbb{R}$ is such that

- rows are indexed by *prefixes* $u \in \mathcal{P}$
- columns are indexes by *suffixes* $v \in \mathcal{S}$
- entries are evaluations $\mathbf{H}_f(u, v) = f(uv)$

Example: $\Sigma = \{a, b\}$ and $f(x) = \#$ of a's in x

$$\mathcal{P} = \{a, b, aa\} \text{ (rows)} \qquad \mathcal{S} = \{\epsilon, a, b\} \text{ (columns)}$$

$$\mathbf{H}_f = \begin{array}{c} \\ a \\ b \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} \epsilon & a & b \end{array} \\ \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 2 \end{bmatrix} \end{array}$$

Note: Entrywise redundancies
$w = u_1 v_1 = u_2 v_2 \Rightarrow \mathbf{H}_f(u_1, v_1) = \mathbf{H}_f(u_2, v_2) = f(w)$

Spectral Learning for WFA

In the particular case of WFA

- Hankel matrices play the role of $\mathbf{M}$
- Redundancies (entrywise & rank) are relevant
- f computed by WFA with $\mathrm{rank}(\mathbf{H}_f)$ states (when $\mathcal{P}$, $\mathcal{S}$ big enough)

Example: $\Sigma = \{a, b\}$ and $f(x) = \#$ of a's in x

$\mathcal{P} = \{a, b, aa\}$ (rows)
$\mathcal{S} = \{\epsilon, a, b\}$ (columns)
$\mathrm{rank}(\mathbf{H}_f) = 2$

$$\mathbf{H}_f = \begin{array}{c} \\ a \\ b \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} \epsilon & a & b \end{array} \\ \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 2 & 3 & 2 \end{bmatrix} \end{array}$$

But: how do you obtain a Hankel matrix in the regression setting?

Missing Entries

In usual applications...

entries in $\mathbf{H}_f$ are empirical counts, e.g. $f(x) = \Pr[x]$

$$\left\{ \begin{array}{c} \text{aa, b, bab, a,} \\ \text{b, a, ab, aa,} \\ \text{ba, b, aa, a,} \\ \text{aa, bab, b, aa} \end{array} \right\} \longrightarrow \begin{array}{c} \\ \epsilon \\ a \\ b \\ ba \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \left[\begin{array}{cc} .19 & .25 \\ .31 & .06 \\ .06 & .00 \\ .00 & .13 \end{array} \right] \end{array}$$

But in this case...

entries in $\mathbf{H}_f$ are labels observed in the sample

$$\left\{ \begin{array}{c} (\text{bab},1) \\ (\text{bbb},0) \\ (\text{aaa},3) \\ (\text{a},1) \\ (\text{ab},1) \\ (\text{aa},2) \\ (\text{aba},2) \\ (\text{bb},0) \end{array} \right\} \longrightarrow \begin{array}{c} \\ a \\ b \\ aa \\ ab \\ ba \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} \epsilon & a & b \end{array} \\ \left[\begin{array}{ccc} 1 & 2 & 1 \\ * & * & 0 \\ 2 & 3 & * \\ 1 & 2 & * \\ * & * & 1 \\ 0 & * & 0 \end{array} \right] \end{array}$$

Constrained Matrix Completion

Solution: Apply *matrix completion* to $\hat{\mathrm{H}}_{\mathrm{f}}$
[CR09,CP10,CT10,FSSS11,Recht11,FS11,NW12]

But: Constrain completed matrix to be *Hankel*

Algorithm: Use convex optimization

$$\hat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{H} \in \mathbb{H}} \ \ell(\mathbf{H}; S) + \lambda \cdot R(\mathbf{H})$$

- Loss ℓ controls agreement of $\mathbf{H}$ with sample
- Regularizer R controls complexity of $\mathbf{H}$ (e.g. schatten norm)
- $\mathbf{H} \in \mathbb{H}$ imposes *convex constains* (equalities between entries)

(Double) Role of Regularization:

- Solve ill-posedness of matrix completion problem
- Less complex $\hat{\mathbf{H}}$ will lead to simpler WFA

A Family of Algorithms

$$\left\{\begin{array}{c}(\text{bab},1)\\(\text{baab},2)\\(\text{bbb},0)\\(\text{aaa},3)\\(\text{a},1)\\(\text{ab},1)\\\vdots\end{array}\right\} \xrightarrow[\text{completion}]{\text{matrix}} \begin{array}{c|ccc} & \epsilon & a & b \\ \hline a & 1 & 2 & 1 \\ b & 0.1 & 0.7 & 0 \\ aa & 2.0 & 3 & 2.0 \\ ab & 1 & 2 & 0.9 \\ ba & 0.7 & 0.9 & 1 \\ bb & 0 & 1 & 0 \end{array} \xrightarrow[\text{learning}]{\text{spectral}} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\}\rangle$$

Family of algorithms parametrized by:

- Choice of rows and columns in $\mathbf{H}$
- A constrained matrix completion algorithm
- Regularization parameters

Question: Can these algorithms provably succeed?

Generalization Bound

Hypotheses:

- Reasonable assumptions on distribution $(x, y) \sim \mathcal{D}$
- Completion loss $\ell(\mathbf{H}; S) = \sum_{(x,y)\in S} |f_{\mathbf{H}}(x) - y|$
- Completion regularizer $R(\mathbf{H}) = \|\mathbf{H}\|_F^2$

Theorem: with high probability over $S \sim \mathcal{D}^m$, the output f_S of the algorithm satisfies:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[|f_S(x) - y|\right] \leqslant \hat{\mathbb{E}}_S[|f_S(x) - y|] + O\left(\frac{\ln m}{m^{1/3}}\right)$$

Proof: joint *stability analysis* of matrix completion and spectral learning

Want to Know More?

Poster T47

Spectral Learning of General Weighted Automata via Constrained Matrix Completion

Borja Balle[1] Mehryar Mohri[2]

[1]Universitat Politècnica de Catalunya

[2]Courant Institute of Mathematical Sciences and Google Research

NIPS 2012, Lake Tahoe