Singular Value Automata and Approximate Minimization

Borja Balle

Amazon Research Cambridge[1]

Weighted Automata: Theory and Applications — May 2018

[1] Based on work completed before joining Amazon

What Is This About?

Analytic Automata Theory

More prosaically:

- The use of tools from mathematical analysis to study questions in automata theory, specifically questions related to approximation and learning
- Based on joint work with: X. Carreras, P. Gourdeau, M. Mohri, P. Panangaden, D. Precup, G. Rabusseau, A. Quattoni
- Key references: **[Bal13, BPP17]**

Keep It Real!

$$\mathbb{R}$$

More precisely:

- Everything works for complex numbers
- Some things work for arbitrary fields
- Virtually nothing works for general semi-rings

Outline

Outline

1. Weighted Languages, Weighted Automata, and Hankel Matrices
2. Perturbation Bounds Between Representations
3. Singular Value Automata: Definition
4. Singular Value Automata: Computation
5. Approximate Minimization via SVA Truncation
6. Concluding Remarks

The Big Picture

$$
\begin{tikzcd}
& \text{Automaton} \arrow[dl] & \\
\text{Language} \arrow[rr] & & \text{Hankel} \arrow[ul]
\end{tikzcd}
$$

Weighted Languages

research

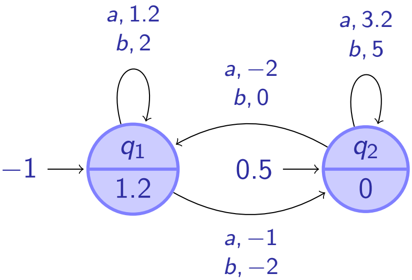
$$f : \Sigma^\star \to \mathbb{R} \quad , \qquad f \in \mathbb{R}^{\Sigma^\star}$$

Notation

- Finite alphabet Σ
- Free monoid $\Sigma^\star$
- Empty string ϵ
- String length $|x|$
- String concatenation $xy = x \cdot y$

Weighted Finite Automata (WFA)

Graphical Representation



Algebraic Representation

$$\boldsymbol{\alpha} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} 1.2 \\ 0 \end{bmatrix}$$

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

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

Weighted Finite Automaton

A WFA A with $n = |A|$ states is a tuple $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\}_{\sigma \in \Sigma} \rangle$ where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ and $\boldsymbol{A}_\sigma \in \mathbb{R}^{n \times n}$

Language of a WFA

With every WFA $A = \langle \alpha, \beta, \{\mathbf{A}_\sigma\} \rangle$ with n states we associate a weighted language $f_A : \Sigma^\star \to \mathbb{R}$ given by

$$f_A(x_1 \cdots x_T) = \sum_{q_0, q_1, \ldots, q_T \in [n]} \alpha(q_0) \left(\prod_{t=1}^{T} \mathbf{A}_{x_t}(q_{t-1}, q_t) \right) \beta(q_T)$$

$$= \alpha^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_T} \beta = \alpha^\top \mathbf{A}_x \beta$$

Recognizable/Rational Languages

A weighted language $f : \Sigma^\star \to \mathbb{R}$ is recognizable/rational if there exists a WFA A such that $f = f_A$. The smallest number of states of such a WFA is $\mathrm{rank}(f)$. A WFA A is minimal if $|A| = \mathrm{rank}(f_A)$.

Observation: The minimal A is not unique. Take any invertible matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, then

$$\alpha^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_T} \beta = (\alpha^\top \mathbf{Q})(\mathbf{Q}^{-1} \mathbf{A}_{x_1} \mathbf{Q}) \cdots (\mathbf{Q}^{-1} \mathbf{A}_{x_T} \mathbf{Q})(\mathbf{Q}^{-1} \beta)$$

Hankel Matrices

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

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

Fliess–Kronecker Theorem [Fli74]

The rank of $\mathbf{H}_f$ is finite if and only if f is rational, in which case $\mathrm{rank}(\mathbf{H}_f) = \mathrm{rank}(f)$

Structure of Low-Rank Hankel Matrices

$$
\mathbf{H}_{f_A} \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star} \qquad \mathbf{P}_A \in \mathbb{R}^{\Sigma^\star \times n} \qquad \mathbf{S}_A \in \mathbb{R}^{n \times \Sigma^\star}
$$

$$
\begin{array}{cc}
 & s \\
p & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & \bullet & \cdots & \cdots \\ & & \vdots & & \end{bmatrix}
\end{array}
=
\begin{array}{cc}
 & \\
 & \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot \end{bmatrix} \\
p &
\end{array}
\begin{array}{c}
s \\
\begin{bmatrix} \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \end{bmatrix}
\end{array}
$$

$$
f_A(p_1 \cdots p_T \cdot s_1 \cdots s_{T'}) = \underbrace{\boldsymbol{\alpha}^\top \mathbf{A}_{p_1} \cdots \mathbf{A}_{p_T}}_{\boldsymbol{\alpha}_A(p)} \; \underbrace{\mathbf{A}_{s_1} \cdots \mathbf{A}_{s_{T'}} \boldsymbol{\beta}}_{\boldsymbol{\beta}_A(s)}
$$

Note: We call $\mathbf{H}_f = \mathbf{P}_A \mathbf{S}_A$ the *forward-backward factorization* induced by A

Structure of Shifted Hankel Matrices

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

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

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

$$\mathbf{H}_\sigma = \begin{array}{cc} & s \\ p & \left[\begin{array}{ccccc} & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ \cdot & \cdot & f(pas) & \cdot & \cdot \\ & & \cdot & & \end{array}\right] \end{array} = \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \bullet & \bullet & \bullet \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot \end{bmatrix}$$

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

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

Aside: Moore–Penrose Pseudo-inverse

For any $\mathbf{M} \in \mathbb{R}^{n\times m}$ there exists a unique *pseudo-inverse* $\mathbf{M}^+ \in \mathbb{R}^{m\times n}$ satisfying:

- $\mathbf{M}\mathbf{M}^+\mathbf{M} = \mathbf{M}$, $\mathbf{M}^+\mathbf{M}\mathbf{M}^+ = \mathbf{M}^+$, and $\mathbf{M}^+\mathbf{M}$ and $\mathbf{M}\mathbf{M}^+$ are symmetric
- If $\mathrm{rank}(\mathbf{M}) = n$ then $\mathbf{M}\mathbf{M}^+ = \mathbf{I}$, and if $\mathrm{rank}(\mathbf{M}) = m$ then $\mathbf{M}^+\mathbf{M} = \mathbf{I}$
- If $\mathbf{M}$ is square and invertible then $\mathbf{M}^+ = \mathbf{M}^{-1}$

Given a system of linear equations $\mathbf{M}\mathbf{u} = \mathbf{v}$, the following is satisfied:

$$\mathbf{M}^+\mathbf{v} = \operatorname*{argmin}_{\mathbf{u}\in \operatorname{argmin} \|\mathbf{M}\mathbf{u}-\mathbf{v}\|_2} \|\mathbf{u}\|_2 \ .$$

In particular:

- If the system is completely determined, $\mathbf{M}^+\mathbf{v}$ solves the system
- If the system is underdetermined, $\mathbf{M}^+\mathbf{v}$ is the solution with smallest norm
- If the system is overdetermined, $\mathbf{M}^+\mathbf{v}$ is the minimum norm solution to the least-squares problem $\min \|\mathbf{M}\mathbf{u} - \mathbf{v}\|_2$

From Finite Hankel Matrix to WFA

Suppose $f : \Sigma^\star \to \mathbb{R}$ has rank n and $\varepsilon \in \mathcal{P}, \mathcal{S} \subset \Sigma^\star$ are such that the sub-block $\mathbf{H} \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}}$ of $\mathbf{H}_f$ satisfies $\mathrm{rank}(\mathbf{H}) = n$.

Let $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ be obtained as follows:

1. Compute a rank factorization $\mathbf{H} = \mathbf{PS}$; i.e. $\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{S}) = \mathrm{rank}(\mathbf{H})$
2. Let $\boldsymbol{\alpha}^\top$ (resp. $\boldsymbol{\beta}$) be the ϵ-row of $\mathbf{P}$ (resp. ϵ-column of $\mathbf{S}$)
3. Let $\mathbf{A}_\sigma = \mathbf{P}^+ \mathbf{H}_\sigma \mathbf{S}^+$, where $\mathbf{H}_\sigma \in \mathbb{R}^{\mathcal{P} \cdot \sigma \times \mathcal{S}}$ is a sub-block of $\mathbf{H}_f$

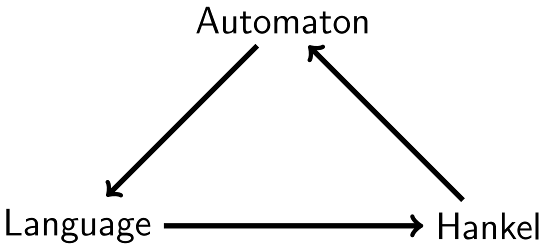
Claim The resulting WFA computes f and is minimal

Proof

- Suppose $\tilde{A} = \langle \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \{\tilde{\mathbf{A}}_\sigma\} \rangle$ is a minimal WFA for f.
- It suffices to show there exists an invertible $\mathbf{Q} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{\alpha}^\top = \tilde{\boldsymbol{\alpha}}^\top \mathbf{Q}$, $\mathbf{A}_\sigma = \mathbf{Q}^{-1} \tilde{\mathbf{A}}_\sigma \mathbf{Q}$ and $\boldsymbol{\beta} = \mathbf{Q}^{-1} \tilde{\boldsymbol{\beta}}$.
- By minimality $\tilde{A}$ induces a rank factorization $\mathbf{H} = \tilde{\mathbf{P}}\tilde{\mathbf{S}}$ and also $\mathbf{H}_\sigma = \tilde{\mathbf{P}}\tilde{\mathbf{A}}_\sigma\tilde{\mathbf{S}}$.
- Since $\mathbf{A}_\sigma = \mathbf{P}^+ \mathbf{H}_\sigma \mathbf{S}^+ = \mathbf{P}^+ \tilde{\mathbf{P}} \tilde{\mathbf{A}}_\sigma \tilde{\mathbf{S}} \mathbf{S}^+$, take $\mathbf{Q} = \tilde{\mathbf{S}}\mathbf{S}^+$.
- Check $\mathbf{Q}^{-1} = \mathbf{P}^+ \tilde{\mathbf{P}}$ since $\mathbf{P}^+ \tilde{\mathbf{P}} \tilde{\mathbf{S}} \mathbf{S}^+ = \mathbf{P}^+ \mathbf{H} \mathbf{S}^+ = \mathbf{P}^+ \mathbf{P} \mathbf{S} \mathbf{S}^+ = \mathbf{I}$.

Outline

The Big Picture



Norms on WFA

research

Weighted Finite Automaton

A WFA with n states is a tuple $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\}_{\sigma \in \Sigma} \rangle$ where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ and $\mathbf{A}_\sigma \in \mathbb{R}^{n \times n}$

Let $p, q \in [1, \infty]$ be Hölder conjugate $\frac{1}{p} + \frac{1}{q} = 1$.

The (p, q)-norm of a WFA A is given by

$$\|A\|_{p,q} = \max \left\{ \|\boldsymbol{\alpha}\|_p, \|\boldsymbol{\beta}\|_q, \max_{\sigma \in \Sigma} \|\mathbf{A}_\sigma\|_q \right\} \quad ,$$

where $\|\mathbf{A}_\sigma\|_q = \sup_{\|\mathbf{v}\|_q \leqslant 1} \|\mathbf{A}_\sigma \mathbf{v}\|_q$ is the q-induced norm.

Example For probabilistic automata $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ with $\boldsymbol{\alpha}$ probability distribution, $\boldsymbol{\beta}$ acceptance probabilities, $\mathbf{A}_\sigma$ row (sub-)stochastic matrices we have $\|A\|_{1,\infty} = 1$

Perturbation Bounds: Automaton→Language [Bal13]

research

Suppose $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\}\rangle$ and $A' = \langle \boldsymbol{\alpha}', \boldsymbol{\beta}', \{\mathbf{A}'_\sigma\}\rangle$ are WFA with n states satisfying $\|A\|_{p,q} \leqslant \rho$, $\|A'\|_{p,q} \leqslant \rho$, $\max\left\{\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p, \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_q, \max_{\sigma\in\Sigma} \|\mathbf{A}_\sigma - \mathbf{A}'_\sigma\|_q\right\} \leqslant \Delta$.

Claim The following holds for any $x \in \Sigma^\star$:

$$|f_A(x) - f_{A'}(x)| \leqslant (|x| + 2)\rho^{|x|+1}\Delta \quad .$$

Proof By induction on $|x|$ we first prove $\|\mathbf{A}_x - \mathbf{A}'_x\|_q \leqslant |x|\rho^{|x|-1}\Delta$:

$$\|\mathbf{A}_{x\sigma} - \mathbf{A}'_{x\sigma}\|_q \leqslant \|\mathbf{A}_x - \mathbf{A}'_x\|_q\|\mathbf{A}_\sigma\|_q + \|\mathbf{A}'_x\|_q\|\mathbf{A}_\sigma - \mathbf{A}'_\sigma\|_q \leqslant |x|\rho^{|x|}\Delta + \rho^{|x|}\Delta = (|x| + 1)\rho^{|x|}\Delta \quad .$$

$$\begin{aligned}
|f_A(x) - f_{A'}(x)| &= |\boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta} - \boldsymbol{\alpha}'^\top \mathbf{A}'_x \boldsymbol{\beta}'| \leqslant |\boldsymbol{\alpha}^\top(\mathbf{A}_x\boldsymbol{\beta} - \mathbf{A}'_x\boldsymbol{\beta}')| + |(\boldsymbol{\alpha} - \boldsymbol{\alpha}')^\top \mathbf{A}'_x\boldsymbol{\beta}'| \\
&\leqslant \|\boldsymbol{\alpha}\|_p\|\mathbf{A}_x\boldsymbol{\beta} - \mathbf{A}'_x\boldsymbol{\beta}'\|_q + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p\|\mathbf{A}'_x\boldsymbol{\beta}'\|_q \\
&\leqslant \|\boldsymbol{\alpha}\|_p\|\mathbf{A}_x\|_q\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_q + \|\boldsymbol{\alpha}\|_p\|\mathbf{A}_x - \mathbf{A}'_x\|_q\|\boldsymbol{\beta}'\|_q + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p\|\mathbf{A}'_x\|_q\|\boldsymbol{\beta}'\|_q \\
&\leqslant \rho^{|x|+1}\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_q + \rho^2\|\mathbf{A}_x - \mathbf{A}'_x\|_q + \rho^{|x|+1}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p \\
&\leqslant \rho^{|x|+1}\Delta + \rho^2\rho^{|x|-1}|x|\Delta + \rho^{|x|+1}\Delta \quad .
\end{aligned}$$

Norms on Languages

- L_p norms ($p \in [1, \infty]$), γ-discounted L_p norms ($\gamma \in (0,1)$)

$$\|f\|_p = \left(\sum_x |f(x)|^p\right)^{1/p} \qquad \|f\|_{p,\gamma} = \left(\sum_x \gamma^{p|x|}|f(x)|^p\right)^{1/p}$$

- Dirichlet norm

$$\|f\|_D = \left(\sum_x (|x|+1)|f(x)|^2\right)^{1/2}$$

- Bisimulation norms **[FZ14, BGP17]**

$$\|f\|_{\infty,\gamma} = \sup_{x \in \Sigma^\star} \gamma^{|x|}|f(x)| \qquad \|f\|_B = \sup_{x \in \Sigma^\infty} \sum_{k \geqslant 0} \gamma^k |f(x_{\leqslant k})|$$

Aside: Banach and Hilbert Spaces

- A (possibly infinite-dimensional) vector space $\mathcal{X}$ equipped with a norm $\| \bullet \| : \mathcal{X} \to [0, \infty)$ is a *Banach space* if the pair $(\mathcal{X}, \| \bullet \|)$ is complete, i.e. Cauchy sequences converge.
  - Examples: $\ell_p = \{f : \Sigma^\star \to \mathbb{R} : \|f\|_p < \infty\}$
  - Exercise: the set of rational $f \in \ell_p$ is dense in ℓ_p for any $p \in [1, \infty]$
- A (real) *Hilbert space* is a Banach space $(\mathcal{X}, \| \bullet \|)$ equipped with an inner product $\langle \bullet, \bullet \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\|\mathbf{v}\| = \sqrt{\langle v, v \rangle}$
  - Example: ℓ_2 with $\|f\|_2^2 = \langle f, f \rangle = \sum_{x \in \Sigma^\star} f(x)^2$
  - Example $\ell_D = \{f : \|f\|_D < \infty\}$ with $\|f\|_D^2 = \langle f, f \rangle_D = \sum_{x \in \Sigma^\star} (|x| + 1) f(x)^2$
- A Hilbert space is *separable* if it admits a countable orthonormal basis.
  - Examples: ℓ_2 and ℓ_D are separable

Perturbation Bounds: Language→Hankel

research

Consider the Hilbert space $\ell_D = \{f : \Sigma^\star \to \mathbb{R} : \|f\|_D < \infty\}$ with the Dirichlet inner product

$$\langle f, g\rangle_D = \sum_{x \in \Sigma^\star} (|x| + 1) f(x) g(x) \quad .$$

Consider the Frobenius norm on matrices $\mathbf{T} \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ given by

$$\|\mathbf{T}\|_F = \sqrt{\sum_{x,y \in \Sigma^\star} \mathbf{T}(x,y)^2} \quad .$$

Claim If $f, f' \in \ell_D$ are two weighted languages such that $\|f - f'\|_D \leqslant \Delta$, then their corresponding Hankel matrices satisfy $\|\mathbf{H}_f - \mathbf{H}_{f'}\|_F \leqslant \Delta$.

Proof

$$\|\mathbf{H}_f - \mathbf{H}_{f'}\|_F^2 = \sum_{x,y \in \Sigma^\star} (\mathbf{H}_f(x,y) - \mathbf{H}_{f'}(x,y))^2 = \sum_{x,y \in \Sigma^\star} (f(x \cdot y) - f'(x \cdot y))^2$$

$$= \sum_{z \in \Sigma^\star} (|z| + 1)(f(z) - f'(z))^2 = \|f - f'\|_D^2$$

Aside: Singular Value Decomposition (SVD)

For any $\mathbf{M} \in \mathbb{R}^{n \times m}$ with $\mathrm{rank}(\mathbf{M}) = k$ there exists a *singular value decomposition*

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{i=1}^{k} \mathfrak{s}_i \mathbf{u}_i \mathbf{v}_i^\top$$

- $\mathbf{D} \in \mathbb{R}^{k \times k}$ diagonal contains k sorted *singular values* $\mathfrak{s}_1 \geqslant \mathfrak{s}_2 \geqslant \cdots \geqslant \mathfrak{s}_k > 0$
- $\mathbf{U} \in \mathbb{R}^{n \times k}$ contains k *left singular vectors*, i.e. orthonormal columns $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$
- $\mathbf{V} \in \mathbb{R}^{m \times k}$ contains k *right singular vectors*, i.e. orthonormal columns $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$

Properties of SVD

- $\mathbf{M} = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{D}^{1/2}\mathbf{V}^\top)$ is a rank factorization
- Can be used to compute the pseudo-inverse as $\mathbf{M}^+ = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top$
- Provides optimal low-rank approximations. For $k' < k$, $\mathbf{M}_{k'} = \mathbf{U}_{k'}\mathbf{D}_{k'}\mathbf{V}_{k'}^\top = \sum_{i=1}^{k'} \mathfrak{s}_i \mathbf{u}_i \mathbf{v}_i^\top$ satisfies

$$\mathbf{M}_{k'} \in \operatorname*{argmin}_{\mathrm{rank}(\hat{M}) \leqslant k'} \|\mathbf{M} - \hat{\mathbf{M}}\|_2$$

Perturbation Bounds: Hankel→Automaton [Bal13]

research

- Suppose $f : \Sigma^\star \to \mathbb{R}$ has rank n and $\varepsilon \in \mathcal{P}, \mathcal{S} \subset \Sigma^\star$ are such that the sub-block $\mathbf{H} \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}}$ of $\mathbf{H}_f$ satisfies $\mathrm{rank}(\mathbf{H}) = n$
- Let $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ be obtained as follows:
  1. Compute the SVD factorization $\mathbf{H} = \mathbf{PS}$; i.e. $\mathbf{P} = \mathbf{UD}^{1/2}$ and $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{V}^\top$
  2. Let $\boldsymbol{\alpha}^\top$ (resp. $\boldsymbol{\beta}$) be the ϵ-row of $\mathbf{P}$ (resp. ϵ-column of $\mathbf{S}$)
  3. Let $\mathbf{A}_\sigma = \mathbf{P}^+ \mathbf{H}_\sigma \mathbf{S}^+$, where $\mathbf{H}_\sigma \in \mathbb{R}^{\mathcal{P} \cdot \sigma \times \mathcal{S}}$ is a sub-block of $\mathbf{H}_f$
- Suppose $\hat{\mathbf{H}} \in \mathbb{R}^{\mathcal{P} \times \mathcal{S}}$ and $\hat{\mathbf{H}}_\sigma \in \mathbb{R}^{\mathcal{P} \cdot \sigma \times \mathcal{S}}$ satisfy $\max\{\|\mathbf{H} - \hat{\mathbf{H}}\|_2, \max_\sigma \|\mathbf{H}_\sigma - \hat{\mathbf{H}}_\sigma\|_2\} \leqslant \Delta$
- Let $\hat{A} = \langle \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \{\hat{\mathbf{A}}_\sigma\} \rangle$ be obtained as follows:
  1. Compute the SVD rank-n approximation $\hat{\mathbf{H}} \approx \hat{\mathbf{P}}\hat{\mathbf{S}}$; i.e. $\hat{\mathbf{P}} = \hat{\mathbf{U}}_n \hat{\mathbf{D}}_n^{1/2}$ and $\hat{\mathbf{S}} = \hat{\mathbf{D}}_n^{1/2} \hat{\mathbf{V}}_n^\top$
  2. Let $\hat{\boldsymbol{\alpha}}^\top$ (resp. $\hat{\boldsymbol{\beta}}$) be the ϵ-row of $\hat{\mathbf{P}}$ (resp. ϵ-column of $\hat{\mathbf{S}}$)
  3. Let $\hat{\mathbf{A}}_\sigma = \hat{\mathbf{P}}^+ \hat{\mathbf{H}}_\sigma \hat{\mathbf{S}}^+$

Claim For any pair of Hölder conjugate (p, q) we have

$$\max\{\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_p, \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_q, \max_\sigma \|\mathbf{A}_\sigma - \hat{\mathbf{A}}_\sigma\|_q\} \leqslant \mathcal{O}(\Delta)$$

Applications and Limitations of Perturbation Bounds

Applications

- Analysis of machine learning algorithms for WFA **[BM12, BCLQ14, BM17]**
- Statistical properties of classes of WFA (e.g. Rademacher complexity) **[BM15, BM18]**
- Continuity of operations on WFA and rational languages **[BGP17]**

Limitations

- Automaton$\rightarrow$Language: grow with $|x|$, depend on representation chosen for A
- Language$\rightarrow$Hankel: only applies to restricted choice of norms (?)
- Hankel$\rightarrow$Automaton: depends on algorithm, cumbersome to prove

Outline

Motivation: Approximate Minimization

- Suppose f is a weighted language with $\text{rank}(f) = n$ and $\|f\| < \infty$
- Problem Given $\hat{n} < n$ find $\hat{f}$ with $\text{rank}(\hat{f}) = \hat{n}$ such that

$$\|f - \hat{f}\| \approx \min_{\text{rank}(f') \leqslant \hat{n}} \|f - f'\|$$

- Typically, f is given by a minimal WFA A and the output is a WFA $\hat{A}$ with $|\hat{A}| = \hat{n}$
- The techniques described so far are too brittle to solve this problem!

Aside: Operators on Hilbert Spaces

- Let $\mathcal{X}_1$, $\mathcal{X}_2$ be a separable Hilbert spaces. Any linear operator $\mathbf{T} : \mathcal{X}_1 \to \mathcal{X}_2$ can be represented as an infinite matrix
- A linear operator $\mathbf{T} : \mathcal{X}_1 \to \mathcal{X}_2$ is *bounded* if $\|\mathbf{T}\|_{\mathrm{op}} = \sup_{\|\mathbf{v}\|_{\mathcal{X}_1} \leqslant 1} \|\mathbf{T}\mathbf{v}\|_{\mathcal{X}_2} < \infty$
- The adjoint $\mathbf{T}^* : \mathcal{X}_2 \to \mathcal{X}_1$ of a bounded linear operator $\mathbf{T}$ is given by $\langle \mathbf{T}\mathbf{u}, \mathbf{v} \rangle_{\mathcal{X}_2} = \langle \mathbf{u}, \mathbf{T}^*\mathbf{v} \rangle_{\mathcal{X}_1}$
- A bounded linear operator $\mathbf{T}$ is *compact* if it is the limit of a sequence of finite-rank operators (w.r.t. the topology induced by $\| \bullet \|_{\mathrm{op}}$).
  - Example: all finite-rank operators are compact
- Compact linear operators $\mathbf{T}$ admit SVD (a.k.a. Hilbert–Schmidt decomposition)

$$\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^* = \sum_{i=1}^{k} \mathfrak{s}_i \mathbf{u}_i \langle \mathbf{v}_i, \bullet \rangle_{\mathcal{X}_1} \quad .$$

  Here $k = \mathrm{rank}(\mathbf{T}) \leqslant \infty$, and if $k = \infty$ then $\lim_{i \to \infty} \mathfrak{s}_i = 0$.
- Finite-rank bounded operators $\mathbf{T}$ admit a pseudo-inverse $\mathbf{T}^+$

Hankel Operators

research

A Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ can be interpreted as a linear operator $\mathbf{H}_f : \mathbb{R}^{\Sigma^\star} \to \mathbb{R}^{\Sigma^\star}$:

$$(\mathbf{H}_f g)(x) = \sum_{y \in \Sigma^\star} f(x \cdot y) g(y) \quad .$$

- Fliess–Kronecker: Finite rank if and only if f rational
- When does it admit an SVD? When it is a compact operator on a Hilbert space!

Shift Characterization

- Define the forward/backward left/right shift operators $\mathbf{L}_\sigma, \mathbf{L}_\sigma^*, \mathbf{R}_\sigma, \mathbf{R}_\sigma^* : \mathbb{R}^{\Sigma^\star} \to \mathbb{R}^{\Sigma^\star}$ as: $(\mathbf{L}_\sigma^* f)(x) = f(\sigma x)$, $(\mathbf{R}_\sigma^* f)(x) = f(x\sigma)$

$$(\mathbf{L}_\sigma f)(x) = \begin{cases} f(\sigma^{-1}x) & x_1 = \sigma \\ 0 & \text{otherwise} \end{cases} \qquad (\mathbf{R}_\sigma f)(x) = \begin{cases} f(x\sigma^{-1}) & x_{|x|} = \sigma \\ 0 & \text{otherwise} \end{cases}$$

- Exercise A linear operator $\mathbf{T} : \mathbb{R}^{\Sigma^\star} \to \mathbb{R}^{\Sigma^\star}$ is Hankel if and only if $\mathbf{R}_\sigma^* \mathbf{T} = \mathbf{T}\mathbf{L}_\sigma$, $\forall \sigma \in \Sigma$

Aside: Operator-Theoretic Proof of Fliess' Theorem

research

Claim Suppose $\mathbf{H}_f : \ell_2 \to \ell_2$ is bounded and has finite rank n. Then there exists a WFA $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ with n states such that $f_A = f$

Proof

Take a rank factorization $\mathbf{H}_f = \mathbf{PS}$ and note $\mathbf{P}$ and $\mathbf{S}$ are bounded and finite rank.
Build the automaton A by taking:

- $\boldsymbol{\alpha}^\top$ the ϵ-row of $\mathbf{P}$; i.e. $\boldsymbol{\alpha}^\top = \mathbf{P}(\varepsilon, -)$
- $\boldsymbol{\beta}$ the ϵ-column of $\mathbf{S}$; i.e. $\boldsymbol{\beta} = \mathbf{S}(-, \varepsilon)$
- $\mathbf{A}_\sigma = \mathbf{S}\mathbf{L}_\sigma\mathbf{S}^+$

It suffices to show that for any $x \in \Sigma^\star$ we have $\boldsymbol{\alpha}^\top \mathbf{A}_x = \mathbf{P}(x, -)$. By induction on length of x:

$$
\begin{aligned}
\boldsymbol{\alpha}^\top \mathbf{A}_x \mathbf{A}_\sigma &= \mathbf{P}(x,-)\mathbf{S}\mathbf{L}_\sigma\mathbf{S}^+ = \Pi_x \mathbf{P}\mathbf{S}\mathbf{L}_\sigma\mathbf{S}^+ = \Pi_x \mathbf{H}_f \mathbf{L}_\sigma \mathbf{S}^+ = \Pi_x \mathbf{R}_\sigma^* \mathbf{H}_f \mathbf{S}^+ \\
&= \Pi_x \mathbf{R}_\sigma^* \mathbf{P}\mathbf{S}\mathbf{S}^+ = \Pi_x \mathbf{R}_\sigma^* \mathbf{P} = \Pi_{x\sigma} \mathbf{P} = \mathbf{P}(x\sigma, -)
\end{aligned}
$$

Which Hankel Operators Admit an SVD?

A Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ can be interpreted as a linear operator $\mathbf{H}_f : \mathbb{R}^{\Sigma^\star} \to \mathbb{R}^{\Sigma^\star}$:

$$(\mathbf{H}_f g)(x) = \sum_{y \in \Sigma^\star} f(x \cdot y) g(y) \ .$$

- Fliess–Kronecker: Finite rank if and only if f rational
- When does it admit an SVD? When it is a compact operator on a Hilbert space!
- Finite rank operators are compact if and only if they are bounded: $\|\mathbf{H}_f\|_{op} = \sup_{\|g\|_2 \leqslant 1} \|\mathbf{H}_f g\|_2 < \infty$
- When is a finite rank Hankel operator bounded?

Boundedness of ℓ_2 and Dirichlet Norms

research

Claim Suppose $f : \Sigma^\star \to \mathbb{R}$ is rational. Then $\|f\|_2 < \infty$ if and only if $\|f\|_D < \infty$

Proof One direction is easy:

$$\|f\|_2^2 = \sum_{x \in \Sigma^\star} f(x)^2 \leqslant \sum_{x \in \Sigma^\star} (|x| + 1) f(x)^2 = \|f\|_D^2 \ .$$

The other direction is more technical. Let $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ be a minimal WFA for f^2 with n states. Then one can show that the spectral radius of $\mathbf{A} = \sum_\sigma \mathbf{A}_\sigma$ satisfies $\rho = \rho(\mathbf{A}) < 1$ (see **[BPP17]**).

$$\begin{aligned} \sum_{x \in \Sigma^t} f(x)^2 &= \sum_{x \in \Sigma^t} \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta} = \boldsymbol{\alpha}^\top (\mathbf{A}_{\sigma_1} + \cdots + \mathbf{A}_{\sigma_k}) \cdots (\mathbf{A}_{\sigma_1} + \cdots + \mathbf{A}_{\sigma_k}) \boldsymbol{\beta} \\ &= \boldsymbol{\alpha}^\top \mathbf{A}^t \boldsymbol{\beta} \leqslant \mathcal{O}(t^n \rho^t) \ . \end{aligned}$$

Therefore, since $\rho < 1$ we have

$$\|f\|_D^2 = \sum_{x \in \Sigma^\star} (|x| + 1) f(x)^2 = \sum_{t \geqslant 0} \sum_{x \in \Sigma^t} (t+1) \boldsymbol{\alpha}^\top \mathbf{A}^t \boldsymbol{\beta} \leqslant \sum_{t \geqslant 0} \mathcal{O}(t^{n+1} \rho^t) < \infty \ .$$

Bounded Hankel Operators of Finite Rank

Let $\mathbf{H}_f : \ell_2 \to \ell_2$ be a finite rank Hankel operator.

Theorem The operator $\mathbf{H}_f$ is bounded if and only if $f \in \ell_2$.

Proof Since f is the first row of $\mathbf{H}_f$, from $\mathbf{H}_f$ bounded to $\|f\|_2 < \infty$ is easy:

$$\infty > \|\mathbf{H}_f\|_{op} = \sup_{\|g\|_2 \leqslant 1} \|\mathbf{H}_f g\|_2 \geqslant \|\mathbf{H}_f \mathbf{e}_\epsilon\|_2 = \|f\|_2 \ .$$

The other direction uses the boundedness of the Dirichlet norm: let $\|g\|_2 \leqslant 1$, then

$$\begin{aligned}
\|H_f g\|_2^2 &= \sum_{x \in \Sigma^\star} \left(\sum_{y \in \Sigma^\star} f(x \cdot y) g(y) \right)^2 = \sum_{x \in \Sigma^\star} \langle \mathbf{L}_x^* f, g \rangle^2 \\
&\leqslant \|g\|_2^2 \sum_{x \in \Sigma^\star} \|\mathbf{L}_x^* f\|_2^2 \leqslant \sum_{x \in \Sigma^\star} \|\mathbf{L}_x^* f\|_2^2 \\
&= \sum_{x \in \Sigma^\star} \sum_{y \in \Sigma^\star} f(x \cdot y)^2 = \sum_{z \in \Sigma^\star} (|z| + 1) f(z)^2 = \|f\|_D^2 < \infty \ .
\end{aligned}$$

Are We Done Yet?

Approximate Minimization Strategy

1. Take rational f with $\mathrm{rank}(f) = n$ and $\|f\|_2 < \infty$
2. Since $\mathbf{H}_f : \ell_2 \to \ell_2$ is compact, it admits an SVD
$$\mathbf{H}_f = \sum_{i=1}^{n} \mathfrak{s}_i \mathbf{u}_i \langle \mathbf{v}_i, \bullet \rangle \ .$$
3. Given $\hat{n} < n$ take the corresponding low-rank approximation $\hat{\mathbf{H}}$
$$\hat{\mathbf{H}} = \sum_{i=1}^{\hat{n}} \mathfrak{s}_i \mathbf{u}_i \langle \mathbf{v}_i, \bullet \rangle \ .$$
4. Compute a WFA $\hat{A}$ from $\hat{\mathbf{H}}$ $\leftarrow$ **NOT NECESSARILY HANKEL!**
5. Bound the error between f and $\hat{f} = f_{\hat{A}}$ as
$$\|f - \hat{f}\|_2 \leqslant \|\mathbf{H}_f - \hat{\mathbf{H}}\|_{op} = \mathfrak{s}_{\hat{n}+1} \ .$$

Duality Between Rank Factorization and Minimal WFA

Well-known fact: If $\mathbf{M}$ has rank n and $\mathbf{M} = \mathbf{P}\mathbf{S} = \mathbf{P}'\mathbf{S}'$ are two rank factorizations, then there exists invertible $\mathbf{Q} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{P}' = \mathbf{P}\mathbf{Q} \qquad \mathbf{S}' = \mathbf{Q}^{-1}\mathbf{S}$$

Well-known fact: If $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ and $A' = \langle \boldsymbol{\alpha}', \boldsymbol{\beta}', \{\mathbf{A}'_\sigma\} \rangle$ are minimal WFA for f of rank n, then there exists invertible $\mathbf{Q} \in \mathbb{R}^{n \times n}$ such that

$$\boldsymbol{\alpha}'^\top = \boldsymbol{\alpha}^\top \mathbf{Q} \qquad \boldsymbol{\beta}' = \mathbf{Q}^{-1}\boldsymbol{\beta} \qquad \mathbf{A}'_\sigma = \mathbf{Q}^{-1}\mathbf{A}_\sigma\mathbf{Q}$$

Less-known fact: From the proof of the Fliess–Kronecker theorem applied to f of rank n one obtains a bijection

$$\{(\mathbf{P}, \mathbf{S}) : \mathbf{H}_f = \mathbf{P}\mathbf{S}, \mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{S}) = n\} \quad \leftrightarrow \quad \{A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle : f_A = f, |A| = n\}$$

Singular Value Automata

- Let A be a minimal WFA with n states computing f
- Definition A is a *singular value automaton* (SVA) if the forward-backward factorization $\mathbf{H}_f = \mathbf{P}_A \mathbf{S}_A$ comes from a singular value decomposition, i.e. $\mathbf{P}_A = \mathbf{U}\mathbf{D}^{1/2}$, $\mathbf{S}_A = \mathbf{D}^{1/2}\mathbf{V}^\top$, with $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$ and $\mathbf{D} = \mathrm{diag}(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$ with $\mathfrak{s}_1 \geqslant \cdots \geqslant \mathfrak{s}_n > 0$
- Theorem Every rational f with $\|f\|_2 < \infty$ admits an SVA
- The SVA of f is "as unique" as the SVD of $\mathbf{H}_f$
  - Example: if all inequalities between singular values are strict, SVD is unique up to sign changes in pairs of associated left/right singular vectors $\Rightarrow$ SVA unique up to sign changes in pairs of associated initial/final weights
- Given a minimal WFA $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\}\rangle$ for f with $\|f\|_2 < \infty$ there exists an invertible $\mathbf{Q} \in \mathbb{R}^{n \times n}$ such that $A^{\mathbf{Q}} = \langle \mathbf{Q}^\top \boldsymbol{\alpha}, \mathbf{Q}^{-1}\boldsymbol{\beta}, \{\mathbf{Q}^{-1}\mathbf{A}_\sigma \mathbf{Q}\}\rangle$ is an SVA for f
- Definition could be changed to have $\mathbf{P}_A = \mathbf{U}$ and $\mathbf{S}_A = \mathbf{D}\mathbf{V}^\top$, or $\mathbf{P}_A = \mathbf{U}\mathbf{D}$ and $\mathbf{S}_A = \mathbf{V}^\top$. But the current one makes computation of $\mathbf{Q}$ above more "symmetric"

Why Are SVA Special?

research

- It *orthogonalizes* the states of a WFA!
- Suppose $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_\sigma\} \rangle$ is an SVA with n states for f inducing the SVD

$$\mathbf{H}_f = \sum_{i=1}^{n} \mathfrak{s}_i \mathbf{u}_i \langle \mathbf{v}_i, \bullet \rangle \enspace .$$

- For $i \in [n]$ let $A_i = \langle \boldsymbol{\alpha}, \mathbf{e}_i, \{\mathbf{A}_\sigma\} \rangle$ where $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)$ is the ith coordinate vector
- The language f_i of A_i is given by $f_i(x) = \boldsymbol{\alpha}^\top \mathbf{A}_x \mathbf{e}_i = \boldsymbol{\alpha}_A(x)^\top [i]$; i.e. is the "memory" of state i after reading x
- The language f_i is also the ith column of the forward matrix $\mathbf{P}_A = \mathbf{U}\mathbf{D}^{1/2}$; i.e. $f_i = \sqrt{\mathfrak{s}_i} \mathbf{u}_i$
- Since the columns of $\mathbf{U}$ are orthonormal, the languages f_i and f_j with $i \neq j$ are orthogonal

Outline

The Gramians of a WFA

research

- Let A be a minimal WFA for f with $n = \mathrm{rank}(f)$ inducing the rank factorization $\mathbf{H}_f = \mathbf{P}\mathbf{S}$ (i.e. $\mathbf{P} = \mathbf{P}_A$ and $\mathbf{S} = \mathbf{S}_A$)
- The *reachability Gramian* of A is the (possibly infinite) $n \times n$ matrix $\mathbf{G}_p = \mathbf{P}^\top \mathbf{P}$

$$\mathbf{G}_p = \mathbf{P}^\top \mathbf{P} = \sum_{x \in \Sigma^\star} \mathbf{P}(x, -)^\top \mathbf{P}(x, -) = \sum_{x \in \Sigma^\star} \left(\alpha^\top \mathbf{A}_x\right)^\top \left(\alpha^\top \mathbf{A}_x\right)$$

- The *observability Gramian* of A is the (possibly infinite) $n \times n$ matrix $\mathbf{G}_s = \mathbf{S}\mathbf{S}^\top$ given by

$$\mathbf{G}_s = \mathbf{S}\mathbf{S}^\top = \sum_{x \in \Sigma^\star} \mathbf{S}(-, x)\mathbf{S}(-, x)^\top = \sum_{x \in \Sigma^\star} \left(\mathbf{A}_x \beta\right) \left(\mathbf{A}_x \beta\right)^\top$$

Existence of the Gramians

Let A be a minimal WFA for f with $n = \mathrm{rank}(f)$ inducing the rank factorization $\mathbf{H}_f = \mathbf{P}\mathbf{S}$ (i.e. $\mathbf{P} = \mathbf{P}_A$ and $\mathbf{S} = \mathbf{S}_A$)

Claim The Gramians of A are finite if and only if $\|f\|_2 < \infty$

Proof (one direction only)

Suppose $\|f\|_2 < \infty$ and let $A' = A^{\mathbf{Q}} = \langle \mathbf{Q}^\top \alpha, \mathbf{Q}^{-1}\beta, \{\mathbf{Q}^{-1}\mathbf{A}_\sigma \mathbf{Q}\}\rangle$ be an SVA for f
Observe the Gramians $\mathbf{G}'_p$ and $\mathbf{G}'_s$ of A' exist since

$$\mathbf{G}'_p = \mathbf{P}_{A'}^\top \mathbf{P}_{A'} = \mathbf{D}^{1/2}\mathbf{U}^\top \mathbf{U}\mathbf{D}^{1/2} = \mathbf{D}$$
$$\mathbf{G}'_s = \mathbf{S}_{A'}\mathbf{S}_{A'}^\top = \mathbf{D}^{1/2}\mathbf{V}^\top \mathbf{V}\mathbf{D}^{1/2} = \mathbf{D}$$

On the other hand, since $\mathbf{P}_{A'} = \mathbf{P}_A\mathbf{Q}$ and $\mathbf{S}_{A'} = \mathbf{Q}^{-1}\mathbf{S}_A$ we have

$$\mathbf{G}'_p = \mathbf{Q}^\top \mathbf{G}_p \mathbf{Q} \qquad \mathbf{G}'_s = \mathbf{Q}^{-\top}\mathbf{G}_s\mathbf{Q}^{-1}$$

Therefore $\mathbf{G}_p$ and $\mathbf{G}_s$ must be finite

From Gramians to SVA

research

- Let A be a minimal WFA for f with $\|f\|_2 < \infty$
- Suppose we have the Gramians of A: $\mathbf{G}_p$ and $\mathbf{G}_s$
- Recall from the previous proof that
  - If A' is SVA then $\mathbf{G}'_p = \mathbf{G}'_s = \mathbf{D} = \mathrm{diag}(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$
  - If $A' = A^{\mathbf{Q}}$ then $\mathbf{G}'_p = \mathbf{Q}^\top \mathbf{G}_p \mathbf{Q}$ and $\mathbf{G}'_s = \mathbf{Q}^{-\top} \mathbf{G}_s \mathbf{Q}^{-1}$
- Claim The following algorithm returns $\mathbf{Q}$ such that $A^{\mathbf{Q}}$ is an SVA
  1. Compute the Cholesky decompositions $\mathbf{G}_p = \mathbf{L}_p \mathbf{L}_p^\top$ and $\mathbf{G}_s = \mathbf{L}_s \mathbf{L}_s^\top$
  2. Compute the SVD decomposition $\mathbf{L}_p^\top \mathbf{L}_s = \mathbf{U}\mathbf{D}\mathbf{V}^\top$
  3. Let $\mathbf{Q} = \mathbf{L}_p^{-\top} \mathbf{U} \mathbf{D}^{1/2}$
- In particular, the $\mathbf{D}$ in this algorithm is the matrix of singular values of $\mathbf{H}_f$
- See proof in [BPP17]

Computing Norms Using Gramians

Suppose A is a minimal WFA for f with $\|f\|_2 < \infty$.
Let $\mathbf{G}_p$ and $\mathbf{G}_s$ be the Gramians of A.
Then the following hold:

- $\|f\|_2^2 = \boldsymbol{\alpha}^\top \mathbf{G}_s \boldsymbol{\alpha} = \boldsymbol{\beta}^\top \mathbf{G}_p \boldsymbol{\beta}$
- $\|f\|_D^2 = \|\mathbf{H}_f\|_F^2 = \mathrm{Tr}(\mathbf{G}_p \mathbf{G}_s)$
- $\|\mathbf{H}_f\|_{op}^2 = \rho(\mathbf{G}_p \mathbf{G}_s) = \max\{|\lambda| : \det(\mathbf{G}_p \mathbf{G}_s - \lambda \mathbf{I}) = 0\}$

Computing the Gramians Using Fixed-Points

research

Let A be a minimal WFA for f with $\|f\|_2 < \infty$.

Claim $\mathbf{X} = \mathbf{G}_p$ and $\mathbf{Y} = \mathbf{G}_s$ are solutions of the fixed-point equations

$$\mathbf{X} = F_p(\mathbf{X}) = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_\sigma \mathbf{A}_\sigma^\top \mathbf{X} \mathbf{A}_\sigma \qquad \mathbf{Y} = F_s(\mathbf{Y}) = \boldsymbol{\beta}\boldsymbol{\beta}^\top + \sum_\sigma \mathbf{A}_\sigma \mathbf{Y} \mathbf{A}_\sigma^\top$$

Proof Recall $\mathbf{G}_p = \mathbf{P}_A^\top \mathbf{P}_A = \sum_{x \in \Sigma^\star} \mathbf{P}_A(x,-)\mathbf{P}_A(x,-)^\top$ and $\mathbf{P}_A(x,-) = \boldsymbol{\alpha}^\top \mathbf{A}_x$. Therefore:

$$\begin{aligned}
\mathbf{G}_p &= \sum_{x \in \Sigma^\star} (\mathbf{A}_x^\top \boldsymbol{\alpha})(\boldsymbol{\alpha}^\top \mathbf{A}_x) = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_{x \in \Sigma^+} (\mathbf{A}_x^\top \boldsymbol{\alpha})(\boldsymbol{\alpha}^\top \mathbf{A}_x) \\
&= \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_{\sigma \in \Sigma} \sum_{x \in \Sigma^\star} \mathbf{A}_\sigma^\top (\mathbf{A}_x^\top \boldsymbol{\alpha})(\boldsymbol{\alpha}^\top \mathbf{A}_x) \mathbf{A}_\sigma \\
&= \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_{\sigma \in \Sigma} \mathbf{A}_\sigma^\top \left(\sum_{x \in \Sigma^\star} (\mathbf{A}_x^\top \boldsymbol{\alpha})(\boldsymbol{\alpha}^\top \mathbf{A}_x) \right) \mathbf{A}_\sigma = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_{\sigma \in \Sigma} \mathbf{A}_\sigma^\top \mathbf{G}_p \mathbf{A}_\sigma
\end{aligned}$$

Solving the Fixed-Point Equations

- Recall the reachability Gramian $\mathbf{G}_p$ is a solution of

$$\mathbf{X} = F_p(\mathbf{X}) = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \sum_\sigma \mathbf{A}_\sigma^\top \mathbf{X} \mathbf{A}_\sigma$$

- Let ρ be the spectral radius of $\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma$, where $\otimes$ denotes the Kronecker product (i.e. $\mathbf{A}_\sigma \otimes \mathbf{A}_\sigma \in \mathbb{R}^{n^2 \times n^2}$)
- We distinguish two cases. If $\rho < 1$:
  - $\mathbf{X} = F_p(\mathbf{X})$ has a *unique* solution
  - Can be found by solving the linear system with n^2 unknowns obtained through vectorization: $\mathrm{vec}(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top) = \boldsymbol{\alpha} \otimes \boldsymbol{\alpha}$ and $\mathrm{vec}(\mathbf{A}_\sigma^\top \mathbf{X} \mathbf{A}_\sigma) = (\mathbf{A}_\sigma \otimes \mathbf{A}_\sigma)^\top \mathrm{vec}(\mathbf{X})$
- If $\rho \geqslant 1$:
  - $\mathbf{X} = F_p(\mathbf{X})$ might have multiple solutions (there is at least one because $\mathbf{G}_p$ is defined)
  - In this case rephrase the problem: $\mathbf{G}_p$ is the least positive semi-definite solution of the linear matrix inequality $\mathbf{X} \succeq F_p(\mathbf{X})$
  - The solution can be found by semi-definite programming

Computing SVA: Summary

Suppose A is a WFA computing a function f. To compute an SVA for f do:

1. Test if $\|f\|_2 < \infty$
2. Minimize A if necessary
3. Compute Gramians $\mathbf{G}_p$ and $\mathbf{G}_s$ (using linear solver or semi-definite solver)
4. Find change of basis $\mathbf{Q}$ through Cholesky and SVD of finite matrices
5. Return $A^{\mathbf{Q}}$

Final remarks

- Runs in time polynomial in $|A|$ and $|\Sigma|$
- Easy to implement in Python or MATLAB

Outline

1. Weighted Languages, Weighted Automata, and Hankel Matrices
2. Perturbation Bounds Between Representations
3. Singular Value Automata: Definition
4. Singular Value Automata: Computation
5. Approximate Minimization via SVA Truncation
6. Concluding Remarks

Approximate Minimization with SVA

- Suppose f is a weighted language with $\mathrm{rank}(f) = n$ and $\|f\|_2 < \infty$. Let $\mathfrak{s}_i$ be the singular values of $\mathbf{H}_f$
- Problem Given $\hat{n} < n$ find $\hat{f}$ with $\mathrm{rank}(\hat{f}) = \hat{n}$ such that

$$\|f - \hat{f}\|_2 \approx \min_{\mathrm{rank}(f') \leqslant \hat{n}} \|f - f'\|_2$$

- SVA Solution Compute SVA A for f and obtain $\hat{A}$ by removing the last $n - \hat{n}$ states

$$\|f - \hat{f}\|_2^2 \leqslant \sum_{i=\hat{n}+1}^{n} \mathfrak{s}_i^2$$

- Lower Bound Considering approximation in terms of $\| \bullet \|_D$ instead of $\| \bullet \|_2$:

$$\min_{\mathrm{rank}(f') \leqslant \hat{n}} \|f - f'\|_D^2 \geqslant \sum_{i=\hat{n}+1}^{n} \mathfrak{s}_i^2$$

Intuition for Removing the Last States from an SVA

- Suppose $A = \langle \alpha, \beta, \{\mathbf{A}_\sigma\} \rangle$ is an SVA. Since the Gramians satisfy $\mathbf{G}_p = \mathbf{G}_s = \mathbf{D} = \mathrm{diag}(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$, we have

$$\mathbf{D} = \alpha\alpha^\top + \sum_\sigma \mathbf{A}_\sigma^\top \mathbf{D} \mathbf{A}_\sigma$$

$$\mathbf{D} = \beta\beta^\top + \sum_\sigma \mathbf{A}_\sigma \mathbf{D} \mathbf{A}_\sigma^\top$$

- By looking at the diagonal entries in these equations we can deduce

$$|\mathbf{A}_\sigma(i,j)| \leqslant \sqrt{\frac{\min\{\mathfrak{s}_i, \mathfrak{s}_j\}}{\max\{\mathfrak{s}_i, \mathfrak{s}_j\}}}$$

- For example, connections between the first and last state are weak: $|\mathbf{A}_\sigma(1,n)|, |\mathbf{A}_\sigma(n,1)| \leqslant \sqrt{\mathfrak{s}_n/\mathfrak{s}_1}$
- See **[BPP15]** for a "pedestrian" bound for $\|f - \hat{f}\|_2$ based on this idea

Analysis of SVA Approximate Minimization

research

SVA

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{bmatrix} ,$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix} ,$$

$$\mathbf{A}_\sigma = \begin{bmatrix} \mathbf{A}_\sigma^{(11)} & \mathbf{A}_\sigma^{(12)} \\ \mathbf{A}_\sigma^{(21)} & \mathbf{A}_\sigma^{(22)} \end{bmatrix}$$

Truncated SVA

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{\alpha}^{(1)} \\ \mathbf{0} \end{bmatrix} = \boldsymbol{\Pi}\boldsymbol{\alpha} ,$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix} = \boldsymbol{\beta} ,$$

$$\hat{\mathbf{A}}_\sigma = \begin{bmatrix} \mathbf{A}_\sigma^{(11)} & \mathbf{0} \\ \mathbf{A}_\sigma^{(21)} & \mathbf{0} \end{bmatrix} = \mathbf{A}_\sigma \boldsymbol{\Pi}$$

$$\boldsymbol{\Pi} = \begin{bmatrix} \mathbf{I}_{\hat{n}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Analysis

- Let A be SVA for f and $\hat{A}$ truncated SVA computing $\hat{f}$
- Show $\|\hat{f}\|_2 \leqslant \|f\|_2$ (see **[BPP17]**)
- Show $\|f - \hat{f}\|_2 \leqslant \mathfrak{s}_{\hat{n}+1}^2 + \cdots + \mathfrak{s}_n^2$ *(organic free-range proof on the board)*

Outline

1. Weighted Languages, Weighted Automata, and Hankel Matrices
2. Perturbation Bounds Between Representations
3. Singular Value Automata: Definition
4. Singular Value Automata: Computation
5. Approximate Minimization via SVA Truncation
6. Concluding Remarks

The Tree Case

- Take a ranked alphabet $\Sigma = \Sigma_0 \cup \Sigma_1 \cup \cdots$
- A weighted tree automaton with n states is a tuple $A = \langle \boldsymbol{\alpha}, \{\mathbf{T}_\tau\}_{\tau \in \Sigma_{\geqslant 1}}, \{\boldsymbol{\beta}_\sigma\}_{\sigma \in \Sigma_0} \rangle$ where

$$\boldsymbol{\alpha}, \boldsymbol{\beta}_\sigma \in \mathbb{R}^n \qquad \mathbf{T}_\tau \in (\mathbb{R}^n)^{\otimes \mathrm{rk}(\tau)+1}$$

- A defines a function $f_A = \mathrm{Trees}_\Sigma \to \mathbb{R}$ through recursive vector-tensor contractions
- There exists an analogue of the Hankel matrix for $f : \mathrm{Trees}_\Sigma \to \mathbb{R}$ where rows are indexed by contexts and columns by trees
- The same ideas lead to a notion of *singular value tree automata* **[RBC16]**
- In this case the computation of the Gramians is already a highly non-trivial problem

The One Symbol Case

- When $|\Sigma| = 1$, $\Sigma^\star = \mathbb{N}$ and one recovers the classical Hankel operators studied in complex analysis and the impulse responses studied in control theory and signal processing
- A new perspective in terms of functions of one complex variable arises from the power-series point of view: for $z \in \mathbb{C}$ with small enough modulus

$$f(z) = \sum_{k \geqslant 0} a_k z^k = \sum_{k \geqslant 0} \boldsymbol{\alpha}(z\mathbf{A})^k \boldsymbol{\beta} = \boldsymbol{\alpha}^\top (\mathbf{I} - z\mathbf{A})^{-1} \boldsymbol{\beta} = \frac{p(z)}{q(z)}$$

- $\mathbb{N}$ can be embedded into a locally compact Abelian group $\mathbb{Z}$, ℓ_2 gets a new definition in terms of Fourier analysis, Hankel operators get a new definition in terms of Hardy spaces, etc.
- Example: Nehari's theorem says that $\|\mathbf{H}_f\|_{op} = \sup_{|z|<1} |f(z)|$
- Suggested readings: Peller's "Hankel Operators and Their Applications" **[Pel12]** and Fuhrmann's "A Polynomial Approach to Linear Algebra" **[Fuh11]**

Open Problems

- Complexity of testing $\|f\|_p < R$, computing and approximating ℓ_p and other norms on languages
- Complexity of optimal approximate minimization in terms of $\| \bullet \|_2$
- Quality of approximation of SVA truncation in terms of $\| \bullet \|_2$ or analysis of approximation in terms of $\| \bullet \|_D$
- Approximate minimization with other norms

Conclusions

research

- Analytic automata theory is a vastly understudied area, rich in interesting open problems (for the mathematically adventurous)
- Singular value automata provide a powerful canonical form for WFA over the reals
- Approximate minimization is a generalization of automata minimization with connections to machine learning

research

Thanks!

References I

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

B. Balle, X. Carreras, F.M. Luque, and A. Quattoni.
Spectral learning of weighted automata: A forward-backward perspective.
Machine Learning, 2014.

B. Balle, P. Gourdeau, and P. Panangaden.
Bisimulation metrics for weighted automata.
In *ICALP*, 2017.

B. Balle and M. Mohri.
Spectral learning of general weighted automata via constrained matrix completion.
In *NIPS*, 2012.

B. Balle and M. Mohri.
On the rademacher complexity of weighted automata.
In *ALT*, 2015.

References II

B. Balle and O.-A. Maillard.
Spectral learning from a single trajectory under finite-state policies.
In *ICML*, 2017.

B. Balle and M. Mohri.
Generalization Bounds for Learning Weighted Automata.
Theoretical Computer Science, 716:89–106, 2018.

B. Balle, P. Panangaden, and D. Precup.
A canonical form for weighted automata and applications to approximate minimization.
In *LICS*, 2015.

Borja Balle, Prakash Panangaden, and Doina Precup.
Singular value automata and approximate minimization.
CoRR, abs/1711.05994, 2017.

M. Fliess.
Matrices de Hankel.
Journal de Mathématiques Pures et Appliquées, 1974.

References III

Paul A Fuhrmann.
A polynomial approach to linear algebra.
Springer Science & Business Media, 2011.

Yuan Feng and Lijun Zhang.
When equivalence and bisimulation join forces in probabilistic automata.
In *International Symposium on Formal Methods*, pages 247–262. Springer, 2014.

Vladimir Peller.
Hankel operators and their applications.
Springer Science & Business Media, 2012.

G. Rabusseau, B. Balle, and S. B. Cohen.
Low-rank approximation of weighted tree automata.
In *AISTATS*, 2016.

Singular Value Automata and Approximate Minimization

Borja Balle

Amazon Research Cambridge[2]

Weighted Automata: Theory and Applications — May 2018

[2]Based on work completed before joining Amazon