Low-Rank Approximation of Weighted Tree Automata
(Supplementary Material)

Guillaume Rabusseau
Aix-Marseille University

Borja Balle
Lancaster University

Shay B. Cohen
University of Edinburgh

1 Proof of Theorem 2

Theorem. *Let $f : \mathfrak{T} \to \mathbb{R}$ be rational. If $\mathbf{H}_f = \mathbf{PS}$ is a rank factorization, then there exists a minimal WTA A computing f such that $\mathbf{P}_A = \mathbf{P}$ and $\mathbf{S}_A = \mathbf{S}$.*

Proof. Let $n = rank(f)$. Let B be an arbitrary minimal WTA computing f. Suppose B induces the rank factorization $\mathbf{H}_f = \mathbf{P}'\mathbf{S}'$. Since the columns of both $\mathbf{P}$ and $\mathbf{P}'$ are basis for the column-span of $\mathbf{H}_f$, there must exists a change of basis $\mathbf{Q} \in \mathbb{R}^{n\times n}$ between $\mathbf{P}$ and $\mathbf{P}'$. That is, $\mathbf{Q}$ is an invertible matrix such that $\mathbf{P}'\mathbf{Q} = \mathbf{P}$. Furthermore, since $\mathbf{P}'\mathbf{S}' = \mathbf{H}_f = \mathbf{PS} = \mathbf{P}'\mathbf{QS}$ and $\mathbf{P}'$ has full column rank, we must have $\mathbf{S}' = \mathbf{QS}$, or equivalently, $\mathbf{Q}^{-1}\mathbf{S}' = \mathbf{S}$. Thus, we let $A = B^{\mathbf{Q}}$, which immediately verifies $f_A = f_B = f$. It remains to be shown that A induces the rank factorization $\mathbf{H}_f = \mathbf{PS}$. Note that when proving the equivalence $f_A = f_B$ we already showed $\boldsymbol{\omega}_A(t) = \mathbf{Q}^{-1}\boldsymbol{\omega}_B(t)$, which means we have $\mathbf{S}_A = \mathbf{Q}^{-1}\mathbf{S}' = \mathbf{S}$. To show $\mathbf{P}_A = \mathbf{P}'\mathbf{Q}$ we need to show that for any $c \in \mathfrak{C}$ we have $\boldsymbol{\alpha}_A(c)^\top = \boldsymbol{\alpha}_B(c)^\top\mathbf{Q}$. This will immediately follow if we show that $\boldsymbol{\Xi}_A(c) = \mathbf{Q}^{-1}\boldsymbol{\Xi}_B(c)\mathbf{Q}$. If we proceed by induction on $\mathrm{drop}(c)$, we see the case $c = *$ is immediate, and for $c = (c', t)$ we get

$$
\begin{aligned}
\boldsymbol{\Xi}_A((c',t)) &= (\boldsymbol{\mathcal{T}}(\mathbf{Q}^{-\top}, \mathbf{Q}, \mathbf{Q}))(\mathbf{I}, \boldsymbol{\Xi}_A(c'), \boldsymbol{\omega}_A(t)) \\
&= (\boldsymbol{\mathcal{T}}(\mathbf{Q}^{-\top}, \mathbf{Q}, \mathbf{Q}))(\mathbf{I}, \mathbf{Q}^{-1}\boldsymbol{\Xi}_B(c')\mathbf{Q}, \mathbf{Q}^{-1}\boldsymbol{\omega}_B(t)) \\
&= \boldsymbol{\mathcal{T}}(\mathbf{Q}^{-\top}, \boldsymbol{\Xi}_B(c')\mathbf{Q}, \boldsymbol{\omega}_B(t)) \\
&= \mathbf{Q}^{-1}\boldsymbol{\mathcal{T}}(\mathbf{I}, \boldsymbol{\Xi}_B(c'), \boldsymbol{\omega}_B(t))\mathbf{Q} \ .
\end{aligned}
$$

Applying the same argument mutatis mutandis for $c = (t, c')$ completes the proof. □

2 Proof of Theorem 3

Theorem. *If $f : \mathfrak{T}_\Sigma \to \mathbb{R}$ is rational and strongly convergent, then $\mathbf{H}_f$ admits a singular value decomposition.*

Proof. The result will follow if we show that $\mathbf{H}_f$ is the matrix of a compact operator between Hilbert spaces [2]. We start by defining the Hilbert spaces of square-summable series indexed by trees and contexts. Given two functions $g, g' : \mathfrak{T}_\Sigma \to \mathbb{R}$ we define their inner product to be $\langle g, g'\rangle_{\mathfrak{T}} = \sum_{t\in\mathfrak{T}_\Sigma} g(t)g'(t)$ (whenever the sum converges). Let $\|g\|_{\mathfrak{T}} = \sqrt{\langle g, g\rangle_{\mathfrak{T}}}$ be the induced norm. We denote by $\ell^2_{\mathfrak{T}}$ be the real vector space of functions $\{g : \mathfrak{T} \to \mathbb{R} | \|g\|_{\mathfrak{T}} < \infty\}$, which is a separable Hilbert space because the set $\mathfrak{T}$ is countable. Similarly, given functions $g, g' : \mathfrak{C}_\Sigma \to \mathbb{R}$ we define an inner product $\langle g, g'\rangle_{\mathfrak{C}} = \sum_{c\in\mathfrak{C}_\Sigma} g(t)g'(t)$, a norm $\|g\|_{\mathfrak{C}} = \sqrt{\langle g, g\rangle_{\mathfrak{C}}}$, and a separable Hilbert space $\ell^2_{\mathfrak{C}} = \{g : \mathfrak{C} \to \mathbb{R} | \|g\|_{\mathfrak{C}} < \infty\}$. With this notation it is possible to see that $\mathbf{H}_f$ is the matrix under the standard basis on $\ell^2_{\mathfrak{T}}$ and $\ell^2_{\mathfrak{C}}$ of the operator $H_f : \ell^2_{\mathfrak{T}} \to \ell^2_{\mathfrak{C}}$ given by $(H_f g)(c) = \sum_{t\in\mathfrak{T}_\Sigma} f(c[t])g(t)$. Since f is rational, $\mathbf{H}_f$ is a finite-rank matrix and therefore H_f is a finite-rank operator. Thus, to show the compactness of H_f it only remains to show that H_f is bounded.

Given $f \in \ell^2_{\mathfrak{T}}$ and $c \in \mathfrak{C}_\Sigma$ we define a new function $f_c \in \ell^2_{\mathfrak{T}}$ given by $f_c(t) = f(c[t])$ for $t \in \mathfrak{T}_\Sigma$. Now let $g \in \ell^2_{\mathfrak{T}}$ with $\|g\|_{\mathfrak{T}} = 1$ and recall H_f is bounded if $\|H_f g\|_{\mathfrak{C}} < \infty$ for every $g \in \ell^2_{\mathfrak{T}}$ with $\|g\|_{\mathfrak{T}} = 1$. To show that H_f is bounded observe that we have:

$$
\begin{aligned}
\|H_f g\|^2_{\mathfrak{C}} &= \sum_{c\in\mathfrak{C}_\Sigma} (H_f g)(c)^2 = \sum_{c\in\mathfrak{C}_\Sigma} \left(\sum_{t\in\mathfrak{T}_\Sigma} f(c[t])g(t) \right)^2 \\
&= \sum_{c\in\mathfrak{C}_\Sigma} \langle f_c, g\rangle^2_{\mathfrak{T}} \le \|g\|^2_{\mathfrak{T}} \sum_{c\in\mathfrak{C}_\Sigma} \|f_c\|^2_{\mathfrak{T}} \\
&= \sum_{c\in\mathfrak{C}_\Sigma} \sum_{t\in\mathfrak{T}_\Sigma} f_c(t)^2 = \sum_{c\in\mathfrak{C}_\Sigma} \sum_{t\in\mathfrak{T}_\Sigma} f(c[t])^2 \\
&= \sum_{t\in\mathfrak{T}_\Sigma} |t| f(t)^2 \le \sup_{t\in\mathfrak{T}_\Sigma} |f(t)| \cdot \sum_{t\in\mathfrak{T}_\Sigma} |t||f(t)| \\
&< \infty \ ,
\end{aligned}
$$

where we used the Cauchy–Schwarz inequality, and the fact that $\sup_{t\in\mathfrak{T}_\Sigma} |f(t)|$ is bounded when f is strongly convergent. □

3 Proof of Theorem 5

Theorem. *Let $F : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ be the mapping defined by $F(\mathbf{v}) = \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \mathbf{v}, \mathbf{v}) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes}$. Then the following hold:*

(i) $\mathbf{s}$ is a fixed-point of F; i.e. $F(\mathbf{s}) = \mathbf{s}$.

(ii) $\mathbf{0}$ is in the basin of attraction of $\mathbf{s}$; i.e. $\lim_{k\to\infty} F^k(\mathbf{0}) = \mathbf{s}$.

(iii) The iteration defined by $\mathbf{s}_0 = \mathbf{0}$ and $\mathbf{s}_{k+1} = F(\mathbf{s}_k)$ converges linearly to $\mathbf{s}$; i.e. there exists $0 < \rho < 1$ such that $\|\mathbf{s}_k - \mathbf{s}\|_2 \le \mathcal{O}(\rho^k)$.

Proof. (i) We have $\boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \mathbf{s}, \mathbf{s}) = \sum_{t,t'\in\mathfrak{T}} \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \boldsymbol{\omega}^{\otimes}(t), \boldsymbol{\omega}^{\otimes}(t')) = \sum_{t,t'\in\mathfrak{T}} \boldsymbol{\omega}^{\otimes}((t,t')) = \sum_{t\in\mathfrak{T}^{\ge 1}} \boldsymbol{\omega}^{\otimes}(t)$ where $\mathfrak{T}^{\ge 1}$ is the set of trees of depth at least one. Hence $F(\mathbf{s}) = \sum_{t\in\mathfrak{T}^{\ge 1}} \boldsymbol{\omega}^{\otimes}(t) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes} = \mathbf{s}$.

(ii) Let $\mathfrak{T}^{\le k}$ denote the set of all trees with depth at most k. We prove by induction on k that $F^k(\mathbf{0}) = \sum_{t\in\mathfrak{T}^{\le k}} \boldsymbol{\omega}^{\otimes}(t)$, which implies that $\lim_{k\to\infty} F^k(\mathbf{0}) = \mathbf{s}$. This is straightforward for $k = 0$. Assuming it is true for all naturals up to $k-1$, we have

$$\begin{aligned} F^k(\mathbf{0}) &= \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, F^{k-1}(\mathbf{0}), F^{k-1}(\mathbf{0})) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes} \\ &= \sum_{t,t'\in\mathfrak{T}^{\le k-1}} \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \boldsymbol{\omega}^{\otimes}(t), \boldsymbol{\omega}^{\otimes}(t')) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes} \\ &= \sum_{t,t'\in\mathfrak{T}^{\le k-1}} \boldsymbol{\omega}^{\otimes}((t,t')) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes} \\ &= \sum_{t\in\mathfrak{T}^{\le k}} \boldsymbol{\omega}^{\otimes}(t) \ . \end{aligned}$$

(iii) Let $\mathbf{E}$ be the Jacobian of F around $\mathbf{s}$, we show that the spectral radius $\rho(\mathbf{E})$ of $\mathbf{E}$ is less than one, which implies the result by Ostrowski's theorem (see [4, Theorem 8.1.7]).

Since A is minimal, there exists trees $t_1, \cdots, t_n \in \mathfrak{T}$ and contexts $c_1, \cdots, c_n \in \mathfrak{C}$ such that both $\{\boldsymbol{\omega}(t_i)\}_{i\in[n]}$ and $\{\boldsymbol{\alpha}(c_i)\}_{i\in[n]}$ are sets of linear independent vectors in $\mathbb{R}^n$ [1]. Therefore, the sets $\{\boldsymbol{\omega}(t_i)\otimes\boldsymbol{\omega}(t_j)\}_{i,j\in[n]}$ and $\{\boldsymbol{\alpha}(c_i)\otimes\boldsymbol{\alpha}(c_j)\}_{i,j\in[n]}$ are sets of linear independent vectors in $\mathbb{R}^{n^2}$. Let $\mathbf{v} \in \mathbb{R}^{n^2}$ be an eigenvector of $\mathbf{E}$ with eigenvalue $\lambda \ne 0$, and let $\mathbf{v} = \sum_{i,j\in[n]} \beta_{i,j} (\boldsymbol{\omega}(t_i)\otimes\boldsymbol{\omega}(t_j))$ be its expression in terms of the basis given by $\{\boldsymbol{\omega}(t_i)\otimes\boldsymbol{\omega}(t_j)\}$. For any vector $\mathbf{u} \in \{\boldsymbol{\alpha}(c_i)\otimes\boldsymbol{\alpha}(c_j)\}$ we have

$$\begin{aligned} \lim_{k\to\infty} \mathbf{u}^\top \mathbf{E}^k \mathbf{v} &\le \lim_{k\to\infty} |\mathbf{u}^\top \mathbf{E}^k \mathbf{v}| \\ &\le \sum_{i,j\in[n]} |\beta_{i,j}| \lim_{k\to\infty} |\mathbf{u}^\top \mathbf{E}^k (\boldsymbol{\omega}(t_i)\otimes\boldsymbol{\omega}(t_j))| = 0 \ , \end{aligned}$$

where we used Lemma 1 in the last step. Since this is true for any vector $\mathbf{u}$ in the basis $\{\boldsymbol{\alpha}(c_i)\otimes\boldsymbol{\alpha}(c_j)\}$, we have $\lim_{k\to\infty} \mathbf{E}^k\mathbf{v} = \lim_{k\to\infty} |\lambda|^k \mathbf{v} = \mathbf{0}$, hence $|\lambda| < 1$. This reasoning holds for any eigenvalue of $\mathbf{E}$, hence $\rho(\mathbf{E}) < 1$. □

Lemma 1. *Let $A = \langle \boldsymbol{\alpha}, \boldsymbol{\mathcal{T}}, \{\boldsymbol{\omega}_\sigma\}\rangle$ be a minimal WTA of dimension n computing the strongly convergent function f, and let $\mathbf{E} \in \mathbb{R}^{n^2\times n^2}$ be the Jacobian around $\mathbf{s} = \sum_{t\in\mathfrak{T}} \boldsymbol{\omega}(t)\otimes\boldsymbol{\omega}(t)$ of the mapping $F : \mathbf{v} \to \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \mathbf{v}, \mathbf{v}) + \sum_{\sigma\in\Sigma} \boldsymbol{\omega}_\sigma^{\otimes}$. Then for any $c_1, c_2 \in \mathfrak{C}$ and any $t_1, t_2 \in \mathfrak{T}$ we have $\lim_{k\to\infty} |(\boldsymbol{\alpha}(c_1)\otimes\boldsymbol{\alpha}(c_2))^\top \mathbf{E}^k (\boldsymbol{\omega}(t_1)\otimes\boldsymbol{\omega}(t_2))| = 0$.*

Proof. Let $\boldsymbol{\Xi}^{\otimes} : \mathfrak{C} \to \mathbb{R}^{n^2\times n^2}$ be the context mapping associated with the WTA $A^{\otimes}$; i.e. $\boldsymbol{\Xi}^{\otimes} = \boldsymbol{\Xi}_{A^{\otimes}}$. We start by proving by induction on $\mathrm{drop}(c)$ that $\boldsymbol{\Xi}^{\otimes}(c) = \boldsymbol{\Xi}(c)\otimes\boldsymbol{\Xi}(c)$ for all $c \in \mathfrak{C}$. Let $\mathfrak{C}^d$ denote the set of contexts $c \in \mathfrak{C}$ with $\mathrm{drop}(c) = d$. The statement is trivial for $c \in \mathfrak{C}^0$. Assume the statement is true for all naturals up to $d-1$ and let $c = (t, c') \in \mathfrak{C}^d$ for some $t \in \mathfrak{T}$ and $c' \in \mathfrak{C}^{d-1}$. Then using our inductive hypothesis we have that

$$\begin{aligned} \boldsymbol{\Xi}^{\otimes}(c) &= \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}_{n^2}, \boldsymbol{\omega}(t)\otimes\boldsymbol{\omega}(t), \boldsymbol{\Xi}(c')\otimes\boldsymbol{\Xi}(c')) \\ &= \boldsymbol{\mathcal{T}}(\mathbf{I}_n, \boldsymbol{\omega}(t), \boldsymbol{\Xi}(c')) \otimes \boldsymbol{\mathcal{T}}(\mathbf{I}_n, \boldsymbol{\omega}(t), \boldsymbol{\Xi}(c')) \\ &= \boldsymbol{\Xi}(c)\otimes\boldsymbol{\Xi}(c) \ . \end{aligned}$$

The case $c = (c', t)$ follows from an identical argument.

Next we use the multi-linearity of F to expand $F(\mathbf{s}+\mathbf{h})$ for a vector $\mathbf{h} \in \mathbb{R}^{n^2}$. Keeping the terms that are linear in $\mathbf{h}$ we obtain that $\mathbf{E} = \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \mathbf{s}, \mathbf{I}) + \boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I}, \mathbf{I}, \mathbf{s})$. It follows that $\mathbf{E} = \sum_{c\in\mathfrak{C}^1} \boldsymbol{\Xi}^{\otimes}(c)$, and it can be shown by induction on k that $\mathbf{E}^k = \sum_{c\in\mathfrak{C}^k} \boldsymbol{\Xi}^{\otimes}(c)$.

Writing $d_c = \min(\mathrm{drop}(c_1), \mathrm{drop}(c_2))$ and $d_t =$

$\min(\mathrm{depth}(t_1), \mathrm{depth}(t_2))$, we can see that

$$\begin{aligned}
&\left|(\boldsymbol{\alpha}(c_1)\otimes\boldsymbol{\alpha}(c_2))^\top \mathbf{E}^k(\boldsymbol{\omega}(t_1)\otimes\boldsymbol{\omega}(t_2))\right| \\
&\quad= \left|\sum_{c\in\mathfrak{C}^k}(\boldsymbol{\alpha}(c_1)\otimes\boldsymbol{\alpha}(c_2))^\top \boldsymbol{\Xi}^{\otimes}(c)(\boldsymbol{\omega}(t_1)\otimes\boldsymbol{\omega}(t_2))\right| \\
&\quad= \left|\sum_{c\in\mathfrak{C}^k}(\boldsymbol{\alpha}(c_1)^\top\boldsymbol{\Xi}(c)\boldsymbol{\omega}(t_1))\cdot(\boldsymbol{\alpha}(c_2)^\top\boldsymbol{\Xi}(c)\boldsymbol{\omega}(t_2))\right| \\
&\quad= \left|\sum_{c\in\mathfrak{C}^k} f(c_1[c[t_1]])f(c_2[c[t_2]])\right| \\
&\quad\le \left(\sum_{c\in\mathfrak{C}^k}|f(c_1[c[t_1]])|\right)\left(\sum_{c\in\mathfrak{C}^k}|f(c_2[c[t_2]])|\right) \\
&\quad\le \left(\sum_{t\in\mathfrak{T}^{\ge d_c+d_t+k}}|t||f(t)|\right)^2 ,
\end{aligned}$$

which tends to 0 with $k\to\infty$ since f is strongly convergent. To prove the last inequality, check that any tree of the form $t'=c[c'[t]]$ satisfies $\mathrm{depth}(t')\ge \mathrm{drop}(c)+\mathrm{drop}(c')+\mathrm{depth}(t)$, and that for fixed $c\in\mathfrak{C}$ and $t,t'\in\mathfrak{T}$ we have $|\{c'\in\mathfrak{C} : c[c'[t]]=t'\}|\le|t'|$ (indeed, a factorization $t'=c[c'[t]]$ is fixed once the root of t is chosen in t', which can be done in at most $|t'|$ different ways). □

4 Proof of Theorem 6

Theorem. *There exists $0<\rho<1$ such that after k iterations in Algorithm 2, the approximations $\hat{\mathbf{G}}_{\mathfrak{C}}$ and $\hat{\mathbf{G}}_{\mathfrak{T}}$ satisfy $\|\mathbf{G}_{\mathfrak{C}}-\hat{\mathbf{G}}_{\mathfrak{C}}\|_F\le\mathcal{O}(\rho^k)$ and $\|\mathbf{G}_{\mathfrak{T}}-\hat{\mathbf{G}}_{\mathfrak{T}}\|_F\le\mathcal{O}(\rho^k)$.*

Proof. The result for the Gram matrix $\mathbf{G}_{\mathfrak{T}}$ directly follows from Theorem 5. We now show how the error in the approximation of $\mathbf{G}_{\mathfrak{T}}=\mathrm{reshape}(\mathbf{s},n\times n)$ affects the approximation of $\mathbf{q}=(\boldsymbol{\alpha}^{\otimes})^\top(\mathbf{I}-\mathbf{E})^{-1}=\mathrm{vec}(\mathbf{G}_{\mathfrak{C}})$. Let $\hat{\mathbf{s}}\in\mathbb{R}^n$ be such that $\|\mathbf{s}-\hat{\mathbf{s}}\|\le\varepsilon$, let $\hat{\mathbf{E}}=\boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I},\hat{\mathbf{s}},\mathbf{I})+\boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I},\mathbf{I},\hat{\mathbf{s}})$ and let $\mathbf{q}=(\boldsymbol{\alpha}^{\otimes})^\top(\mathbf{I}-\hat{\mathbf{E}})^{-1}$. We first bound the distance between $\mathbf{E}$ and $\hat{\mathbf{E}}$. We have

$$\begin{aligned}
\|\mathbf{E}-\hat{\mathbf{E}}\|_F &= \|\boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I},\mathbf{s}-\hat{\mathbf{s}},\mathbf{I})+\boldsymbol{\mathcal{T}}^{\otimes}(\mathbf{I},\mathbf{I},\mathbf{s}-\hat{\mathbf{s}})\|_F \\
&\le 2\|\boldsymbol{\mathcal{T}}^{\otimes}\|_F\|\mathbf{s}-\hat{\mathbf{s}}\| \\
&= \mathcal{O}(\varepsilon) ,
\end{aligned}$$

where we used the bounds $\|\boldsymbol{\mathcal{T}}(\mathbf{I},\mathbf{I},\mathbf{v})\|_F\le\|\boldsymbol{\mathcal{T}}\|_F\|\mathbf{v}\|$ and $\|\boldsymbol{\mathcal{T}}(\mathbf{I},\mathbf{v},\mathbf{I})\|_F\le\|\boldsymbol{\mathcal{T}}\|_F\|\mathbf{v}\|$.

Let $\delta=\|\mathbf{E}-\hat{\mathbf{E}}\|$ and let σ be the smallest nonzero eigenvalue of the matrix $\mathbf{I}-\mathbf{E}$. It follows from [3, Equation (7.2)] that if $\delta<\sigma$ then $\|(\mathbf{I}-\mathbf{E})^{-1}-(\mathbf{I}-\hat{\mathbf{E}})^{-1}\|\le\delta/(\sigma(\sigma-\delta))$. Since $\delta=\mathcal{O}(\varepsilon)$ from our previous bound, the condition $\delta\le\sigma/2$ will be eventually satisfied as $\varepsilon\to 0$, in which case we can conclude that

$$\begin{aligned}
\|\mathbf{G}_{\mathfrak{C}}-\hat{\mathbf{G}}_{\mathfrak{C}}\|_F &= \|\mathbf{q}-\hat{\mathbf{q}}\| \\
&\le \|(\mathbf{I}-\mathbf{E})^{-1}-(\mathbf{I}-\hat{\mathbf{E}})^{-1}\|\|\boldsymbol{\alpha}^{\otimes}\| \\
&\le \frac{2\delta}{\sigma^2}\|\boldsymbol{\alpha}^{\otimes}\| \\
&= \mathcal{O}(\varepsilon) .
\end{aligned}$$

□

5 Proof of Theorem 4

Let $A=\langle\boldsymbol{\alpha},\boldsymbol{\mathcal{T}},\{\boldsymbol{\omega}_\sigma\}_{\sigma\in\Sigma}\rangle$ be a SVTA with n states realizing a function f and let $\mathfrak{s}_1\ge\mathfrak{s}_2\ge\cdots\ge\mathfrak{s}_n$ be the singular values of the Hankel matrix $\mathbf{H}_f$.

Theorem 4 relies on the following lemma, which explores the consequences that the fixed-point equations used to compute $\mathbf{G}_{\mathfrak{T}}$ and $\mathbf{G}_{\mathfrak{C}}$ have for an SVTA.

Lemma 2. *For all $i\in[n]$, the following hold:*

1. $\mathfrak{s}_i=\sum_{\sigma\in\Sigma}\boldsymbol{\omega}_\sigma(i)^2+\sum_{j,k=1}^n\boldsymbol{\mathcal{T}}(i,j,k)^2\mathfrak{s}_j\mathfrak{s}_k$,
2. $\mathfrak{s}_i=\boldsymbol{\alpha}(j)^2+\sum_{j,k=1}^n(\boldsymbol{\mathcal{T}}(j,i,k)^2+\boldsymbol{\mathcal{T}}(j,k,i)^2)\mathfrak{s}_j\mathfrak{s}_k$.

Proof. Let $\mathbf{G}_{\mathfrak{T}}$ and $\mathbf{G}_{\mathfrak{C}}$ be the Gram matrices associated with the rank factorization of $\mathbf{H}_f$. Since A is a SVTA we have $\mathbf{G}_{\mathfrak{T}}=\mathbf{G}_{\mathfrak{C}}=\mathbf{D}$ where $\mathbf{D}=diag(\mathfrak{s}_1,\cdots,\mathfrak{s}_n)$ is a diagonal matrix with the Hankel singular values on the diagonal. The first equality then follows from the following fixed point characterization of $\mathbf{G}_{\mathfrak{T}}$:

$$\begin{aligned}
\mathbf{G}_{\mathfrak{T}} &= \sum_{t\in\mathfrak{T}}\boldsymbol{\omega}(t)\boldsymbol{\omega}(t)^\top \\
&= \sum_{\sigma\in\Sigma}\boldsymbol{\omega}_\sigma\boldsymbol{\omega}_\sigma^\top \\
&\quad+\sum_{t_1,t_2\in\mathfrak{T}}\boldsymbol{\mathcal{T}}(\mathbf{I},\boldsymbol{\omega}(t_1),\boldsymbol{\omega}(t_2))\boldsymbol{\mathcal{T}}(\mathbf{I},\boldsymbol{\omega}(t_1),\boldsymbol{\omega}(t_2))^\top \\
&= \sum_{\sigma\in\Sigma}\boldsymbol{\omega}_\sigma\boldsymbol{\omega}_\sigma^\top+\mathbf{T}_{(1)}(\mathbf{G}_{\mathfrak{T}}\otimes\mathbf{G}_{\mathfrak{T}})\mathbf{T}_{(1)}^\top ,
\end{aligned}$$

(where $\mathbf{T}_(i)$ denotes the matricization of the tensor $\boldsymbol{\mathcal{T}}$ along the ith mode). The second equality follows from the following fixed point characterization of $\mathbf{G}_{\mathfrak{C}}$:

$$\begin{aligned}\mathbf{G}_{\mathfrak{C}} &= \sum_{c \in \mathfrak{C}} \boldsymbol{\alpha}(c)\boldsymbol{\alpha}(c)^\top \\ &= \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \\ &+ \sum_{c\in\mathfrak{C}, t\in\mathfrak{T}} \boldsymbol{\mathcal{T}}(\boldsymbol{\alpha}(c), \boldsymbol{\omega}(t), \mathbf{I})\boldsymbol{\mathcal{T}}(\boldsymbol{\alpha}(c), \boldsymbol{\omega}(t), \mathbf{I})^\top \\ &+ \sum_{c\in\mathfrak{C}, t\in\mathfrak{T}} \boldsymbol{\mathcal{T}}(\boldsymbol{\alpha}(c), \mathbf{I}, \boldsymbol{\omega}(t))\boldsymbol{\mathcal{T}}(\boldsymbol{\alpha}(c), \mathbf{I}, \boldsymbol{\omega}(t))^\top \\ &= \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \\ &+ \mathbf{T}_{(2)}(\mathbf{G}_{\mathfrak{C}} \otimes \mathbf{G}_{\mathfrak{T}})\mathbf{T}_{(2)}^\top \\ &+ \mathbf{T}_{(3)}(\mathbf{G}_{\mathfrak{C}} \otimes \mathbf{G}_{\mathfrak{T}})\mathbf{T}_{(3)}^\top \ .\end{aligned}$$

□

Theorem. *For any $t \in \mathfrak{T}$, $c \in \mathfrak{C}$ and $i, j, k \in [n]$ the following hold:*

- $|\boldsymbol{\omega}(t)_i| \leq \sqrt{\mathfrak{s}_i}$,
- $|\boldsymbol{\alpha}(c)_i| \leq \sqrt{\mathfrak{s}_i}$, *and*
- $|\boldsymbol{\mathcal{T}}(i,j,k)| \leq \min\{\frac{\sqrt{\mathfrak{s}_i}}{\sqrt{\mathfrak{s}_j}\sqrt{\mathfrak{s}_k}}, \frac{\sqrt{\mathfrak{s}_j}}{\sqrt{\mathfrak{s}_i}\sqrt{\mathfrak{s}_k}}, \frac{\sqrt{\mathfrak{s}_k}}{\sqrt{\mathfrak{s}_i}\sqrt{\mathfrak{s}_j}}\}$.

Proof. The third point is a direct consequence of the previous Lemma. For the first point, let $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the SVD of $\mathbf{H}_f$. Since A is a SVTA we have

$$\boldsymbol{\omega}(t)_i^2 = (\mathbf{D}^{1/2}\mathbf{V}^\top)_{i,t}^2 = \mathfrak{s}_i \mathbf{V}(t,i)^2$$

and since the rows of $\mathbf{V}$ are orthonormal we have $\mathbf{V}(t,i)^2 \leq 1$.

The inequality for contexts is proved similarly by reasoning on the rows of $\mathbf{U}\mathbf{D}^{1/2}$. □

References

[1] R. Bailly, A. Habrard, and F. Denis. A spectral approach for probabilistic grammatical inference on trees. In *Proceedings of ALT*, 2010.

[2] J. B. Conway. *A course in functional analysis.* Springer, 1990.

[3] L. El Ghaoui. Inversion error, condition number, and approximate inverses of uncertain matrices. *Linear algebra and its applications*, 2002.

[4] J. M. Ortega. *Numerical analysis: a second course.* Siam, 1990.