Low-Rank Approximation of Weighted Tree Automaton
Guillaume Rabusseau (Aix-Marseille University) | Borja Balle (Lancaster University) | Shay B. Cohen (University of Edinburgh)

Overview
- Weighted Tree Automata: model subsuming PCFG and L-PCFG.
- Aim: model reduction to speed-up inference (e.g. parsing).
- Method: similar to PCA for weighted context-free grammar.
- Iterative algorithm that resembles the power method for SVD.

Tensor Networks

Weighted Tree Automata
- A weighted tree automaton (WTA) is a tuple \( \langle \alpha, \mathcal{T}, (\omega_\sigma)_{\sigma \in \Sigma} \rangle \) where
  - \( \alpha \in \mathbb{R}^n \): vector of initial weights
  - \( \mathcal{T} \in \mathbb{R}^{n \times n \times n} \): tensor of transition weights
  - \( \omega_\sigma \in \mathbb{R}^{n \times n} \): vector of final weights associated with \( \sigma \in \Sigma \)
- Bottom up computation. \( \omega_A \) maps trees in \( \mathcal{T} \) to vectors in \( \mathbb{R}^n \):

\[
\omega_A(w_\alpha, w_\beta, w_\gamma) = \mathcal{T}(T_\alpha, T_\beta, T_\gamma)
\]

- The WTA \( \mathcal{A} \) compute the function \( f_{\mathcal{A}} : t \mapsto \alpha^T \omega_{\mathcal{A}}(t) \in \mathbb{R} \).
- Contexts are trees with a hole:
  - \( c = \)
  - \( t = \)
  - \( a \)
  - \( b \)
  - \( d \)
  - \( c[t] = \)

\[
\alpha_A \text{ maps contexts in } C_\Sigma \text{ to vectors in } \mathbb{R}^n:
\]

\[
\alpha_A(a, b) = \mathcal{T}_\alpha(a, b).
\]

- For any tree context \( c \in C_\Sigma \) and tree \( t \in \mathcal{T}_\Sigma \):
  \[
  f_{\mathcal{A}}(c[t]) = \alpha_A(c)^T \omega_{\mathcal{A}}(t).
  \]

Rank Factorization of the Hankel Matrix
- The rank of a function \( f : T_\Sigma \to \mathbb{R} \) is the number of states of the smallest WTA computing \( f \).
- The Hankel matrix \( H_f \in \mathbb{R}^{C_\Sigma \times T_\Sigma} \) is the bi-infinite matrix defined by
  \[
  H_f(c, t) = f(c[t]) \text{ for any } c \in C_\Sigma, \ t \in T_\Sigma.
  \]

Theorem (Bozapalidis and Louscou-Bozapalidou [1983]).
For any \( f : T_\Sigma \to \mathbb{R} \) we have \( \text{rank}(f) = \text{rank}(H_f) \).

- A WTA \( \mathcal{A} \) with \( n \) states induces a rank \( n \) factorization \( H_f = P_A S_A \):
  - \( P_A \in \mathbb{R}^{C_\Sigma \times n} \) defined by \( P_A(c,:) = \alpha_A(c) \)
  - \( S_A \in \mathbb{R}^{n \times T_\Sigma} \) defined by \( S_A(c,:) = \omega_A(c) \).
- One-to-one correspondence between rank factorizations of \( H_f \) and WTAs computing \( f \):

\[
\Rightarrow \quad \text{Iterative algorithm converging exponentially fast:}
\]

\[
\text{Theorem. There exists } 0 < \rho < 1 \text{ such that after } k \text{ iterations, the approximations } G_{C, \Sigma} \text{ and } G_{T, \Sigma} \text{ satisfy } \|G_{C, \Sigma} - G_{C, \Sigma}^{(k)}\| \leq O(\rho^k) \text{ and } \|G_{T, \Sigma} - G_{T, \Sigma}^{(k)}\| \leq O(\rho^k).
\]

Low-Rank Approximation of WTA
Problem: Given a WTA \( \mathcal{A} \) with \( n \) states, find a WTA with \( \hat{n} < n \) states that is close to \( \mathcal{A} \).

Idea: (following Balle et al. [2015] in the string case)
- Find a WTA \( \hat{\mathcal{A}} \) such that the rank-\( n \) factorization \( H_f = P_{\hat{\mathcal{A}}} S_{\hat{\mathcal{A}}} \) coincides with the SVD of \( H_f = U \Sigma V^T \):
  \[
  P_{\hat{\mathcal{A}}} = U^T \sqrt{\Sigma_{1: \hat{n}}} \quad \text{and} \quad S_{\hat{\mathcal{A}}} = V \sqrt{\Sigma_{1: \hat{n}}}. \tag{1}
  \]
- The \( i \)-th state of \( \hat{\mathcal{A}} \) corresponds to the \( i \)-th singular value of \( H_f \).
- Remove the \( n - \hat{n} \) states corresponding to the smallest singular values.

We call a WTA satisfying (1) a singular value tree automaton (SVTA).

Computing the SVTA
Given a WTA \( \mathcal{A} = (\alpha, \mathcal{T}, (\omega_\sigma)_{\sigma \in \Sigma}) \) how to find an SVTA computing the same function?
- Obvious way: compute the SVD of the bi-infinite matrix \( H_f = \mathcal{P} \).
- In practice:
  - Gram matrices \( \mathcal{G}_T = U \mathcal{P} \mathcal{G} R \mathcal{P} \mathcal{G} T \in \mathbb{R}^{n \times n} \) and \( \mathcal{G}_T = \mathcal{G}_T \mathcal{G}_T^T \in \mathbb{R}^{n \times n} \).
  - Compute \( \mathcal{Q} \) from the eigendecompositions of \( \mathcal{G}_T \) and \( \mathcal{G}_T^T \).
- The WTA \( \hat{\mathcal{A}} = (Q^T \alpha, \mathcal{T}(Q^{-T} \mathcal{Q}, \mathcal{Q}), (Q^{-T} \omega_r)) \) is a SVTA computing \( f \).

Computing the Gram Matrices
- Leverage the fixed point equations:

\[
\mathcal{G}_T = \sum_{\sigma \in \Sigma} \omega_\sigma \omega_\sigma^T + T(1) (\mathcal{G}_T \mathcal{G}_T^T) T(1) \quad \text{and} \quad \mathcal{G}_C = \alpha \alpha^T + T(2) (\mathcal{G}_C \mathcal{G}_T T(2) + T(3) (\mathcal{G}_C \mathcal{G}_T \mathcal{G}_T T(3)^T)
\]

Experiments
- PCFG with \( n = 211 \) nonterminals learnt from the annotated corpus of German newspaper texts NERGA [Sklut et al., 1997].
- Comparison with Cohen et al. [2013] (decomposition of the tensors of the PCFG).
- Number of parameters: \( \hat{n}^2 \) for a WTA with \( \hat{n} \) states, and \( 3 \hat{n} n \) when the tensor \( \mathcal{T} \) is approximated with a tensor of CP-rank \( R \).

References