Unsupervised Spectral Learning of WCFG as Low-rank Matrix Completion

Raphaël Bailly† Xavier Carreras† Franco M. Luque‡ Ariadna Quattoni†
† Universitat Politècnica de Catalunya ‡ Universidad Nacional de Córdoba y CONICET

--- Abstract ---
- **Goal**: induce a WCFG from observed strings (e.g. language modeling)
- **Contribution**: a spectral method for unsupervised WCFG learning

--- Ingredients ---
- Convex optimization of a Hankel matrix for WCFG
- Linear constraints to characterize WCFG Hankel matrices
- Low-rank objective, derived from the spectral approach

--- Weighted Context-Free Grammars ---
- **Definition**: WCFG $G = \langle \Sigma, \Pi, \alpha, (\beta_i)_i, A \rangle$
  - Alphabet $\Sigma$, $n$ states (states are non-terminals)
  - Initial vector $\alpha_0 \in \mathbb{R}^n$
  - Terminal vectors $\beta_a \in \mathbb{R}^n$ for $a \in \Sigma$
  - Bilinear operator $A : \mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$
- **Grammar function** $g_G : \Sigma^* \rightarrow \mathbb{R}$
  - $g_G(x) = \alpha_0 \beta_0(x)$
- **Inside function** $\beta : \Sigma^+ \rightarrow \mathbb{R}^n$
  - $\beta(G(x)) = \sum_{x \in \Sigma^+} A(\beta(G(x_1)) \otimes \beta(G(x_2)))$
- **Outside function** $\alpha : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}^n$
  - $\alpha(G(x \cdot z)) = \sum_{x \in \Sigma^*} \sum_{z \in \Sigma^*} A(I_n \otimes \beta(G(z_1)))$

--- The Spectral Method ---
1. Compute Hankel matrix $H$ using training data
   - **Supervised**: training data contains derivations, count events
   - **Unsupervised** (this paper): induce $H$ from plain strings
2. Compute the SVD of $H_{\text{Grammar}} = U \Lambda U^T$
3. Create a rank-$n$ factorization of $H_{\text{Grammar}} \approx F_n B_n$
   - $F_n = U_n A_n$
   - $B_n = V_n$
4. Compute $G$:
   - $\alpha_G = H_{\text{Grammar}} B_n$
   - $\beta_G = F_n H_{\text{Grammar}}^T$
   - $A = F_n H_{\text{Grammar}}^T B_n \otimes B_n$

--- Hankel Matrices for WCFG ---
- **Definition**: $H(G, (\theta, i)) \in \mathbb{R}^{n \times n}$
  - $H(G, (\theta, i)) = g(G(x_i))$
- **Inside strings** $I$
  - $\alpha_I = H (\theta, i)$
  - $\beta_I = F \beta_G$
- **Outside contexts** $O$
  - $\alpha_O = \alpha_I \cdot \beta_I$

--- Hankel Induction ---
- Observable statistics (for full strings):
  - $H(x, (\theta, i)) = p(x)$

- **WCGP constraints**:
  - Hankel constraints
    - $H(x, (y_1, y_2)) = H(x, y_1, y_2) = H(x, y_2, y_1) = H(y_1, x, y_2) = H(y_2, x, y_1)$
- **Inside constraints**
  - $H(x, (\theta, i)) = \sum_{x \in \Sigma^+} H(x, (\theta, i_2))$
- **Outside constraints**
  - $H(x, z, i) = \sum_{z_1 \in \Sigma^+} H(x, i_2, z, z_1) + \sum_{z_2 \in \Sigma^+} H(x, z_1, z_2, i)$
- **Property**: any $H$ satisfying constraints corresponds to a WCFG

--- Convex Optimization ---
- **minimize** $H$
  - **subject to** $|O \cdot \ell - z|_2 \leq \mu$
  - $K \cdot \ell = 0$
  - $|H|_2 \leq 1$

--- Synthetic Experiments ---
- Random synthetic targets
- Fixed inside-outside patterns in Hankel
- Comparison to EM and a supervised spectral method

--- Experiments with Dyck Languages ---
- The target grammar is:
  - $S \rightarrow SS \ (0.2) \ | a \ b \ c \ (0.4) \ | a \ b \ (0.4)$
  - **e.g.** $p(aab) = 0.16$, $p(aaa) = 0$
- We vary the size of the Hankel
- Comparison to spectral algorithm for Weighted Automata

--- On WSJ-10 data ---

--- Acknowledgements ---
This work was supported by a Google Research Award, and by projects XLike (FP7-288342), ERA-Net CHISTERA VISIN, TACARDI (TIN2012-38523-C02-02), BASMATI (TIN2011-27479-C04-03), SGR-GPLN (2009-SGR-1082) and SGR-LARCA (2009-SGR-1428). Xavier Carreras was supported by the Ramón y Cajal program of the Spanish Government (RYC-2008-02233). Franco M. Luque was supported by a postdoctoral fellowship from CONICET of the Argentinian Government.