Identifiability and Unmixing of Latent Parse Trees

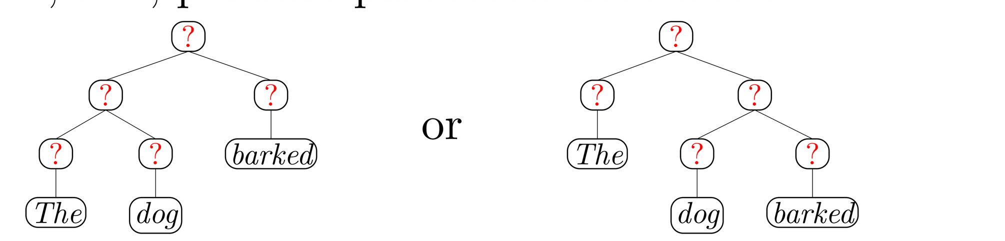
Daniel Hsu — Microsoft Research
Sham M. Kakade — Microsoft Research
Percy Liang — Stanford University

Overview

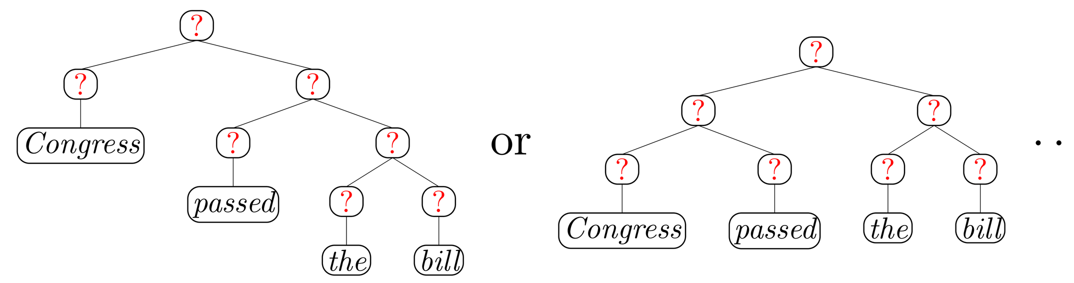
Model: $\mathbb{P}_\theta(x, z)$ over parse trees z and sentences x

Goal: given n sentences $x^{(1)}, \ldots, x^{(n)}$, produce parameter estimate $\hat{\theta}$

The dog barked $\Rightarrow$



Congress passed the bill $\Rightarrow$



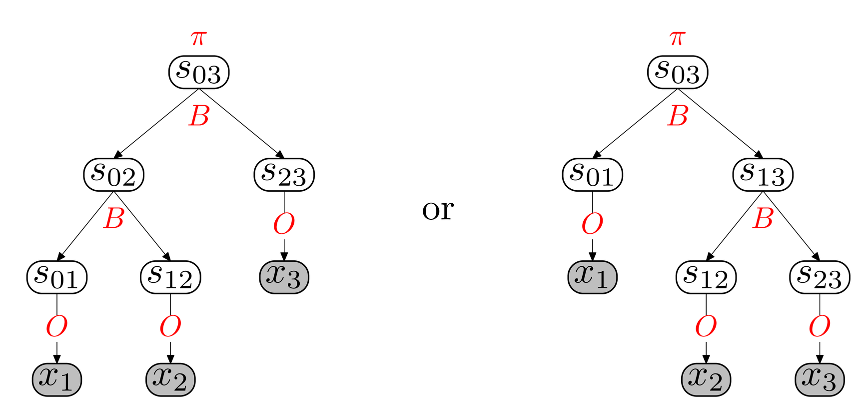
Challenge: tree topology is unobserved and varies across sentences

Two questions:

- Which model families $\mathbb{P}_\theta(x, z)$ are identifiable?
  Our result: **PCFG is not identifiable.**
- How to estimate parameters without local optima issues?
  Our result: new **unmixing** technique works for restricted PCFGs.

Probabilistic Context-Free Grammars (PCFG)

For $L = 3$ words:



Topology$(z) = 1$ — Topology$(z) = 2$

Parameters $\theta = (\pi, B, O)$:

Initial $\pi \in \mathbb{R}^k$: probability of initial state

Binary productions $B \in \mathbb{R}^{k^2 \times k}$: probability of children given parent state

Emissions $O \in \mathbb{R}^{d \times k}$: probability of word given state

Latent parse tree $z = (\text{Topology}(z), \text{latent states } \{s_{[i:j]}\})$

$$\mathbb{P}_\theta(x, z) = |\,\text{Topologies}\,|^{-1} \pi^\top s_{[0:L]} \prod_{[i:m],[m:j]} (s_{[i:m]} \otimes_k s_{[m:j]})^\top B s_{[i:j]} \prod_i x_i^\top O s_{[i-1:i]}$$

Assumption: uniform distribution over binary branching trees

Parameter estimation

Standard approach (maximum likelihood):

Estimator: $\hat{\theta} = \arg\max_\theta \sum_{i=1}^n \log \mathbb{P}_\theta(x)$

Intractable, EM algorithm gets stuck in local optima [Lari & Young, 1990]

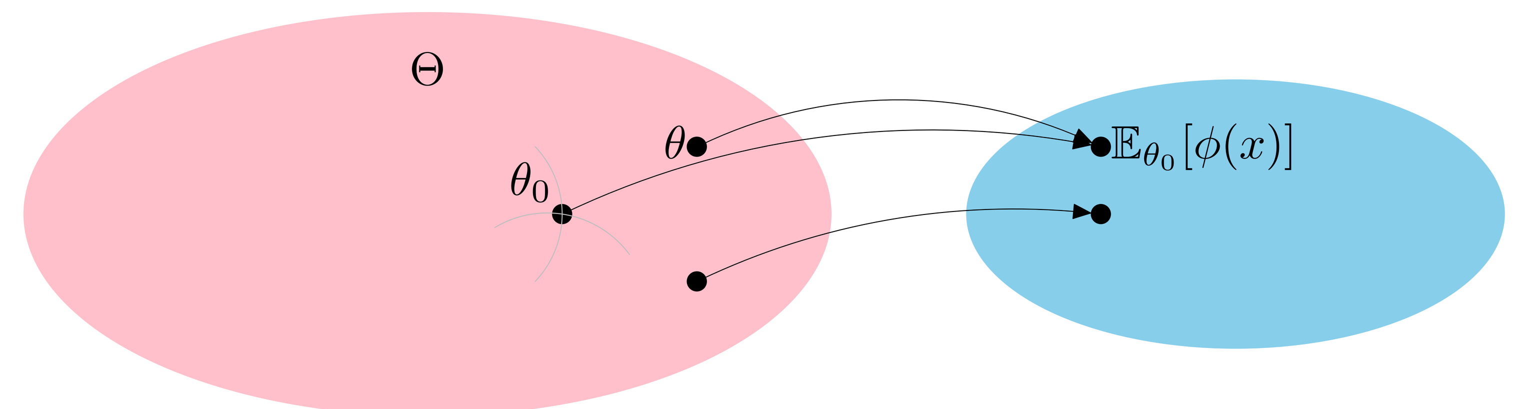
Our strategy (**method of moments**):

Moment function: $\phi(x) \in \mathbb{R}^m$ (e.g., $\phi_{12}(x) = x_1 x_2^\top \in \mathbb{R}^{d \times d}$)

Estimator: $\hat{\theta}$ such that $\mathbb{E}_{\hat{\theta}}[\phi(x)] = \frac{1}{n} \sum_{i=1}^n \phi(x^{(i)})$

Identifiability from moments

Definition (global identifiability): model family $\Theta \subset [0, 1]^p$ is identifiable from a moment function $\phi(x)$ if $S_\Theta(\theta_0) = \{\theta \in \Theta : \mathbb{E}_\theta[\phi(x)] = \mathbb{E}_{\theta_0}[\phi(x)]\}$ is finite for almost every $\theta_0 \in \Theta$; that is: given moments $\mathbb{E}_\theta[\phi(x)]$, possible to recover parameters θ up to a finite equivalence class (e.g., permutation of states)?



General identifiability checker:

1. Choose a **single** $\tilde{\theta} \in \Theta$ uniformly at random.
2. Compute Jacobian matrix $J(\tilde{\theta}) = \frac{\partial \mathbb{E}_\theta[\phi(x)]}{\partial \theta}\big|_{\theta = \tilde{\theta}} \in \mathbb{R}^{m \times p}$.
3. Return identifiable iff $J(\tilde{\theta})$ is full rank.

Theorem: identifiability checker is correct with probability 1.

Significance:

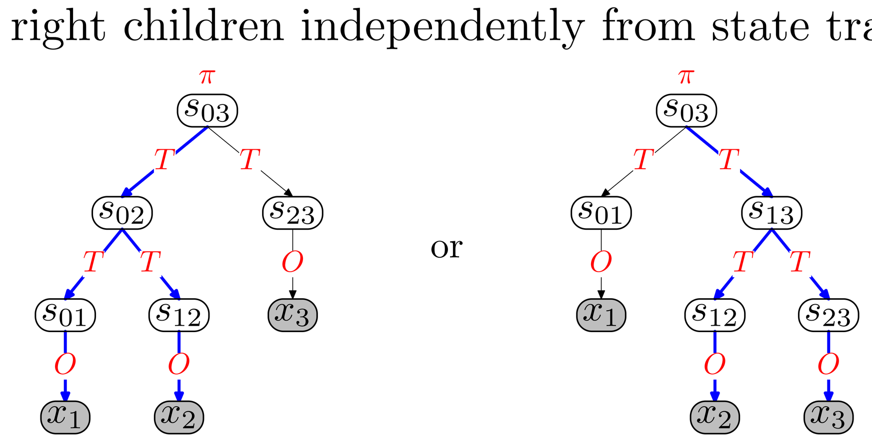
Test random point (cheap, local information) $\Rightarrow$ identifable? (global property)

Intuition: space is nice because moments are polynomials of parameters

Result: **PCFG is not identifiable** from any moments $\phi(x)$ and $L \leq 5$.

Restricted PCFG

Generate left and right children independently from state transitions $T \in \mathbb{R}^{k \times k}$



Topology$(z) = 1$ — Topology$(z) = 2$

Decompose [Anandkumar/Hsu/Kakade 2012]

Purpose: used to recover parameters when tree structure is known.

Unknowns (think of these as parameters):

$M_1, M_2 \in \mathbb{R}^{d \times k}$: matrices with full column rank

$D \in \mathbb{R}^k$: diagonal matrix with distinct diagonal entries

Result: can recover M_1 up to scaling/permutation of columns

$$\textbf{eigenvectors}\Big(\underbrace{\boxed{M_1}\,\boxed{D}\,\boxed{M_2^\top}}_{\text{observed}} \cdot \underbrace{\boxed{M_2^{\top -1}}\,\boxed{M_1^{-1}}}_{\text{observed}} \Big) = M_1$$

Note: above is valid for $k = d$; otherwise, project down to k dimensions.

Unmixing

Known tree structure (for $L = 3$ words):

$$\Psi_{2;\eta} = \mathbb{E}[x_1 (x_2^\top \eta) x_3^\top \mid \text{Topology}(z) = 2] = \underbrace{OT}_{M_1} \underbrace{\text{diag}(T^\top O^\top \eta)}_{D} \underbrace{T^\top \text{diag}(\pi) T^\top O^\top}_{M_2^\top}$$

Compute $\Psi_{2;\eta}$ for two different η, apply Decompose to recover $M_1 = OT$.

Apply simple matrix algebra to extract all parameters $\theta = (\pi, T, O)$.

Unknown tree structure (for $L = 3$ words):

Strategy: reduce to the known tree structure case

$$\underbrace{\begin{pmatrix} \mu_{123;\eta} \\ \mu_{132;\eta} \\ \mu_{231;\eta} \end{pmatrix}}_{\text{observed moments } \mu_{*;\eta}} = \underbrace{\begin{pmatrix} 0.5I & 0.5I & 0 \\ 0 & 0.5I & 0.5I \\ 0.5I & 0 & 0.5I \end{pmatrix}}_{\text{mixing matrix } M} \underbrace{\begin{pmatrix} \Psi_{1;\eta} \\ \Psi_{2;\eta} \\ \Psi_{3;\eta} \end{pmatrix}}_{\text{compound parameters } \Psi_{*;\eta}}.$$

Unknown tree structure (general case):

$$\begin{matrix}\text{moments} \\ \mu_{*;\eta}\end{matrix} \Rightarrow \boxed{\begin{matrix}\text{Solve} \\ \text{linear} \\ \text{system}\end{matrix}} \Rightarrow \begin{matrix}\text{compound} \\ \text{parameters} \\ \Psi_{*;\eta} = M^\dagger \mu_{*;\eta}\end{matrix} \Rightarrow \boxed{\text{Decompose}} \Rightarrow \begin{matrix}\text{parameters} \\ \theta\end{matrix}$$

Proposition (unmixing):

If e_j in row space of M, can recover $\Psi_{j;\eta}$.

Call base algorithm on $\Psi_{j;\eta}$ to recover θ.

All operations involve low-order matrix computations.

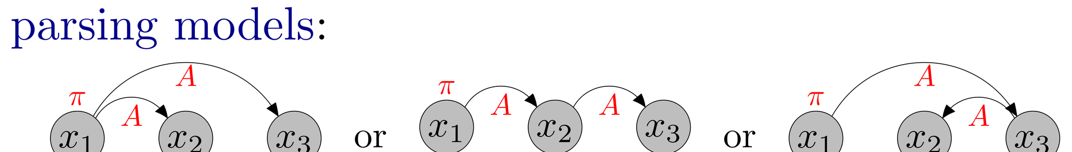
Sample complexity n is polynomial in k, d, L and spectral properties of T, O.

Result: for restricted PCFG, e_2 in row space of M for all L.

Other results

Restricted PCFG	Restricted PCFG (different $T_{\text{left}}, T_{\text{right}}$ transitions)	PCFG
identifiable	identifiable	non-identifiable
unmixing	?	hopeless

Dependency parsing models:



Topology$(z) = 1$ — Topology$(z) = 2$ — Topology$(z) = 3$

Result: identifiable, unmixing works for restricted version

Conclusion

Related work on spectral methods:

HMMs [Hsu/Kakade/Zhang 2009]
Latent tree models with known structure [Parikh/Song/Xing 2011]
Unknown fixed structure [Anandkumar/Chaudhuri/Hsu/Kakade/Song/Zhang 2011]
PCFGs with known tree structure [Cohen/Stratos/Collins/Foster/Ungar 2012]
Recover parameters for HMMs [Anandkumar/Hsu/Kakade 2012]
This work: recover parameters, unknown random structure

Two contributions:

- Identifiability checker: easy method to see if model family identifiable
- Unmixing technique: consistent parameter recovery with random structures