Identifiability and Unmixing of Latent Parse Trees

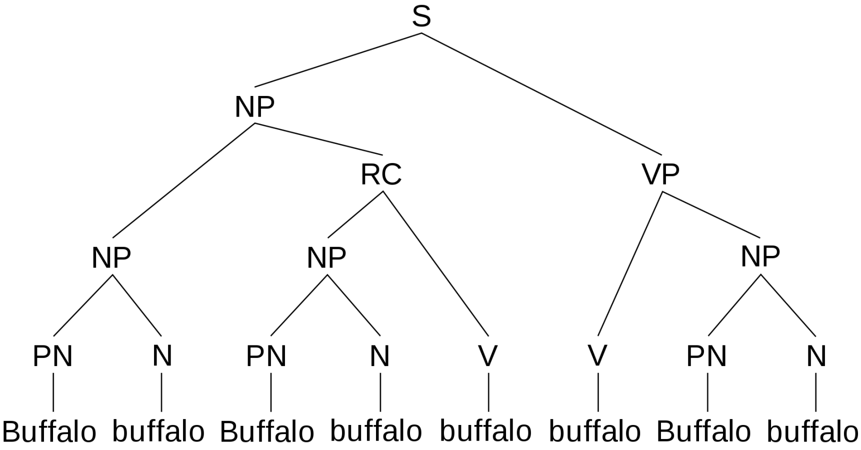
Daniel Hsu, Sham Kakade, Percy Liang

NIPS 2012

Jan Gasthaus

Tea talk
January 8th, 2013

Parsing


S
NP
RC
VP
NP
NP
NP
PN
N
PN
N
V
V
PN
N
Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo

Big Picture

- Generative parsing models define joint distributions $P_\theta(\mathbf{x}, z)$ over sentences $\mathbf{x}$ and their *structure* z.

Big Picture

- Generative parsing models define joint distributions $P_\theta(\mathbf{x}, z)$ over sentences $\mathbf{x}$ and their *structure* z.
- Can we identify θ given only sentences (but *not* their structure, i.e. without supervision)?

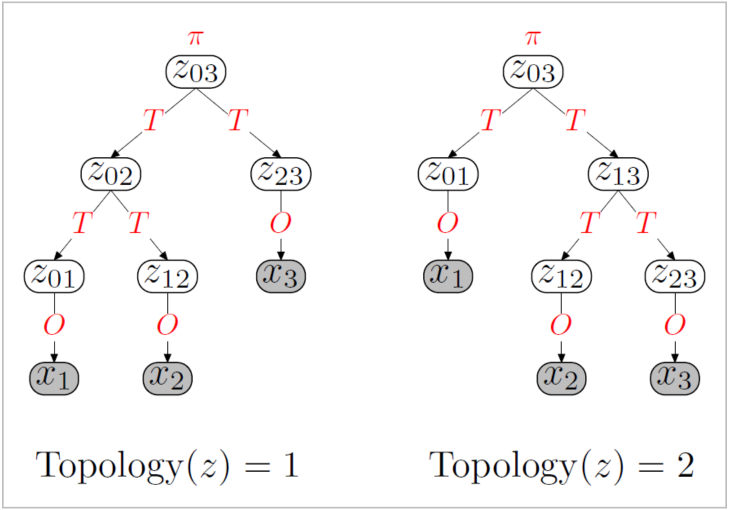
Big Picture

- Generative parsing models define joint distributions $P_\theta(\mathbf{x}, z)$ over sentences **x** and their *structure* z.
- Can we identify θ given only sentences (but *not* their structure, i.e. without supervision)?
- The paper has two parts:
  1. Identifiabilty of several models (PCFGs not identifiable!)

Big Picture

- Generative parsing models define joint distributions $P_\theta(\mathbf{x}, z)$ over sentences **x** and their *structure* z.
- Can we identify θ given only sentences (but *not* their structure, i.e. without supervision)?
- The paper has two parts:
  1. Identifiabilty of several models (PCFGs not identifiable!)
  2. Parameter recovery: unmixing (for restricted PCFGs)

Big Picture

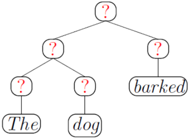


the lady sang

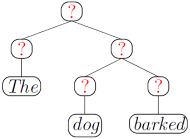
Gatsby likes Bayesians

Big Picture

The dog barked ⇒

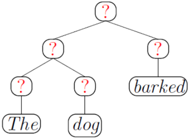


or

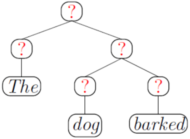


Big Picture

The dog barked $\Rightarrow$



or



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]

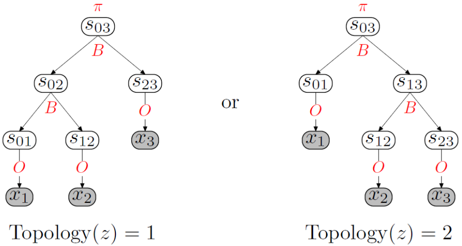
Our strategy (**method of moments**):

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

PCFG model

For $L = 3$ words:



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

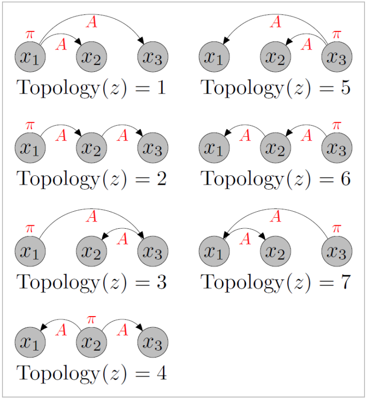
Emissions $O \in \mathbb{R}^{d \times k}$: probability of word given state

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

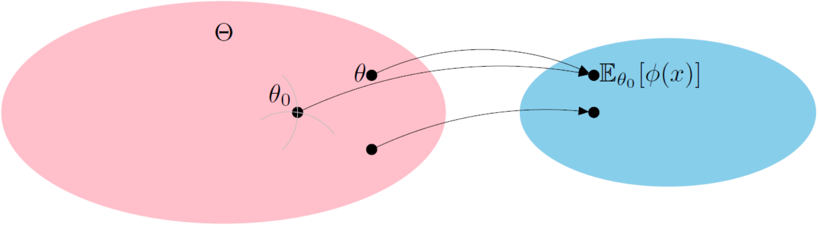
Dependency Grammars



$$\mathbb{P}_\theta(\mathbf{x}, z) = |\,\text{Topologies}\,|^{-1} \pi^\top x_{\text{Root}(z)} \prod_{(i,j)\in z} x_j^\top A_{\text{dir}(i,j)} x_i$$

Identifiability

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)?



Identifiability

- $S_\Theta(\theta_0)$ defined by moment constraints

$$h_{\theta_0}(\theta) = \mu(\theta) - \mu(\theta_0) = 0$$

- Rows of Jacobian of h_{θ_0} are directions of constraint violation

Identifiability

- $S_\Theta(\theta_0)$ defined by moment constraints

$$h_{\theta_0}(\theta) = \mu(\theta) - \mu(\theta_0) = 0$$

- Rows of Jacobian of h_{θ_0} are directions of constraint violation

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 $L \leq 5$.

Identifiability

<table>
<tr><th>Model \ Observation function</th><th>ϕ_{12}</th><th>ϕ_{**}</th><th>ϕ_{123e_1}</th><th>ϕ_{123}</th><th>ϕ_{***e_1}</th><th>ϕ_{***}</th></tr>
<tr><td>PCFG</td><td colspan="6">No, even from ϕ_{all} for $L \in \{3, 4, 5\}$</td></tr>
<tr><td>PCFG-I / PCFG-IE</td><td>No</td><td>Yes iff $L \geq 4$</td><td colspan="4">Yes iff $L \geq 3$</td></tr>
<tr><td>DEP-I</td><td>No</td><td colspan="5">Yes iff $L \geq 3$</td></tr>
<tr><td>DEP-IE / DEP-IES</td><td colspan="6">Yes iff $L \geq 3$</td></tr>
</table>

Figure 2: Local identifiability of parsing models. These findings are given by CHECKIDENTIFIABILITY have the semantics from Theorem 1. These were checked for $d \in \{2, 3, \ldots, 8\}$, $k \in \{2, \ldots, d\}$ (applies only for PCFG models), $L \in \{2, 3, \ldots, 9\}$.

$$\phi_{12}(\mathbf{x}) \stackrel{\text{def}}{=} x_1 \otimes x_2 \qquad \phi_{**}(\mathbf{x}) \stackrel{\text{def}}{=} (x_i \otimes x_j : i, j \in [L])$$

$$\phi_{123}(\mathbf{x}) \stackrel{\text{def}}{=} x_1 \otimes x_2 \otimes x_3 \qquad \phi_{***}(\mathbf{x}) \stackrel{\text{def}}{=} (x_i \otimes x_j \otimes x_k : i, j, k \in [L])$$

$$\phi_{123\eta}(\mathbf{x}) \stackrel{\text{def}}{=} (x_1 \otimes x_2)(\eta^\top x_3) \qquad \phi_{***\eta}(\mathbf{x}) \stackrel{\text{def}}{=} ((x_i \otimes x_j)(\eta^\top x_k) : i, j, k \in [L])$$

$$\phi_{\text{all}}(\mathbf{x}) \stackrel{\text{def}}{=} x_1 \otimes \cdots \otimes x_L$$

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 $M_1 = OT$.

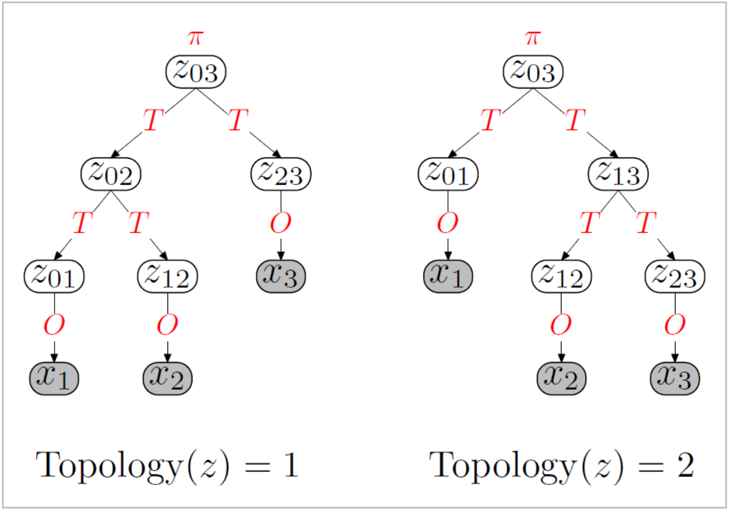
Apply simple matrix algebra to extract all parameters $\theta = (\pi, T, O)$.

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

Strategy: reduce to the known tree structure case

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

Unmixing



π
z03
T
T
z02
z23
T
T
O
z01
z12
x3
O
O
x1
x2
Topology(z) = 1
π
z03
T
T
z01
z13
O
T
T
x1
z12
z23
O
O
x2
x3
Topology(z) = 2

Unmixing

Unknown tree structure (general case):

moments $\mu_{*;\eta}$ $\Rightarrow$ Solve linear system $\Rightarrow$ compound parameters $\Psi_{*;\eta} = M^{\dagger}\mu_{*;\eta}$ $\Rightarrow$ Decompose $\Rightarrow$ parameters θ

Proposition (unmixing):

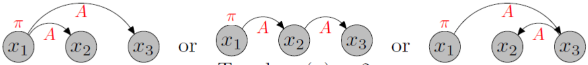
If e_j in row space of M, can recover $\Psi_{j;\eta}$.
Call base algorithm on $\Psi_{j;\eta}$ to recover θ.
All operations involve low-order matrix computations.
Sample complexity n is polynomial in k, d, L and spectral properties of T, O.

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

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

Conclusions

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