Spectral Learning of Refinement HMMs

Karl Stratos[1]

Joint work with Alexander M. Rush[2], Shay B. Cohen[1], and Michael Collins[1]

[1]Columbia University

[2]MIT CSAIL

Spectral Algorithms

Used for estimation of latent-variable models

- Latent-variable PCFGs (Cohen et al., 2012)
- Hidden Markov Models (Hsu et al., 2008)

Basic idea: replace EM with methods based on matrix decompositions, in particular singular value decomposition (SVD)

Guaranteed to learn (unlike EM) under assumptions on singular values in the SVD

Motivating Example: Phoneme Recognition

$a =$	$\ldots$	sil	ao	ao	ao	ao	ao	ao	ao	ow	$\ldots$
$x =$	$\ldots$	353	26	11	12	15	14	17	447	435	$\ldots$

- Label sequence a: phonemes
- Observation sequence x: vector-quantized speech frames
- Task: given a sequence of speech frames, predict the correct sequence of phonemes
- HMMs' strong independence assumption can be limiting

Motivating Example: Phoneme Recognition (cont.)

					↓					
...	sil	ao	ao	ao	ao	ao	ao	ao	ow	...
...	353	2	11	12	15	9	7	900	835	...

Motivating Example: Phoneme Recognition (cont.)

					$\downarrow$					
...	sil	ao	ao	ao	ao^3	ao	ao	ao	ow	...
...	353	2	11	12	15	9	7	900	835	...

ao marked with latent state 3 means...

- The next phoneme is ow

Motivating Example: Phoneme Recognition (cont.)

$$\downarrow$$

...	sil	ao	ao	ao	ao^3	ao	ao	ao	ow	...
...	353	2	11	12	15	9	7	900	835	...

ao marked with latent state 3 means...

- The next phoneme is ow
- The previous phoneme is sil

Motivating Example: Phoneme Recognition (cont.)

					↓					
...	sil	ao	ao	ao	ao^3	ao	ao	ao	ow	...
...	353	2	11	12	15	9	7	900	835	...

ao marked with latent state 3 means...

- The next phoneme is ow
- The previous phoneme is sil
- We are in the middle of the current ao-sequence
- etc.

Motivating Example: Phoneme Recognition (cont.)

...	sil^{1}	ao^{4}	ao^{4}	ao^{4}	ao^{3}	ao^{3}	ao^{1}	ao^{1}	ow^{1}	...
...	353	2	11	12	15	9	7	900	835	...

ao marked with latent state 3 means...

- The next phoneme is ow
- The previous phoneme is sil
- We are in the middle of the current ao-sequence
- etc.

Introduce a latent state for each phoneme

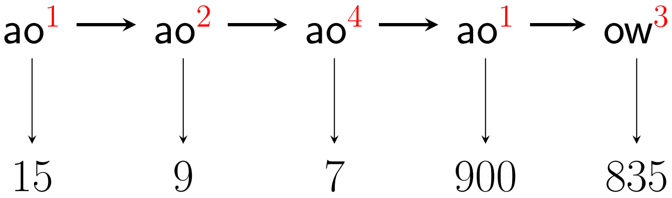
This Work

Spectral algorithm for estimation of Refinement HMMs (R-HMMs)

$$\begin{array}{ccccccccc}
\text{ao}^{1} & \rightarrow & \text{ao}^{2} & \rightarrow & \text{ao}^{4} & \rightarrow & \text{ao}^{1} & \rightarrow & \text{ow}^{3} \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
15 & & 9 & & 7 & & 900 & & 835
\end{array}$$

$$p(15\ 9\ 7\ 900\ 835,\ \text{ao}\ \ \text{ao}\ \ \text{ao}\ \ \text{ao}\ \ \text{ow},\ 1\ 2\ 4\ 1\ 3)$$

Parameter Estimation of Refinement HMM (R-HMM)



$$p(15\ 9\ 7\ 900\ 835,\ \textsf{ao}\ \ \textsf{ao}\ \ \textsf{ao}\ \ \textsf{ao}\ \ \textsf{ow},\ {\color{red}1\ 2\ 4\ 1\ 3})$$
$$= \pi(\textsf{ao}^{\color{red}1}) \times t(\textsf{ao}^{\color{red}2}|\textsf{ao}^{\color{red}1}) \times t(\textsf{ao}^{\color{red}4}|\textsf{ao}^{\color{red}2}) \times t(\textsf{ao}^{\color{red}1}|\textsf{ao}^{\color{red}4}) \times t(\textsf{ow}^{\color{red}3}|\textsf{ao}^{\color{red}1})$$
$$\times\, o(15|\textsf{ao}^{\color{red}1}) \times o(9|\textsf{ao}^{\color{red}2}) \times o(7|\textsf{ao}^{\color{red}4}) \times o(900|\textsf{ao}^{\color{red}1}) \times o(835|\textsf{ow}^{\color{red}3})$$

Goal: **estimate the R-HMM parameters π, t, and o** from samples without observing the latent states

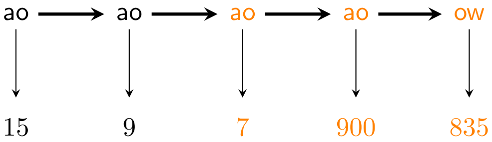
Overview

Introduction

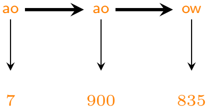
Sketch of the Spectral Algorithm for Parameter Estimation

Experiments

Feature Functions $\phi, \psi, \xi, \upsilon$

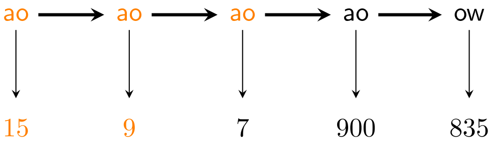


Future f

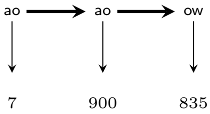


$\phi(f) = [0, 1, 0, \ldots, 0, 0]$

Feature Functions $\phi, \psi, \xi, \upsilon$

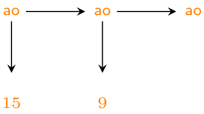


Future f



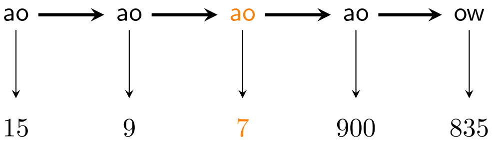
$\phi(f) = [0, 1, 0, \ldots, 0, 0]$

Past p

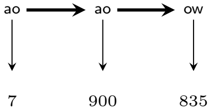


$\psi(p) = [0, 0, 1, \ldots, 0, 1]$

Feature Functions $\phi, \psi, \xi, \upsilon$

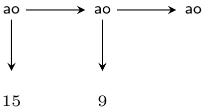


Future f



$\phi(f) = [0, 1, 0, \ldots, 0, 0]$

Past p

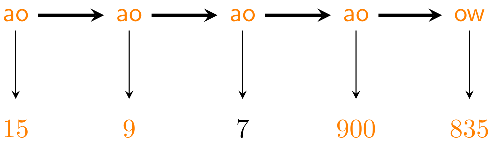


$\psi(p) = [0, 0, 1, \ldots, 0, 1]$

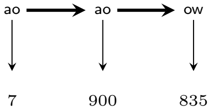
Present r

$\xi(r) = [0, 0, 0, \ldots, 0, 1]$

Feature Functions $\phi, \psi, \xi, \upsilon$

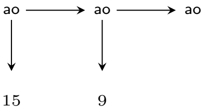


Future f



$\phi(f) = [0, 1, 0, \ldots, 0, 0]$

Past p



$\psi(p) = [0, 0, 1, \ldots, 0, 1]$

Present r

$\xi(r) = [0, 0, 0, \ldots, 0, 1]$

Destiny d

$\upsilon(d) = [1, 1, 0, \ldots, 1, 1]$

Sketch of the Algorithm

Input: skeletal sequences $\{(x^{(i)}, a^{(i)})\}$, number of hidden states m, feature functions $\phi, \psi, \xi, \upsilon$

1. Estimate matrices
   - $\hat{\Omega}_1^a$ = coocurrence b/t future features ϕ and past features ψ
   - $\hat{\Omega}_2^a$ = coocurrence b/t present features ξ and destiny features υ
2. Do an **SVD** on $\hat{\Omega}_1^a$ and $\hat{\Omega}_2^a$. Use the top m singular vectors to project all samples to an m-dimensional space
3. Use the method of moments in this lower-dimensional space to estimate parameters π, t, o **up to linear transformation**

Overview

Experiments

- Phoneme recognition (TIMIT dataset)

$$\begin{array}{cccccccccccc} & & & & & & \downarrow & & & & & \\ a^{(i)} = & \dots & \text{sil} & \text{ao} & \text{ao} & \text{ao} & \text{ao} & \text{ao} & \text{ao} & \text{ao} & \text{ow} & \dots \\ x^{(i)} = & \dots & 353 & 26 & 11 & 12 & 15 & 14 & 17 & 447 & 435 & \dots \end{array}$$

- Feature examples

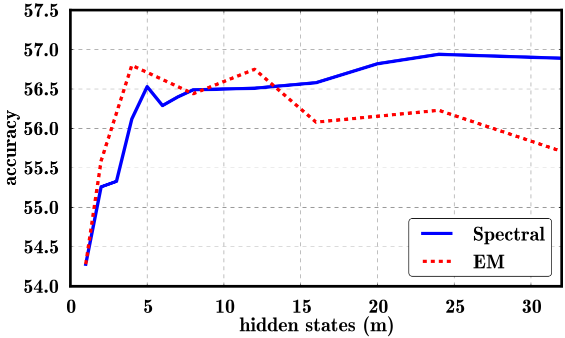
ϕ : next phoneme (ow)

ψ : previous phoneme (sil)

ξ : current observation (15)

υ : relative position within the current phoneme (“middle”)

Result on the Dev Set



Method	Accuracy
EM (4 HIDDEN STATES)	56.80
EM (24 HIDDEN STATES)	56.23
SPECTRAL (24 HIDDEN STATES)	**56.94**

Result on the Test Set

Method	Accuracy
Unigram (no hidden state)	48.04
HMM (no hidden state)	54.08
EM (4 hidden states)	55.49
Spectral(24 hidden states)	**55.82**
HTK	55.70

Conclusion

- Presented a spectral algorithm for estimation of R-HMMs
  - Statistical consistensy
  - No issue with local optima
- Empirically competitive with EM
- Spectral methods are a viable alternative to EM