Where We Are

- Language models
- Text classification
- **Linguistic analysis**
- Generation
Linguistic Analysis: Overview

Every linguistic analyzer is comprised of:

1. Theoretical motivation from linguistics and/or the text domain
2. An algorithm that maps $\mathcal{V}^\dagger$ to some output space $\mathcal{Y}$.
   - In this class, I’ll start with abstract algorithms applicable to many problems.
3. An implementation of the algorithm
   - Once upon a time: rule systems and crafted rules
   - Most common now: supervised learning from annotated data
   - Frontier: less supervision (semi-, un-, distant, . . .)
Sequence Labeling

After text classification ($\mathcal{V}^\dagger \rightarrow \mathcal{L}$), the next simplest type of output is a **sequence labeling**.

$$\langle x_1, x_2, \ldots, x_\ell \rangle \mapsto \langle y_1, y_2, \ldots, y_\ell \rangle$$

Every word (or character) gets a label in $\mathcal{L}$.

Example problems:

- part-of-speech tagging (Church, 1988)
- spelling correction (Kernighan et al., 1990)
- word alignment (Vogel et al., 1996)
- named-entity recognition (Bikel et al., 1999)
- compression (Conroy and O’Leary, 2001)
The Simplest Sequence Labeler

Define features of a labeled word in context: \( \phi(\mathbf{x}, i, y) \).

Train a classifier, e.g.,

\[
\hat{y}_i = \arg\max_{y \in \mathcal{L}} s(\mathbf{x}, i, y) \\
= \arg\max_{y \in \mathcal{L}} \text{linear } \mathbf{w} \cdot \phi(\mathbf{x}, i, y)
\]
The Simplest Sequence Labeler

Define features of a labeled word in context: $\phi(x, i, y)$.

Train a classifier, e.g.,

$$\hat{y}_i = \arg\max_{y \in \mathcal{L}} s(x, i, y)$$

$$= \arg\max_{y \in \mathcal{L}} \mathbf{w} \cdot \phi(x, i, y)$$

Sometimes this works!
The Simplest Sequence Labeler

Define features of a labeled word in context: $\phi(x, i, y)$.

Train a classifier, e.g.,

$$
\hat{y}_i = \arg\max_{y \in \mathcal{L}} s(x, i, y) \\
\text{linear} = \arg\max_{y \in \mathcal{L}} w \cdot \phi(x, i, y)
$$

Sometimes this works!

We can do better when there are predictable relationships between $Y_i$ and $Y_{i+1}$. 
Generative Sequence Labeling: Hidden Markov Models

\[ p(x, y) = \pi_{y_0} \prod_{i=1}^{\ell+1} \theta_{x_i|y_i} \cdot \gamma_{y_i|y_{i-1}} \]

For each state/label \( y \in \mathcal{L} \):

- \( \theta_{*|y} \) is the “emission” distribution
- \( \gamma_{*|y} \) is called the “transition” distribution

We saw this model before (Brown clustering on 1/25). Differences:

- We used “z” before, now it’s “y”
- Before, we wanted to discover each \( y_i \) (“unsupervised”)
- Now, we want to map \( x \mapsto y \), defined within a task (might be supervised or not)
Graphical Representation of Hidden Markov Models

Note: handling of beginning and end of sequence is a bit different than before. From here on, ignore last $x$ since $\theta \circ|\circ = 1$. 
Factor Graph Representation of Hidden Markov Models
A More General Form

Twice now, we’ve made the move from generative models based on repeated “rolls of dice” to discriminative models based on feature representations.

- Language modeling
- Text classification

In the structured case, we can do the same thing.

\[
\arg\max_{y \in \mathcal{L}^{\ell+1}} p(y_0) \prod_{i=1}^{\ell+1} p(x_i, y_i \mid y_{i-1})
\]

\[
= \arg\max_{y \in \mathcal{L}^{\ell+1}} \log p(y_0) + \sum_{i=1}^{\ell+1} \log p(x_i, y_i \mid y_{i-1})
\]

\[
= \arg\max_{y \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} w \cdot \phi(x_i, y_i, y_{i-1})
\]

In this case, each $Y_i$ “interacts” with $Y_{i-1}$ and $Y_{i+1}$ directly.
Structured vs. Not

Each of these has an advantage over the other:
- The HMM lets the different labels “interact.”
- The simple unstructured classifier makes all of $x$ available for every decision.
A More Powerful Solution

Slightly more generally, define features of adjacent labels in context: $\phi(x, i, y, y')$.

Features can depend on *any words at all*; this turns out not to affect asymptotic cost of prediction!
Local Pairwise Classifier

\[(\hat{y}_i, \hat{y}_{i+1}) = \arg\max_{y, y' \in L} w \cdot \phi(x, i, y, y')\]
Local Pairwise Classifier

\[(\hat{y}_i, \hat{y}_{i+1}) = \arg\max_{y,y' \in L} w \cdot \phi(x, i, y, y')\]
Local Pairwise Classifier

\[(\hat{y}_i, \hat{y}_{i+1}) = \text{argmax}_{y,y'\in L} w \cdot \phi(x, i, y, y')\]
Local Pairwise Classifier

\[(\hat{y}_i, \hat{y}_{i+1}) = \arg\max_{y, y' \in \mathcal{L}} w \cdot \phi(x, i, y, y')\]

The problem is with disagreements: what if the $Y_{1:2}$ prediction and the $Y_{2:3}$ prediction do not agree about $Y_2$?
Even More Powerful: “Global” Prediction

As with the pairwise model, define features of adjacent labeled words in context: $\phi(x, i, y, y')$

“Structured” classifier/predictor:

$$\hat{y} = \arg\max_{y \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} w \cdot \phi(x, i, y_i, y_{i-1})$$
Even More Powerful: “Global” Prediction

As with the pairwise model, define features of adjacent labeled words in context: $\phi(x, i, y, y')$

“Structured” classifier/predictor:

$$\hat{y} = \arg\max_{y \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} w \cdot \phi(x, i, y_i, y_{i-1})$$
Even More Powerful: “Global” Prediction

As with the pairwise model, define features of adjacent labeled words in context: \( \phi(x, i, y, y') \)

“Structured” classifier/predictor:

\[
\hat{y} = \arg\max_{y \in \mathcal{L}^{\ell+1}} \sum_{i=1}^{\ell+1} w \cdot \phi(x, i, y_i, y_{i-1})
\]

This is a fundamentally different kind of problem, demanding new:

▶ predicting ("decoding") algorithms
▶ training algorithms (to be discussed later)
We’ll start with the classical HMM, then return later to the featurized case.

\[
\arg\max_{\mathbf{y} \in \mathcal{L}^{\ell+1}} p(y_0) \prod_{i=1}^{\ell+1} p(x_i, y_i \mid y_{i-1})
\]

How to optimize over $|\mathcal{L}|^\ell$ choices without explicit enumeration?
Prediction with HMMs

We’ll start with the classical HMM, then return later to the featurized case.

\[
\arg\max_{y \in \mathcal{L}^{\ell+1}} p(y_0) \prod_{i=1}^{\ell+1} p(x_i, y_i | y_{i-1})
\]

How to optimize over \(|\mathcal{L}|^\ell\) choices without explicit enumeration?

Key: exploit the conditional independence assumptions:

\[
Y_i \perp Y_{1:i-2} \mid Y_{i-1} \\
Y_i \perp Y_{i+2:\ell} \mid Y_{i+1}
\]
Part-of-Speech Tagging Example

|   | l | suspect | the | present | forecast | is | pessimistic | . |
|---|---|---------|-----|---------|----------|----|-------------|
| noun | ● | ● | ● | ● | ● | ● | ● |
| adj. |   | ● | ● | ● | ● |   | ● |
| adv. |   |   |   |   |   |   |   |
| verb |   | ● | ● | ● | ● | ● |   |
| num. | ● |   |   |   |   |   |   |
| det. |   |   |   |   |   | ● |   |
| punc. |   |   |   |   |   |   | ● |

With this very simple tag set, $7^8 = 5.7$ million labelings.
(Even restricting to the possibilities above, 288 labelings.)
Two Obvious Solutions

**Brute force:** Enumerate all solutions, score them, pick the best.

**Greedy:** Pick each $\hat{y}_i$ according to:

$$
\hat{y}_i = \arg\max_{y \in \mathcal{L}} p(y \mid \hat{y}_{i-1}) \cdot p(x_i \mid y)
$$

What’s wrong with these?
We can get an exact solution in polynomial time!

\[ Y_i \perp Y_{1:i-2} \mid Y_{i-1} \]
\[ Y_i \perp Y_{i+2:\ell} \mid Y_{i+1} \]

Given the adjacent labels to \( Y_i \), others do not matter.

Let’s start at the last position, \( \ell \ldots \)
The End of the Sequence

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y^{last}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$p(Y_\ell = y \mid x, y_1:(\ell-1)) = p(Y_\ell = y \mid X_\ell = x_\ell, Y_{\ell-1} = y_{\ell-1}, Y_{\ell+1} = y')$$

$$= \gamma_{y \mid y} \cdot \theta_{x_\ell \mid y} \cdot \gamma_{y \mid y_{\ell-1}}$$

The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$. 
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.
First, think about the score of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

$$s_\ell(y) = \gamma_{|y|} \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y')$$
Recurrence

First, think about the score of the best sequence.

Let \( s_i(y) \) be the score of the best label sequence for \( x_{1:i} \) that ends in \( y \). It is defined recursively:

\[
\begin{align*}
    s_{\ell}(y) &= \gamma_{\ell} y \cdot \theta_{x_{\ell}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \\
    s_{\ell-1}(y) &= \theta_{x_{\ell-1}|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-2}(y')
\end{align*}
\]
Recurrence

First, think about the *score* of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

$$s_{\ell}(y) = \gamma_{y} \cdot \theta_{x_{\ell}|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-1}(y')$$

$$s_{\ell-1}(y) = \theta_{x_{\ell}|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-2}(y')$$

$$s_{\ell-2}(y) = \theta_{x_{\ell}|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-3}(y')$$
Recurrence

First, think about the score of the best sequence.

Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

\[
\begin{align*}
    s_\ell(y) &= \gamma_{x_\ell | y} \cdot \theta_{x_\ell | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{\ell-1}(y') \\
    s_{\ell-1}(y) &= \theta_{x_\ell | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{\ell-2}(y') \\
    s_{\ell-2}(y) &= \theta_{x_\ell | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{\ell-3}(y') \\
    &\quad \vdots \\
    s_i(y) &= \theta_{x_i | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot s_{i-1}(y')
\end{align*}
\]
Recurrence

First, think about the score of the best sequence.
Let $s_i(y)$ be the score of the best label sequence for $x_{1:i}$ that ends in $y$. It is defined recursively:

$$s_\ell(y) = \gamma_{y} \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y')$$

$$s_{\ell-1}(y) = \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-2}(y')$$

$$s_{\ell-2}(y) = \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-3}(y')$$

$$\vdots$$

$$s_i(y) = \theta_{x_i|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')$$

$$\vdots$$

$$s_1(y) = \theta_{x_1|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot \pi_{y'}$$
Viterbi Procedure (Part I: Prefix Scores)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y^{last}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
Viterbi Procedure (Part I: Prefix Scores)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y^{last}$</td>
<td>$s_1(y^{last})$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$s_1(y) = \theta_{x_1 | y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y | y'} \cdot \pi_{y'}$
Viterbi Procedure (Part I: Prefix Scores)

\[
\begin{array}{|c|ccccc|}
\hline
 & x_1 & x_2 & \ldots & x_\ell \\
\hline
y & s_1(y) & s_2(y) & & & \\
\hline
y' & s_1(y') & s_2(y') & & & \\
\hline
\vdots & & & & & \\
\hline
y^{last} & s_1(y^{last}) & s_2(y^{last}) & & & \\
\hline
\end{array}
\]

\[
s_i(y) = \theta_{x_i \mid y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y \mid y'} \cdot s_{i-1}(y')
\]
Viterbi Procedure (Part I: Prefix Scores)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\ldots$</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$s_2(y)$</td>
<td>$\ldots$</td>
<td>$s_\ell(y)$</td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$s_2(y')$</td>
<td>$\ldots$</td>
<td>$s_\ell(y')$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$y^{\text{last}}$</td>
<td>$s_1(y^{\text{last}})$</td>
<td>$s_2(y^{\text{last}})$</td>
<td>$\ldots$</td>
<td>$s_\ell(y^{\text{last}})$</td>
</tr>
</tbody>
</table>

\[
s_\ell(y) = \gamma_{\bigcirc |y} \cdot \theta_{x_\ell |y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y |y'} \cdot s_{\ell-1}(y')
\]
Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y)$
Claim: $\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y)$

\[
\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} \gamma \circ |y \cdot \theta x_\ell | y \cdot \max_{y' \in \mathcal{L}} \gamma y | y' \cdot s_{\ell-1}(y')
\]
Claim: \( \max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y) \)

\[
\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} \gamma_\circ \mid_{y} \cdot \theta_{x_\ell \mid y} \cdot \max_{y' \in \mathcal{L}} \gamma_y \mid_{y'} \cdot s_{\ell-1}(y')
\]

\[
= \max_{y \in \mathcal{L}} \gamma_\circ \mid_{y} \cdot \theta_{x_\ell \mid y} \cdot \max_{y' \in \mathcal{L}} \gamma_y \mid_{y'} \cdot \theta_{x_{\ell-1} \mid y'} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'} \mid_{y''} \cdot s_{\ell-2}(y'')
\]
Claim: \( \max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}^{\ell+1}} p(x, y) \)

\[
\max_{y \in \mathcal{L}} s_\ell(y) = \max_{y \in \mathcal{L}} \gamma \cap |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \\
= \max_{y \in \mathcal{L}} \gamma \cap |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot \theta_{x_{\ell-1}|y'} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'|y''} \cdot s_{\ell-2}(y'') \\
= \max_{y \in \mathcal{L}} \gamma \cap |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot \theta_{x_{\ell-1}|y'} \cdot \max_{y'' \in \mathcal{L}} \gamma_{y'|y''} \cdot \theta_{x_{\ell-2}|y''} \cdot \max_{y''' \in \mathcal{L}} \gamma_{y''|y'''} \cdot s_{\ell-3}(y''')
\]
Claim: \( \max_{y \in L} s_\ell(y) = \max_{y \in L^{\ell+1}} p(x, y) \)

\[
\max_{y \in L} s_\ell(y) = \max_{y \in L} \gamma(y) \cdot \theta x_\ell | y \cdot \max_{y' \in L} \gamma y | y' \cdot s_{\ell-1}(y')
\]

\[
= \max_{y \in L} \gamma(y) \cdot \theta x_\ell | y \cdot \max_{y' \in L} \gamma y | y' \cdot \theta x_{\ell-1} | y' \cdot \max_{y'' \in L} \gamma y' | y'' \cdot s_{\ell-2}(y'')
\]

\[
= \max_{y \in L} \gamma(y) \cdot \theta x_\ell | y \cdot \max_{y' \in L} \gamma y | y' \cdot \theta x_{\ell-1} | y' \cdot \max_{y'' \in L} \gamma y' | y'' \cdot \theta x_{\ell-2} | y'' \cdot \max_{y''' \in L} \gamma y'' | y''' \cdot s_{\ell-3}(y''')
\]

\[
= \max_{y \in L^{\ell+1}} \gamma(y) \cdot \theta x_\ell | y_\ell \cdot \gamma y_\ell | y_{\ell-1} \cdot \theta x_{\ell-1} | y_{\ell-1} \cdot \gamma y_{\ell-1} | y_{\ell-2} \cdot \theta x_{\ell-2} | y_{\ell-2} \cdots \theta x_1 | y_1 \cdot \gamma y_1 | y_0 \cdot \pi y_0
\]
Claim: \( \max_{y \in L} s_\ell(y) = \max_{y \in L^{\ell+1}} p(x, y) \)

\[
\begin{align*}
\max_{y \in L} s_\ell(y) &= \max_{y \in L} \gamma \circ |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot s_{\ell-1}(y') \\
&= \max_{y \in L} \gamma \circ |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot \theta_{x_{\ell-1}|y'} \cdot \max_{y'' \in L} \gamma_{y''|y'} \cdot s_{\ell-2}(y'') \\
&= \max_{y \in L} \gamma \circ |y| \cdot \theta_{x_\ell|y} \cdot \max_{y' \in L} \gamma_{y|y'} \cdot \theta_{x_{\ell-1}|y'} \cdot \max_{y'' \in L} \gamma_{y''|y'} \cdot \theta_{x_{\ell-2}|y''} \cdot \max_{y''' \in L} \gamma_{y'''|y''} \cdot s_{\ell-3}(y''') \\
&= \max_{y \in L^{\ell+1}} \gamma \circ |y_\ell| \cdot \theta_{x_\ell|y_\ell} \cdot \gamma_{y_\ell|y_{\ell-1}} \cdot \theta_{x_{\ell-1}|y_{\ell-1}} \cdot \gamma_{y_{\ell-1}|y_{\ell-2}} \cdot \theta_{x_{\ell-2}|y_{\ell-2}} \cdots \theta_{x_1|y_1} \cdot \gamma_{y_1|y_0} \cdot \pi y_0 \\
&= \max_{y \in L^{\ell+1}} \pi y_0 \prod_{i=1}^{\ell+1} \theta_{x_i|y_i} \cdot \gamma_{y_i|y_{i-1}}
\end{align*}
\]
High-Level View of Viterbi

- The decision about $Y_\ell$ is a function of $y_{\ell-1}$, $x$, and nothing else!
- If, for each value of $y_{\ell-1}$, we knew the best $y_{1:(\ell-1)}$, then picking $y_\ell$ (and $y_{\ell-1}$) would be easy.
- Idea: for each position $i$, calculate the score of the best label prefix $y_{1:i}$ ending in each possible value for $Y_i$.
- With a little bookkeeping, we can then trace backwards and recover the best label sequence.
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\ldots$</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_{\text{last}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>\ldots</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$b_1(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$b_1(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$y^{last}$</td>
<td>$s_1(y^{last})$</td>
<td>$b_1(y^{last})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$s_1(y) = \theta_{x_1\mid y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y\mid y'} \cdot \pi_{y'}$

$b_1(y) = \arg\max_{y' \in \mathcal{L}} \gamma_{y\mid y'} \cdot \pi_{y'}$
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$s_2(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y)$</td>
<td>$b_2(y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$s_2(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y')$</td>
<td>$b_2(y')$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y^{last}$</td>
<td>$s_1(y^{last})$</td>
<td>$s_2(y^{last})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$b_1(y^{last})$</td>
<td>$b_2(y^{last})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
s_i(y) = \theta_{x_i|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot \boxed{s_{i-1}(y')}
\]

\[
b_i(y) = \arg\max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{i-1}(y')
\]
Viterbi Procedure (Part I: Prefix Scores and Backpointers)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$s_1(y)$</td>
<td>$s_2(y)$</td>
<td>...</td>
<td>$s_\ell(y)$</td>
</tr>
<tr>
<td></td>
<td>$b_1(y)$</td>
<td>$b_2(y)$</td>
<td></td>
<td>$b_\ell(y)$</td>
</tr>
<tr>
<td>$y'$</td>
<td>$s_1(y')$</td>
<td>$s_2(y')$</td>
<td>...</td>
<td>$s_\ell(y')$</td>
</tr>
<tr>
<td></td>
<td>$b_1(y')$</td>
<td>$b_2(y')$</td>
<td></td>
<td>$b_\ell(y')$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_{\text{last}}$</td>
<td>$s_1(y_{\text{last}})$</td>
<td>$s_2(y_{\text{last}})$</td>
<td>...</td>
<td>$s_\ell(y_{\text{last}})$</td>
</tr>
<tr>
<td></td>
<td>$b_1(y_{\text{last}})$</td>
<td>$b_2(y_{\text{last}})$</td>
<td></td>
<td>$b_\ell(y_{\text{last}})$</td>
</tr>
</tbody>
</table>

\[ s_\ell(y) = \gamma \circ |y \cdot \theta_{x_\ell|y} \cdot \max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \]

\[ b_\ell(y) = \arg\max_{y' \in \mathcal{L}} \gamma_{y|y'} \cdot s_{\ell-1}(y') \]
Full Viterbi Procedure

Input: $x, \theta, \gamma, \pi$

Output: $\hat{y}$

1. For $i \in \langle 1, \ldots, \ell \rangle$:
   - Solve for $s_i(\ast)$ and $b_i(\ast)$.
     - Special base case for $i = 1$ to handle $\pi$
     - General recurrence for $i \in \langle 2, \ldots, \ell - 1 \rangle$
     - Special case for $i = \ell$ to handle stopping probability

2. $\hat{y}_\ell \leftarrow \operatorname{argmax}_{y \in \mathcal{L}} s_\ell(y)$

3. For $i \in \langle \ell, \ldots, 1 \rangle$:
   - $\hat{y}_{i-1} \leftarrow b(y_i)$
Readings and Reminders

- Collins (2011), which has somewhat different notation; Jurafsky and Martin (2015)
- Submit a suggestion for an exam question by Friday at 5pm.


