Lecture 5: Linear Sequential Models

Vlad Niculae & André Martins

Deep Structured Learning Course, Fall 2019

Today's Roadmap

Today we're starting to talk about **structure**, more specifically sequences:

- Generative sequence models: (hidden) Markov models
- Dynamic programming: the Viterbi and Forward-Backward algorithms
- Viterbi decoding and minimum risk decoding
- Unsupervised learning with the Baum-Welch (EM) algorithm

Outline

So far, we looked at **classification**: simply predicting one-of-K classes.

How about more complicated output spaces?

Named Entity Recognition

McGrath left out of Ireland World Cup squad

Named Entity Recognition

I	O	O	O	I	I	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently:
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$ I,
  $\boldsymbol{x}_2 =$ left; $\boldsymbol{y}_2 =$ O, ...

Named Entity Recognition

I	O	O	O	I	I	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently:
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$ I,
  $\boldsymbol{x}_2 =$ left; $\boldsymbol{y}_2 =$ O, ...

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently:
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
  $\boldsymbol{x}_2 =$ left; $\boldsymbol{y}_2 =$O, ...
- This *labeling scheme* is called BIO (beginning/inside/outside).

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently:
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
  $\boldsymbol{x}_2 =$ left; $\boldsymbol{y}_2 =$O, ...
- This *labeling scheme* is called BIO (beginning/inside/outside).
- Could add more context into $\boldsymbol{x}$, e.g. a window:
  $\boldsymbol{x}_1 = (\$, \textbf{McGrath}, \text{left})$; $\boldsymbol{y}_1 =$B,
  $\boldsymbol{x}_2 = (\text{McGrath}, \textbf{left}, \text{out})$; $\boldsymbol{y}_2 =$O, ...

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently:
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
  $\boldsymbol{x}_2 =$ left; $\boldsymbol{y}_2 =$O, ...
- This *labeling scheme* is called BIO (beginning/inside/outside).
- Could add more context into $\boldsymbol{x}$, e.g. a window:
  $\boldsymbol{x}_1 = (\$, \textbf{McGrath}, \text{left})$; $\boldsymbol{y}_1 =$B,
  $\boldsymbol{x}_2 = (\text{McGrath}, \textbf{left}, \text{out})$; $\boldsymbol{y}_2 =$O, ...
- But the predictions can't depend on one another!
  OI is not allowed; BB is allowed but relatively rare!
  ... what do we *really* want?

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently: this makes a big assumption!
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
- Really, we should think at sentence-level:
  $\boldsymbol{x}_1 =$ McGrath left out of...; $\boldsymbol{y}_1 =$B O O O B B I O
- Can capture more statistics (BB is rare, BBBB even rarer, OO is common...)
- Can we attempt this with our multi-class toolbox (e.g. naïve Bayes?)

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently: this makes a big assumption!
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
- Really, we should think at sentence-level:
  $\boldsymbol{x}_1 =$ McGrath left out of...; $\boldsymbol{y}_1 =$B O O O B B I O
- Can capture more statistics (BB is rare, BBBB even rarer, OO is common...)
- Can we attempt this with our multi-class toolbox (e.g. naïve Bayes?)
  How many sentences have the label B O O O B B I O?

Named Entity Recognition

B	O	O	O	B	B	I	O
McGrath	left	out	of	Ireland	World	Cup	squad

- Classify each word independently: this makes a big assumption!
  $\boldsymbol{x}_1 =$ McGrath; $\boldsymbol{y}_1 =$B,
- Really, we should think at sentence-level:
  $\boldsymbol{x}_1 =$ McGrath left out of...; $\boldsymbol{y}_1 =$B O O O B B I O
- Can capture more statistics (BB is rare, BBBB even rarer, OO is common...)
- Can we attempt this with our multi-class toolbox (e.g. naïve Bayes?)
  How many sentences have the label B O O O B B I O?

rank	sequence	count
1	O	1072
2	B O	663
3	B O O O O O O O	446
4	B O B O	378
5	O B I O B O O	272
	...	
4856	B O O O B B I O	1

We are essentially treating each label sequence $\boldsymbol{y}$ as a distinct object, *ignoring its internal structure!*

We are essentially treating each label sequence $\boldsymbol{y}$ as a distinct object, *ignoring its internal structure!*

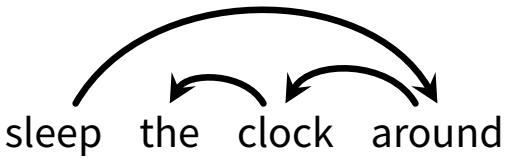
We've never seen sentences labelled `O O O O O O B`,
but we *have* seen sentences labelled `O O O O O B`.
Surely that could help!

Structured Prediction

A framework for handling structured, constrained, inter-dependent outputs.

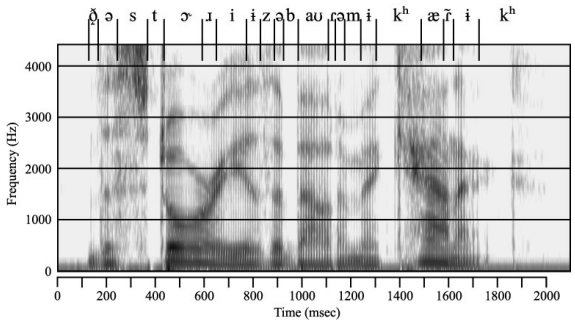
NLP

- Named Entity Recognition
- Machine Translation
- Syntactic Parsing



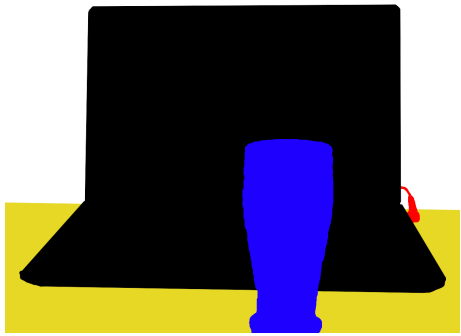
Speech Processing

- Speaker ID
- Speech Recognition



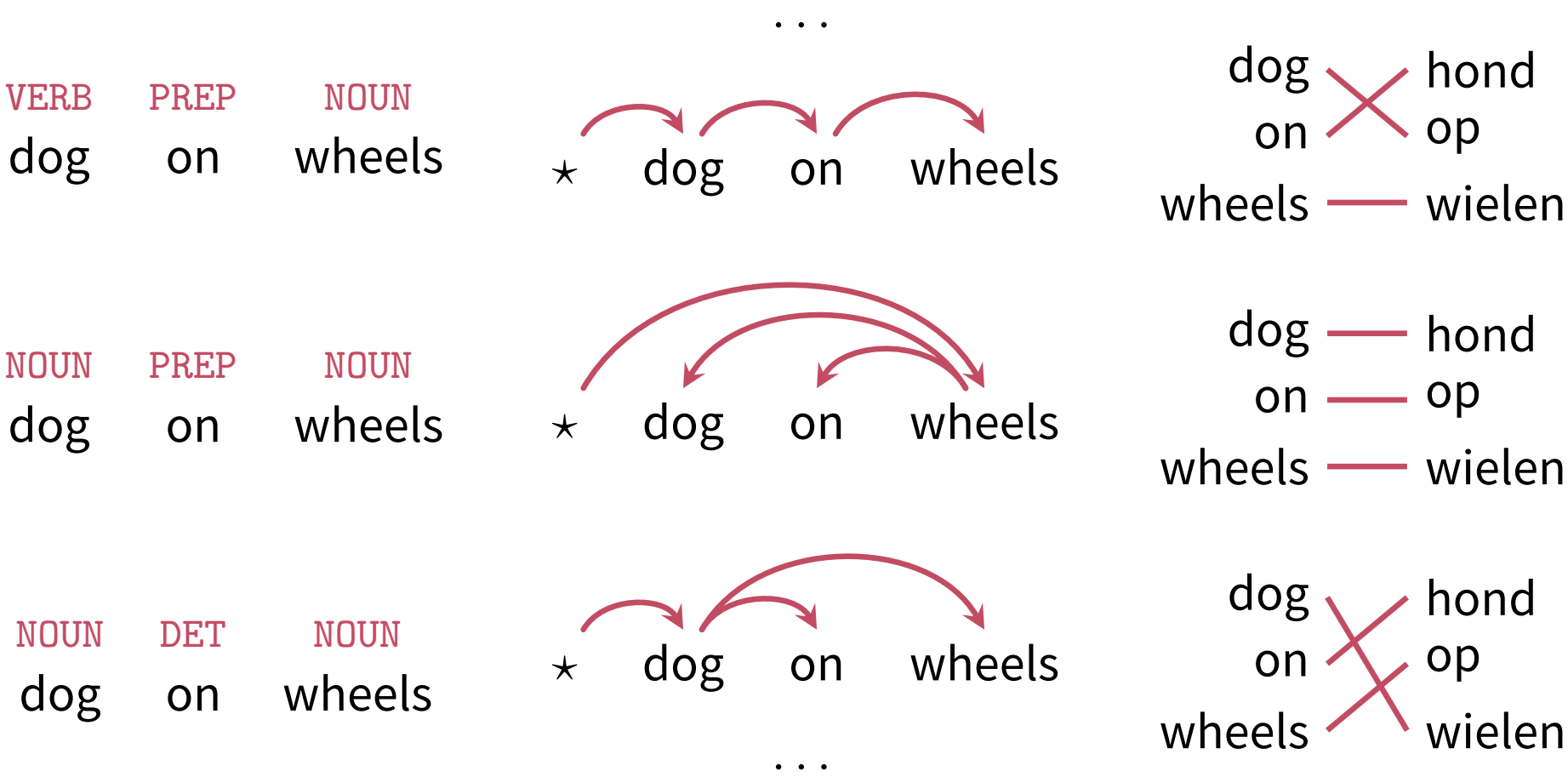
Computer Vision

- Object detection
- Segmentation

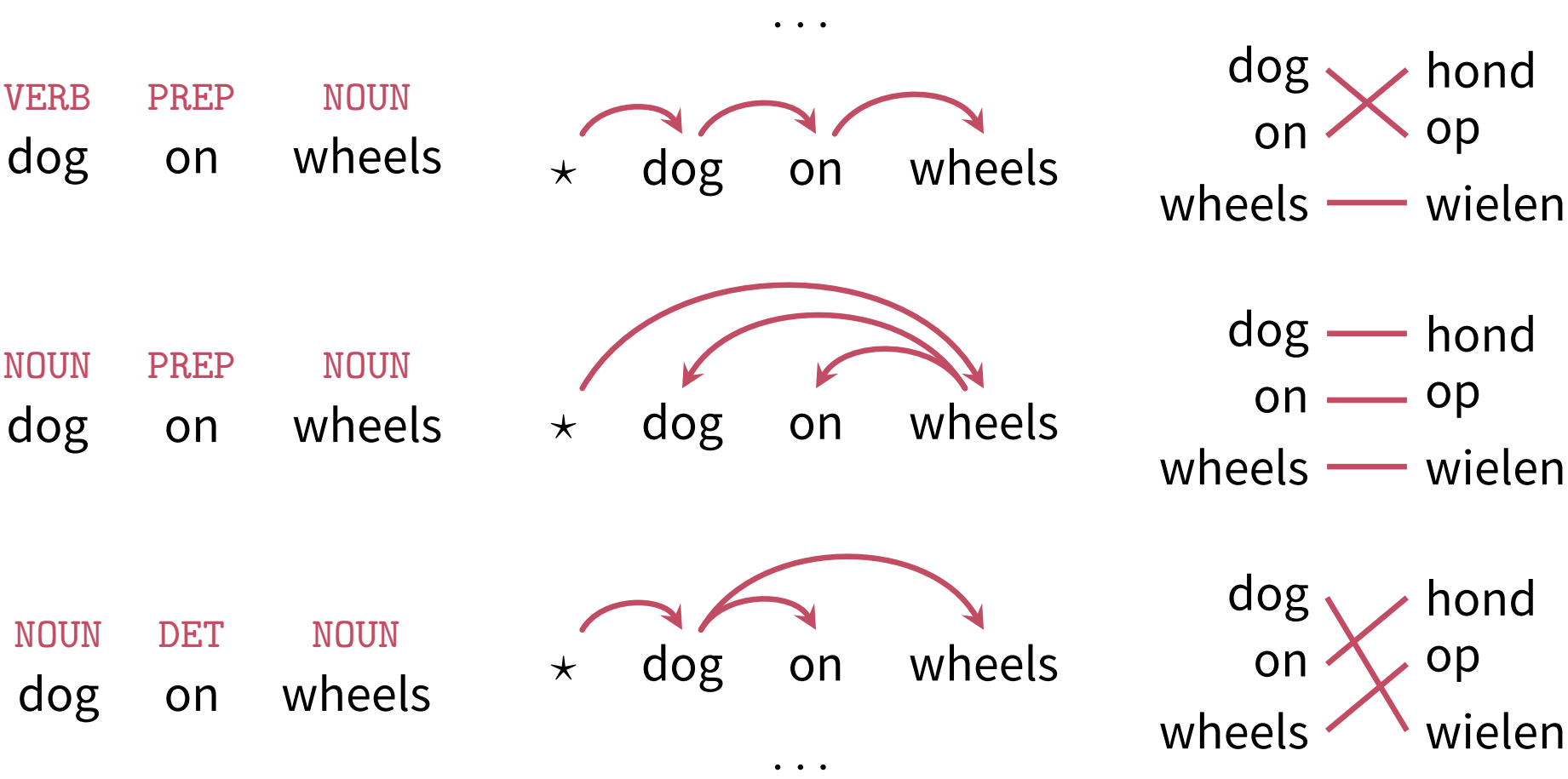


computational biology, robotics / planning, time series forecasting, etc.

Structured Prediction



Structured Prediction



Today, we talk about sequences.

Roadmap: Models for Structured Prediction

Binary/Multi-class	Structured Prediction
Naive Bayes	?
Logistic Regression	?
Perceptron	?
SVMs	?

Outline

Outline

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P\big(\boldsymbol{y} = [\texttt{O} \;\; \texttt{O} \;\; \texttt{O} \;\; \texttt{B}]\big) = ?$

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P\big(\boldsymbol{y} = [\texttt{O} \;\; \texttt{O} \;\; \texttt{O} \;\; \texttt{B}]\big) = ?$

What are all possible labellings?

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P\big(\boldsymbol{y} = [\texttt{O}\ \ \texttt{O}\ \ \texttt{O}\ \ \texttt{B}]\big) = ?$

What are all possible labellings?

Let $\Sigma = \{\texttt{B}, \texttt{I}, \texttt{O}\}$. Then, $\mathcal{Y} = \Sigma^\star = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P(\boldsymbol{y} = [\texttt{O}\ \ \texttt{O}\ \ \texttt{O}\ \ \texttt{B}]) =$?

What are all possible labellings?

Let $\Sigma = \{\texttt{B}, \texttt{I}, \texttt{O}\}$. Then, $\mathcal{Y} = \Sigma^\star = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

Could set $\mathrm{P}(\boldsymbol{y}) \propto \#\boldsymbol{y}$

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P(\boldsymbol{y} = [\texttt{O}\ \texttt{O}\ \texttt{O}\ \texttt{B}]) =$?

What are all possible labellings?

Let $\Sigma = \{\texttt{B}, \texttt{I}, \texttt{O}\}$. Then, $\mathcal{Y} = \Sigma^\star = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

Could set $\mathrm{P}(\boldsymbol{y}) \propto \#\boldsymbol{y}$

rank	sequence	count
1	O	1072
2	B O	663
3	B O O O O O O O	446
4	B O B O	378
5	O B I O B O O	272
	...	
4856	B O O O B B I O	1

Probabilities over sequences

To begin, let's forget for a moment about $\boldsymbol{x}$ and focus on $\boldsymbol{y}$.

$P(\boldsymbol{y} = [\texttt{O}\ \texttt{O}\ \texttt{O}\ \texttt{B}]) = ?$

What are all possible labellings?

Let $\Sigma = \{\texttt{B}, \texttt{I}, \texttt{O}\}$. Then, $\mathcal{Y} = \Sigma^\star = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$

Could set $\mathrm{P}(\boldsymbol{y}) \propto \#\boldsymbol{y}$

rank	sequence	count
1	O	1072
2	B O	663
3	B O O O O O O O	446
4	B O B O	378
5	O B I O B O O	272
	...	
4856	B O O O B B I O	1

Issues:

- $\mathrm{P}(\boldsymbol{y}) = 0$ for **most** $\boldsymbol{y}$!
- No sharing between sequences that are similar!
- Ignores the fact that $\boldsymbol{y} = [y_1, \ldots, y_L]$.

Lower extreme: Bag-of-words model

$$\text{Let } \boldsymbol{y} = [y_1, \ldots, y_L]. \quad \mathrm{P}(\boldsymbol{y}) = \prod_{i=1}^{L} \mathrm{P}(y_i)$$

- Also called "unigram" model
- Assumes every word is generated independently of other words
  therefore, abandons the structure of $\boldsymbol{y}$ entirely.

Lower extreme: Bag-of-words model

$$\text{Let } \boldsymbol{y} = [y_1, \ldots, y_L]. \quad \mathrm{P}(\boldsymbol{y}) = \prod_{i=1}^{L} \mathrm{P}(y_i)$$

- Also called "unigram" model
- Assumes every word is generated independently of other words
  therefore, abandons the structure of $\boldsymbol{y}$ entirely.
- Probability of a string is insensitive to word order:

$$\mathrm{P}\left([\mathtt{O}, \mathtt{B}, \mathtt{O}]\right) = \mathrm{P}\left([\mathtt{B}, \mathtt{O}, \mathtt{O}]\right) = \mathrm{P}(\mathtt{B})\,\mathrm{P}(\mathtt{O})^2$$

Lower extreme: Bag-of-words model

$$\text{Let } \boldsymbol{y} = [y_1, \ldots, y_L]. \quad \mathrm{P}(\boldsymbol{y}) = \prod_{i=1}^{L} \mathrm{P}(y_i)$$

- Also called "unigram" model
- Assumes every word is generated independently of other words
  therefore, abandons the structure of $\boldsymbol{y}$ entirely.
- Probability of a string is insensitive to word order:

$$\mathrm{P}\left([\mathtt{O}, \mathtt{B}, \mathtt{O}]\right) = \mathrm{P}\left([\mathtt{B}, \mathtt{O}, \mathtt{O}]\right) = \mathrm{P}(\mathtt{B})\,\mathrm{P}(\mathtt{O})^2$$

- How many parameters do we need to estimate and how?

Upper extreme: Full history model

Let $\boldsymbol{y} = [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]$.

$\mathsf{P}(\boldsymbol{y}) =$

Upper extreme: Full history model

$$
\begin{aligned}
\text{Let } \boldsymbol{y} = & [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]. \\
\mathrm{P}(\boldsymbol{y}) = & \mathrm{P}(y_1 | \texttt{start}) \\
& \cdot \mathrm{P}(y_2 | \texttt{start}, y_1) \\
& \cdot \ldots \\
& \cdot \mathrm{P}(y_L | \texttt{start}, y_1, y_2, \ldots, y_{L-1}) \\
& \cdot \mathrm{P}(\texttt{end} | \texttt{start}, y_1, y_2, \ldots, y_{L-1}, y_L)
\end{aligned}
$$

Upper extreme: Full history model

$$
\begin{aligned}
\text{Let } \boldsymbol{y} = & [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]. \\
\mathrm{P}(\boldsymbol{y}) = & \mathrm{P}(y_1 | \texttt{start}) \\
& \cdot \mathrm{P}(y_2 | \texttt{start}, y_1) \\
& \cdot \ldots \\
& \cdot \mathrm{P}(y_L | \texttt{start}, y_1, y_2, \ldots, y_{L-1}) \\
& \cdot \mathrm{P}(\texttt{end} | \texttt{start}, y_1, y_2, \ldots, y_{L-1}, y_L) \\
= & \prod_{i=1}^{L+1} \mathrm{P}(y_i | \underbrace{y_1, \ldots, y_{i-1}}_{\boldsymbol{y}'})
\end{aligned}
$$

Upper extreme: Full history model

$$
\begin{aligned}
\text{Let } \boldsymbol{y} =& [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]. \\
\mathrm{P}(\boldsymbol{y}) =& \mathrm{P}(y_1 | \texttt{start}) \\
& \cdot \mathrm{P}(y_2 | \texttt{start}, y_1) \\
& \cdot \ldots \\
& \cdot \mathrm{P}(y_L | \texttt{start}, y_1, y_2, \ldots, y_{L-1}) \\
& \cdot \mathrm{P}(\texttt{end} | \texttt{start}, y_1, y_2, \ldots, y_{L-1}, y_L) \\
=& \prod_{i=1}^{L+1} \mathrm{P}(y_i | \underbrace{y_1, \ldots, y_{i-1}}_{\boldsymbol{y}'})
\end{aligned}
$$

- Each symbol y_i generated based on entire history $\boldsymbol{y}'$.

Upper extreme: Full history model

$$
\begin{aligned}
\text{Let } \boldsymbol{y} =& [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]. \\
\mathrm{P}(\boldsymbol{y}) =& \mathrm{P}(y_1 | \texttt{start}) \\
& \cdot \mathrm{P}(y_2 | \texttt{start}, y_1) \\
& \cdot \ldots \\
& \cdot \mathrm{P}(y_L | \texttt{start}, y_1, y_2, \ldots, y_{L-1}) \\
& \cdot \mathrm{P}(\texttt{end} | \texttt{start}, y_1, y_2, \ldots, y_{L-1}, y_L) \\
=& \prod_{i=1}^{L+1} \mathrm{P}(y_i | \underbrace{y_1, \ldots, y_{i-1}}_{\boldsymbol{y}'})
\end{aligned}
$$

- Each symbol y_i generated based on entire history $\boldsymbol{y}'$.
- Must estimate $\mathrm{P}(y|\boldsymbol{y}')$ for every possible history $\boldsymbol{y}'$!
  ... and we're back where we started: same as counting sequences.

Upper extreme: Full history model

$$
\begin{aligned}
\text{Let } \boldsymbol{y} =& [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]. \\
\mathrm{P}(\boldsymbol{y}) =& \mathrm{P}(y_1 | \texttt{start}) \\
& \cdot \mathrm{P}(y_2 | \texttt{start}, y_1) \\
& \cdot \ldots \\
& \cdot \mathrm{P}(y_L | \texttt{start}, y_1, y_2, \ldots, y_{L-1}) \\
& \cdot \mathrm{P}(\texttt{end} | \texttt{start}, y_1, y_2, \ldots, y_{L-1}, y_L) \\
=& \prod_{i=1}^{L+1} \mathrm{P}(y_i | \underbrace{y_1, \ldots, y_{i-1}}_{\boldsymbol{y}'})
\end{aligned}
$$

- Each symbol y_i generated based on entire history $\boldsymbol{y}'$.
- Must estimate $\mathrm{P}(y|\boldsymbol{y}')$ for every possible history $\boldsymbol{y}'$!
  ... and we're back where we started: same as counting sequences.
- Idea: condition only on the last *few* symbols.

In-between: Markov Models

$$\text{Let } \boldsymbol{y} = [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}].$$

$$\begin{aligned} \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(y_1|\texttt{start}) \cdot \mathrm{P}(y_2|y_1) \cdot \ldots \cdot \mathrm{P}(y_L|y_{L-1}) \cdot \mathrm{P}(\texttt{end}|y_L) \\ &= \prod_{i=1}^{L+1} \mathrm{P}(y_i|y_{i-1}) \end{aligned}$$

- Each symbol only depends on the previous word.
- We estimate transition probabilities $\mathrm{P}(y_i|y_{i-1})$;
  Including initial and final probabilities $\mathrm{P}(y_1|\texttt{start})$ and $\mathrm{P}(\texttt{end}|y_L)$.
- Total number of parameters:

In-between: Markov Models

$$\text{Let } \boldsymbol{y} = [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}].$$

$$\begin{aligned} \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(y_1|\texttt{start}) \cdot \mathrm{P}(y_2|y_1) \cdot \ldots \cdot \mathrm{P}(y_L|y_{L-1}) \cdot \mathrm{P}(\texttt{end}|y_L) \\ &= \prod_{i=1}^{L+1} \mathrm{P}(y_i|y_{i-1}) \end{aligned}$$

- Each symbol only depends on the previous word.
- We estimate transition probabilities $\mathrm{P}(y_i|y_{i-1})$;
  Including initial and final probabilities $\mathrm{P}(y_1|\texttt{start})$ and $\mathrm{P}(\texttt{end}|y_L)$.
- Total number of parameters: $O(|\Sigma|^2)$.

$\mathrm{P}(\texttt{B}\|\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\|\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\|\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\|\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\|\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\|\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\|\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\|\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\|\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\|\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\|\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\|\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\|\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\|\texttt{O}) = .084$

Aside: k^{th} order Markov Models

Let $\boldsymbol{y} = [\texttt{start}, y_1, y_2, \ldots, y_L, \texttt{end}]$. $\quad \mathrm{P}(\boldsymbol{y}) = \prod_{i=1}^{L+1} \mathrm{P}(y_i | y_{i-1}, \ldots, y_{i-k})$

- Each symbol depends on k previous symbols.
- Transition probabilities $\mathrm{P}(y_i | y_{i-1}, \ldots, y_{i-k})$
- Total number of parameters: $O(|\Sigma|^{k+1})$
- **Widely used in language modeling**
  - Here, $\Sigma =$ the vocabulary of English words.
  - Goal: next word prediction; $\mathrm{P}(w | \text{“can we rely”}) = ?$

Estimating Markov Models: Maximum Likelihood

$$P(y_i = \texttt{b} | y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a}, \texttt{b}]}{\sum_{\texttt{b'} \in \Sigma} \#[\texttt{a}, \texttt{b'}]} = \frac{\#[\texttt{a}, \texttt{b}]}{\#\texttt{a}}$$

$P(\texttt{B}\|\texttt{start}) = .393$	$P(\texttt{B}\|\texttt{B}) = .009$	$P(\texttt{B}\|\texttt{I}) = .003$	$P(\texttt{B}\|\texttt{O}) = .102$
$P(\texttt{I}\|\texttt{start}) = .0$	$P(\texttt{I}\|\texttt{B}) = .369$	$P(\texttt{I}\|\texttt{I}) = .178$	$P(\texttt{I}\|\texttt{O}) = .0$
$P(\texttt{O}\|\texttt{start}) = .607$	$P(\texttt{O}\|\texttt{B}) = .610$	$P(\texttt{O}\|\texttt{I}) = .779$	$P(\texttt{O}\|\texttt{O}) = .815$
$P(\texttt{end}\|\texttt{start}) = .0$	$P(\texttt{end}\|\texttt{B}) = .013$	$P(\texttt{end}\|\texttt{I}) = .040$	$P(\texttt{end}\|\texttt{O}) = .084$

Estimating Markov Models: Maximum Likelihood

$$P(y_i = \texttt{b} | y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a}, \texttt{b}]}{\sum_{\texttt{b'} \in \Sigma} \#[\texttt{a}, \texttt{b'}]} = \frac{\#[\texttt{a}, \texttt{b}]}{\#\texttt{a}}$$

$P(\texttt{B}\|\texttt{start}) = .393$	$P(\texttt{B}\|\texttt{B}) = .009$	$P(\texttt{B}\|\texttt{I}) = .003$	$P(\texttt{B}\|\texttt{O}) = .102$
$P(\texttt{I}\|\texttt{start}) = .0$	$P(\texttt{I}\|\texttt{B}) = .369$	$P(\texttt{I}\|\texttt{I}) = .178$	$P(\texttt{I}\|\texttt{O}) = .0$
$P(\texttt{O}\|\texttt{start}) = .607$	$P(\texttt{O}\|\texttt{B}) = .610$	$P(\texttt{O}\|\texttt{I}) = .779$	$P(\texttt{O}\|\texttt{O}) = .815$
$P(\texttt{end}\|\texttt{start}) = .0$	$P(\texttt{end}\|\texttt{B}) = .013$	$P(\texttt{end}\|\texttt{I}) = .040$	$P(\texttt{end}\|\texttt{O}) = .084$

We never saw [O I], so $P(\texttt{I}|\texttt{O}) = 0$.

The model rules out **any sequence containing** [O I]!

$$P([\texttt{start}, \texttt{O}, \texttt{O}, \texttt{I}, \texttt{B}, \texttt{O}, \texttt{end}]) = P(\texttt{O}|\texttt{start}) \cdot P(\texttt{O}|\texttt{O}) \cdot \underbrace{P(\texttt{I}|\texttt{O})}_{=0} \cdot \ldots = 0$$

Estimating Markov Models: Maximum Likelihood

$$P(y_i = \texttt{b} | y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a}, \texttt{b}]}{\sum_{\texttt{b'} \in \Sigma} \#[\texttt{a}, \texttt{b'}]} = \frac{\#[\texttt{a}, \texttt{b}]}{\#\texttt{a}}$$

$P(\texttt{B}\|\texttt{start}) = .393$	$P(\texttt{B}\|\texttt{B}) = .009$	$P(\texttt{B}\|\texttt{I}) = .003$	$P(\texttt{B}\|\texttt{O}) = .102$
$P(\texttt{I}\|\texttt{start}) = .0$	$P(\texttt{I}\|\texttt{B}) = .369$	$P(\texttt{I}\|\texttt{I}) = .178$	$P(\texttt{I}\|\texttt{O}) = .0$
$P(\texttt{O}\|\texttt{start}) = .607$	$P(\texttt{O}\|\texttt{B}) = .610$	$P(\texttt{O}\|\texttt{I}) = .779$	$P(\texttt{O}\|\texttt{O}) = .815$
$P(\texttt{end}\|\texttt{start}) = .0$	$P(\texttt{end}\|\texttt{B}) = .013$	$P(\texttt{end}\|\texttt{I}) = .040$	$P(\texttt{end}\|\texttt{O}) = .084$

We never saw [O I], so $P(\texttt{I}|\texttt{O}) = 0$.
The model rules out **any sequence containing** [O I]!

$$P([\texttt{start}, \texttt{O}, \texttt{O}, \texttt{I}, \texttt{B}, \texttt{O}, \texttt{end}]) = P(\texttt{O}|\texttt{start}) \cdot P(\texttt{O}|\texttt{O}) \cdot \underbrace{P(\texttt{I}|\texttt{O})}_{=0} \cdot \ldots = 0$$

Here, this is correct; other times it may be just due to insufficient data.

Estimating Markov Models: Maximum Likelihood

$$P(y_i = \texttt{b} | y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a}, \texttt{b}]}{\sum_{\texttt{b'} \in \Sigma} \#[\texttt{a}, \texttt{b'}]} = \frac{\#[\texttt{a}, \texttt{b}]}{\#\texttt{a}}$$

$P(\texttt{B}\|\texttt{start}) = .393$	$P(\texttt{B}\|\texttt{B}) = .009$	$P(\texttt{B}\|\texttt{I}) = .003$	$P(\texttt{B}\|\texttt{O}) = .102$
$P(\texttt{I}\|\texttt{start}) = .0$	$P(\texttt{I}\|\texttt{B}) = .369$	$P(\texttt{I}\|\texttt{I}) = .178$	$P(\texttt{I}\|\texttt{O}) = .0$
$P(\texttt{O}\|\texttt{start}) = .607$	$P(\texttt{O}\|\texttt{B}) = .610$	$P(\texttt{O}\|\texttt{I}) = .779$	$P(\texttt{O}\|\texttt{O}) = .815$
$P(\texttt{end}\|\texttt{start}) = .0$	$P(\texttt{end}\|\texttt{B}) = .013$	$P(\texttt{end}\|\texttt{I}) = .040$	$P(\texttt{end}\|\texttt{O}) = .084$

We never saw [O I], so $P(\texttt{I}|\texttt{O}) = 0$.
The model rules out **any sequence containing** [O I]!

$$P([\texttt{start}, \texttt{O}, \texttt{O}, \texttt{I}, \texttt{B}, \texttt{O}, \texttt{end}]) = P(\texttt{O}|\texttt{start}) \cdot P(\texttt{O}|\texttt{O}) \cdot \underbrace{P(\texttt{I}|\texttt{O})}_{=0} \cdot \ldots = 0$$

Here, this is correct; other times it may be just due to insufficient data.
Smoothing (one way): pretend we saw each possible transition once more

$$P(y_i = \texttt{b} | y_{i-1} = \texttt{a}) = \frac{1 + \#[\texttt{a},\texttt{b}]}{\sum_{\texttt{b}} 1 + \#[\texttt{a},\texttt{b'}]} = \frac{1 + \#[\texttt{a},\texttt{b}]}{|\Sigma| + \#\texttt{a}}$$

What can we do with a Markov Model?

$$
\begin{aligned}
P(\texttt{B}|\texttt{start}) &= .393 \\
P(\texttt{I}|\texttt{start}) &= .0 \\
P(\texttt{O}|\texttt{start}) &= .607 \\
P(\texttt{end}|\texttt{start}) &= .0
\end{aligned}
\qquad
\begin{aligned}
P(\texttt{B}|\texttt{B}) &= .009 \\
P(\texttt{I}|\texttt{B}) &= .369 \\
P(\texttt{O}|\texttt{B}) &= .610 \\
P(\texttt{end}|\texttt{B}) &= .013
\end{aligned}
\qquad
\begin{aligned}
P(\texttt{B}|\texttt{I}) &= .003 \\
P(\texttt{I}|\texttt{I}) &= .178 \\
P(\texttt{O}|\texttt{I}) &= .779 \\
P(\texttt{end}|\texttt{I}) &= .040
\end{aligned}
\qquad
\begin{aligned}
P(\texttt{B}|\texttt{O}) &= .102 \\
P(\texttt{I}|\texttt{O}) &= .0 \\
P(\texttt{O}|\texttt{O}) &= .815 \\
P(\texttt{end}|\texttt{O}) &= .084
\end{aligned}
$$

What can we do with a Markov Model?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\ \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}\vert\texttt{start})\,\mathrm{P}(\texttt{O}\vert\texttt{B})\,\mathrm{P}(\texttt{O}\vert\texttt{O})\,\mathrm{P}(\texttt{end}\vert\texttt{O}) \\ &= .393 \cdot .610 \cdot .815 \cdot .084 = .0164\end{aligned}$$

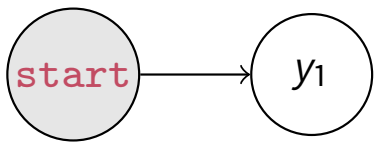
What can we do with a Markov Model?

$$
\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}
$$

- Given a sequence, assess its likelihood under the model.

$$
\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}
$$

- Sample sequences, going from left to right!



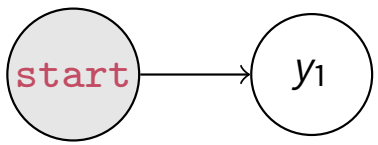
What can we do with a Markov Model?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\ \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}\vert\texttt{start})\,\mathrm{P}(\texttt{O}\vert\texttt{B})\,\mathrm{P}(\texttt{O}\vert\texttt{O})\,\mathrm{P}(\texttt{end}\vert\texttt{O}) \\ &= .393 \cdot .610 \cdot .815 \cdot .084 = .0164\end{aligned}$$

- Sample sequences, going from left to right!



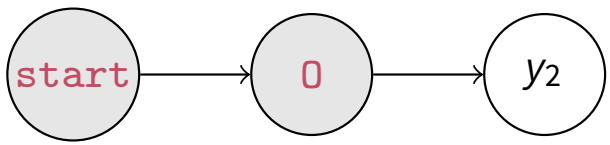
What can we do with a Markov Model?

$$\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}$$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}$$

- Sample sequences, going from left to right!



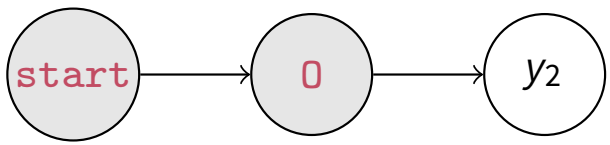
What can we do with a Markov Model?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned} \boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\ \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}\vert\texttt{start})\,\mathrm{P}(\texttt{O}\vert\texttt{B})\,\mathrm{P}(\texttt{O}\vert\texttt{O})\,\mathrm{P}(\texttt{end}\vert\texttt{O}) \\ &= .393 \cdot .610 \cdot .815 \cdot .084 = .0164 \end{aligned}$$

- Sample sequences, going from left to right!



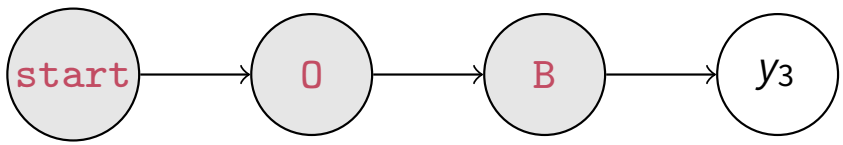
What can we do with a Markov Model?

$$\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}$$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}$$

- Sample sequences, going from left to right!



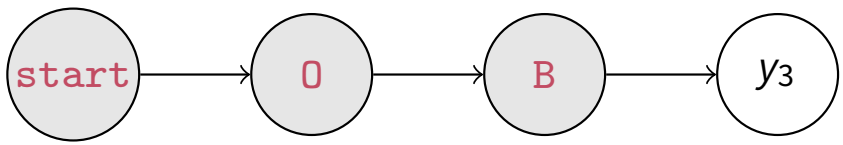
What can we do with a Markov Model?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\ \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}\vert\texttt{start})\,\mathrm{P}(\texttt{O}\vert\texttt{B})\,\mathrm{P}(\texttt{O}\vert\texttt{O})\,\mathrm{P}(\texttt{end}\vert\texttt{O}) \\ &= .393 \cdot .610 \cdot .815 \cdot .084 = .0164\end{aligned}$$

- Sample sequences, going from left to right!



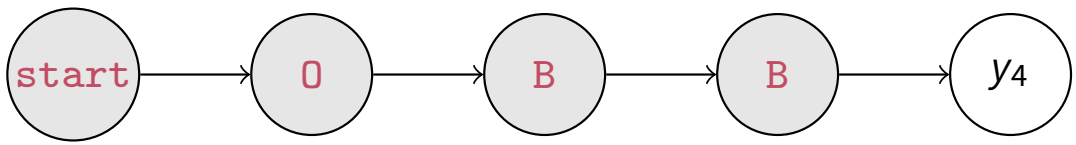
What can we do with a Markov Model?

$$\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}$$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}$$

- Sample sequences, going from left to right!



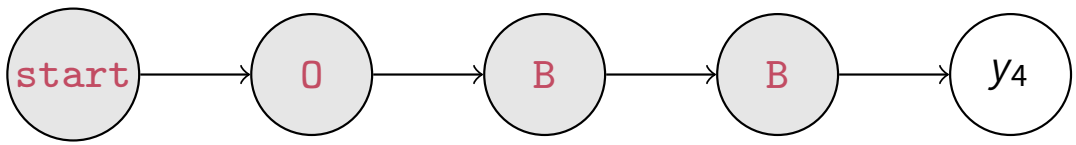
What can we do with a Markov Model?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\ \mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}\vert\texttt{start})\,\mathrm{P}(\texttt{O}\vert\texttt{B})\,\mathrm{P}(\texttt{O}\vert\texttt{O})\,\mathrm{P}(\texttt{end}\vert\texttt{O}) \\ &= .393 \cdot .610 \cdot .815 \cdot .084 = .0164\end{aligned}$$

- Sample sequences, going from left to right!



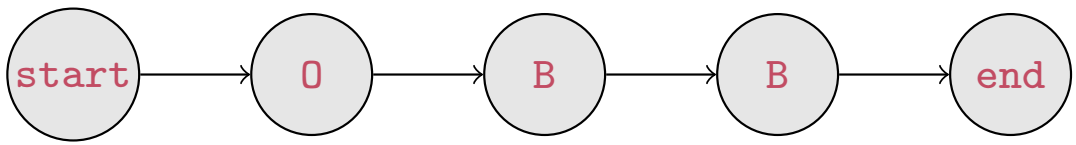
What can we do with a Markov Model?

$$\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}$$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}$$

- Sample sequences, going from left to right!



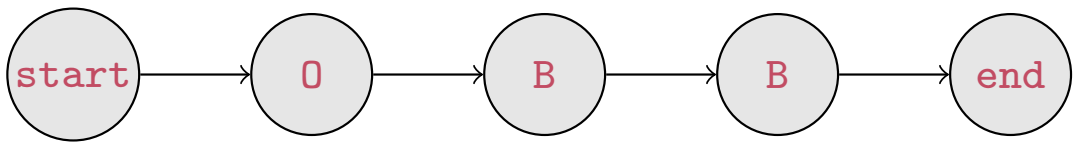
What can we do with a Markov Model?

$$\begin{aligned}
\mathrm{P}(\texttt{B}|\texttt{start}) &= .393 & \mathrm{P}(\texttt{B}|\texttt{B}) &= .009 & \mathrm{P}(\texttt{B}|\texttt{I}) &= .003 & \mathrm{P}(\texttt{B}|\texttt{O}) &= .102 \\
\mathrm{P}(\texttt{I}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{I}|\texttt{B}) &= .369 & \mathrm{P}(\texttt{I}|\texttt{I}) &= .178 & \mathrm{P}(\texttt{I}|\texttt{O}) &= .0 \\
\mathrm{P}(\texttt{O}|\texttt{start}) &= .607 & \mathrm{P}(\texttt{O}|\texttt{B}) &= .610 & \mathrm{P}(\texttt{O}|\texttt{I}) &= .779 & \mathrm{P}(\texttt{O}|\texttt{O}) &= .815 \\
\mathrm{P}(\texttt{end}|\texttt{start}) &= .0 & \mathrm{P}(\texttt{end}|\texttt{B}) &= .013 & \mathrm{P}(\texttt{end}|\texttt{I}) &= .040 & \mathrm{P}(\texttt{end}|\texttt{O}) &= .084
\end{aligned}$$

- Given a sequence, assess its likelihood under the model.

$$\begin{aligned}
\boldsymbol{y} &= [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{O}, \texttt{end}] \\
\mathrm{P}(\boldsymbol{y}) &= \mathrm{P}(\texttt{B}|\texttt{start})\,\mathrm{P}(\texttt{O}|\texttt{B})\,\mathrm{P}(\texttt{O}|\texttt{O})\,\mathrm{P}(\texttt{end}|\texttt{O}) \\
&= .393 \cdot .610 \cdot .815 \cdot .084 = .0164
\end{aligned}$$

- Sample sequences, going from left to right!



- Predict the most likely next symbol (like your phone's autocomplete)!
  (As above, but take max instead of a random sample.

What can we do with a MM? Incomplete data warm-up

$P(\texttt{B}\mid\texttt{start}) = .393$	$P(\texttt{B}\mid\texttt{B}) = .009$	$P(\texttt{B}\mid\texttt{I}) = .003$	$P(\texttt{B}\mid\texttt{O}) = .102$
$P(\texttt{I}\mid\texttt{start}) = .0$	$P(\texttt{I}\mid\texttt{B}) = .369$	$P(\texttt{I}\mid\texttt{I}) = .178$	$P(\texttt{I}\mid\texttt{O}) = .0$
$P(\texttt{O}\mid\texttt{start}) = .607$	$P(\texttt{O}\mid\texttt{B}) = .610$	$P(\texttt{O}\mid\texttt{I}) = .779$	$P(\texttt{O}\mid\texttt{O}) = .815$
$P(\texttt{end}\mid\texttt{start}) = .0$	$P(\texttt{end}\mid\texttt{B}) = .013$	$P(\texttt{end}\mid\texttt{I}) = .040$	$P(\texttt{end}\mid\texttt{O}) = .084$

- What is the probability that the *second* symbol in a sequence is B?

What can we do with a MM? Incomplete data warm-up

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- What is the probability that the *second* symbol in a sequence is B?
  $\mathrm{P}([\texttt{start}, ?, \texttt{B}]) = \mathrm{P}(y_0 = \texttt{start}, y_2 = \texttt{B}) = ?$

What can we do with a MM? Incomplete data warm-up

$P(\texttt{B}\mid\texttt{start}) = .393$	$P(\texttt{B}\mid\texttt{B}) = .009$	$P(\texttt{B}\mid\texttt{I}) = .003$	$P(\texttt{B}\mid\texttt{O}) = .102$
$P(\texttt{I}\mid\texttt{start}) = .0$	$P(\texttt{I}\mid\texttt{B}) = .369$	$P(\texttt{I}\mid\texttt{I}) = .178$	$P(\texttt{I}\mid\texttt{O}) = .0$
$P(\texttt{O}\mid\texttt{start}) = .607$	$P(\texttt{O}\mid\texttt{B}) = .610$	$P(\texttt{O}\mid\texttt{I}) = .779$	$P(\texttt{O}\mid\texttt{O}) = .815$
$P(\texttt{end}\mid\texttt{start}) = .0$	$P(\texttt{end}\mid\texttt{B}) = .013$	$P(\texttt{end}\mid\texttt{I}) = .040$	$P(\texttt{end}\mid\texttt{O}) = .084$

- What is the probability that the *second* symbol in a sequence is B?
  $P([\texttt{start}, ?, \texttt{B}]) = P(y_0 = \texttt{start}, y_2 = \texttt{B}) = ?$
- We must consider all possible choices for y_1. Recall $P(b) = \sum_a P(a, b)$!

$$\begin{aligned} P([\texttt{start}, ?, \texttt{B}]) &= \sum_{\texttt{y} \in \Sigma} P([\texttt{start}, \texttt{y}, \texttt{B}]) \\ &= P([\texttt{start}, \texttt{B}, \texttt{B}]) + P([\texttt{start}, \texttt{I}, \texttt{B}]) + P([\texttt{start}, \texttt{O}, \texttt{B}]) \\ &= .393 \cdot .009 + .0 \cdot .003 + .607 \cdot .102 = .065 \end{aligned}$$

What can we do with a MM? Incomplete data warm-up

$P(\texttt{B}\mid\texttt{start}) = .393$	$P(\texttt{B}\mid\texttt{B}) = .009$	$P(\texttt{B}\mid\texttt{I}) = .003$	$P(\texttt{B}\mid\texttt{O}) = .102$
$P(\texttt{I}\mid\texttt{start}) = .0$	$P(\texttt{I}\mid\texttt{B}) = .369$	$P(\texttt{I}\mid\texttt{I}) = .178$	$P(\texttt{I}\mid\texttt{O}) = .0$
$P(\texttt{O}\mid\texttt{start}) = .607$	$P(\texttt{O}\mid\texttt{B}) = .610$	$P(\texttt{O}\mid\texttt{I}) = .779$	$P(\texttt{O}\mid\texttt{O}) = .815$
$P(\texttt{end}\mid\texttt{start}) = .0$	$P(\texttt{end}\mid\texttt{B}) = .013$	$P(\texttt{end}\mid\texttt{I}) = .040$	$P(\texttt{end}\mid\texttt{O}) = .084$

- What is the probability that the *second* symbol in a sequence is B?
  $P([\texttt{start}, ?, \texttt{B}]) = P(y_0 = \texttt{start}, y_2 = \texttt{B}) = ?$
- We must consider all possible choices for y_1. Recall $P(b) = \sum_a P(a, b)$!

$$\begin{aligned} P([\texttt{start}, ?, \texttt{B}]) &= \sum_{\texttt{y} \in \Sigma} P([\texttt{start}, \texttt{y}, \texttt{B}]) \\ &= P([\texttt{start}, \texttt{B}, \texttt{B}]) + P([\texttt{start}, \texttt{I}, \texttt{B}]) + P([\texttt{start}, \texttt{O}, \texttt{B}]) \\ &= .393 \cdot .009 + .0 \cdot .003 + .607 \cdot .102 = .065 \end{aligned}$$

- How about the proba of B as *third* symbol?

What can we do with a MM? Incomplete data warm-up

$P(\texttt{B}\mid\texttt{start}) = .393$	$P(\texttt{B}\mid\texttt{B}) = .009$	$P(\texttt{B}\mid\texttt{I}) = .003$	$P(\texttt{B}\mid\texttt{O}) = .102$
$P(\texttt{I}\mid\texttt{start}) = .0$	$P(\texttt{I}\mid\texttt{B}) = .369$	$P(\texttt{I}\mid\texttt{I}) = .178$	$P(\texttt{I}\mid\texttt{O}) = .0$
$P(\texttt{O}\mid\texttt{start}) = .607$	$P(\texttt{O}\mid\texttt{B}) = .610$	$P(\texttt{O}\mid\texttt{I}) = .779$	$P(\texttt{O}\mid\texttt{O}) = .815$
$P(\texttt{end}\mid\texttt{start}) = .0$	$P(\texttt{end}\mid\texttt{B}) = .013$	$P(\texttt{end}\mid\texttt{I}) = .040$	$P(\texttt{end}\mid\texttt{O}) = .084$

- What is the probability that the *second* symbol in a sequence is B?
  $P([\texttt{start}, ?, \texttt{B}]) = P(y_0 = \texttt{start}, y_2 = \texttt{B}) = ?$
- We must consider all possible choices for y_1. Recall $P(b) = \sum_a P(a, b)$!

$$\begin{aligned} P([\texttt{start}, ?, \texttt{B}]) &= \sum_{\texttt{y}\in\Sigma} P([\texttt{start}, \texttt{y}, \texttt{B}]) \\ &= P([\texttt{start}, \texttt{B}, \texttt{B}]) + P([\texttt{start}, \texttt{I}, \texttt{B}]) + P([\texttt{start}, \texttt{O}, \texttt{B}]) \\ &= .393 \cdot .009 + .0 \cdot .003 + .607 \cdot .102 = .065 \end{aligned}$$

- How about the proba of B as *third* symbol?

$$P([\texttt{start}, ?, ?, \texttt{B}]) = \sum_{\texttt{y}_1\in\Sigma} \sum_{\texttt{y}_2\in\Sigma} P([\texttt{start}, \texttt{y}_1, \texttt{y}_2, \texttt{B}]) = P([\texttt{start}, \texttt{B}, \texttt{B}, \texttt{B}]) + P([\texttt{start}, \texttt{B}, \texttt{I}, \texttt{B}]) + ..$$

$$\begin{aligned} \text{(regrouping)} \quad &= P([\texttt{start}, ?, \texttt{B}, \texttt{B}]) + P([\texttt{start}, ?, \texttt{I}, \texttt{B}]) + P([\texttt{start}, ?, \texttt{O}, \texttt{B}]) \\ &= P([\texttt{start}, ?, \texttt{B}])\, P(\texttt{B}|\texttt{B}) + P([\texttt{start}, ?, \texttt{I}])\, P(\texttt{B}|\texttt{I}) + P([\texttt{start}, ?, \texttt{O}])\, P(\texttt{B}|\texttt{O}) \\ &= \sum_{\texttt{y}_2\in\Sigma} P([\texttt{start}, ?, \texttt{y}_2])\, P(\texttt{y}_2|\texttt{B}) \qquad \text{(A pattern we'll see again soon!)} \end{aligned}$$

What can we do with a MM?

$P(\texttt{B}\|\texttt{start}) = .393$	$P(\texttt{B}\|\texttt{B}) = .009$	$P(\texttt{B}\|\texttt{I}) = .003$	$P(\texttt{B}\|\texttt{O}) = .102$
$P(\texttt{I}\|\texttt{start}) = .0$	$P(\texttt{I}\|\texttt{B}) = .369$	$P(\texttt{I}\|\texttt{I}) = .178$	$P(\texttt{I}\|\texttt{O}) = .0$
$P(\texttt{O}\|\texttt{start}) = .607$	$P(\texttt{O}\|\texttt{B}) = .610$	$P(\texttt{O}\|\texttt{I}) = .779$	$P(\texttt{O}\|\texttt{O}) = .815$
$P(\texttt{end}\|\texttt{start}) = .0$	$P(\texttt{end}\|\texttt{B}) = .013$	$P(\texttt{end}\|\texttt{I}) = .040$	$P(\texttt{end}\|\texttt{O}) = .084$

- Can we do NER?

start	B	O	O	end
	Halloween	is	coming	

What can we do with a MM?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Can we do NER?

start	B	O	O	end
	Halloween	is	coming	

Not well! — a model of $\mathrm{P}(\boldsymbol{y})$ does not take $\boldsymbol{x}$ into account!

What can we do with a MM?

$\mathrm{P}(\texttt{B}\vert\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\vert\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\vert\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\vert\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\vert\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\vert\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\vert\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\vert\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\vert\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\vert\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\vert\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\vert\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\vert\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\vert\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\vert\texttt{O}) = .084$

- Can we do NER?

start	B	O	O	end
	Halloween	is	coming	

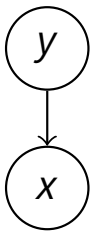
  Not well! — a model of $\mathrm{P}(\boldsymbol{y})$ does not take $\boldsymbol{x}$ into account!

- We want a model that can take $\boldsymbol{x}$ into account, too.

Outline

Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

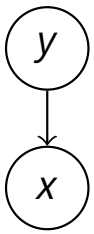
- Observe $\boldsymbol{x}$ and use Bayes' rule:
  $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$$

- For NER, a form of NB:

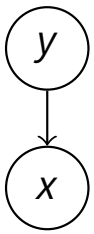
$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	.0
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	.0
$\mathrm{P}(\text{under}\|y) =$	.0	.0	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	.0	.0002	.0
$\mathrm{P}(\text{of}\|y) =$	.0	.0127	.0212
...			

Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:

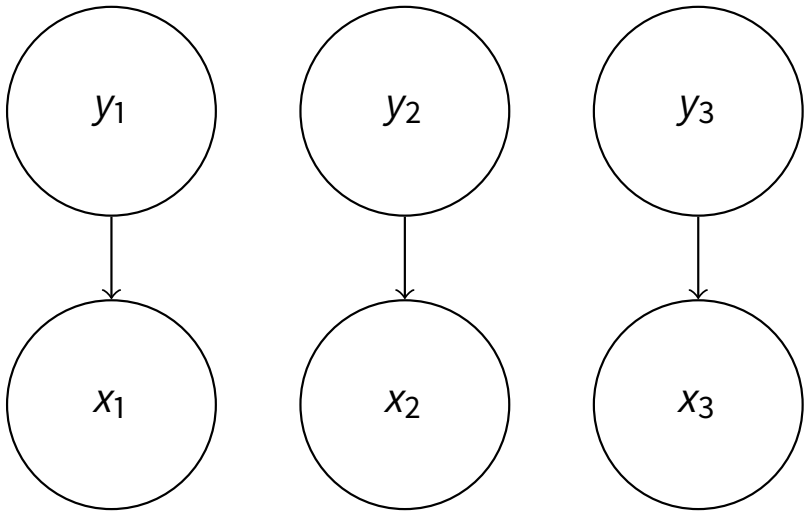
$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$$

- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

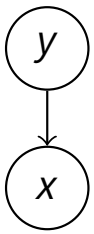
$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	.0
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	.0
$\mathrm{P}(\text{under}\|y) =$	.0	.0	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	.0	.0002	.0
$\mathrm{P}(\text{of}\|y) =$	.0	.0127	.0212
...			



Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

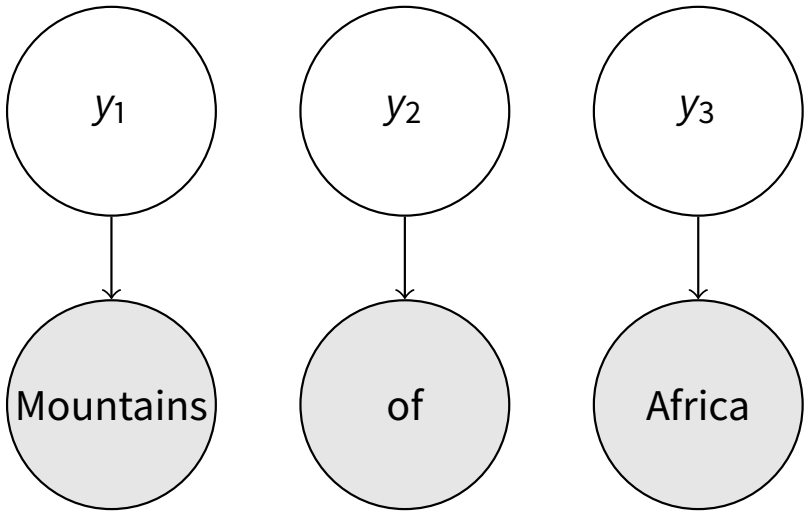
$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:
  $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

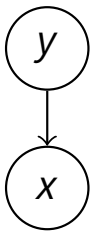
$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	.0
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	.0
$\mathrm{P}(\text{under}\|y) =$	.0	.0	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	.0	.0002	.0
$\mathrm{P}(\text{of}\|y) =$	.0	.0127	.0212
...			



Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:

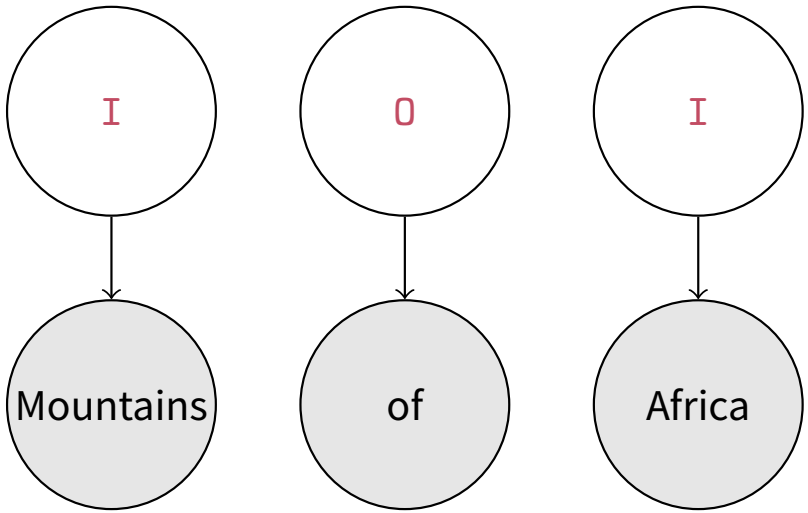
$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$$

- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

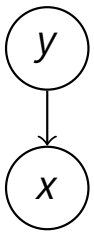
$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	.0
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	.0
$\mathrm{P}(\text{under}\|y) =$	.0	.0	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	.0	.0002	.0
$\mathrm{P}(\text{of}\|y) =$	.0	.0127	.0212
...			



Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

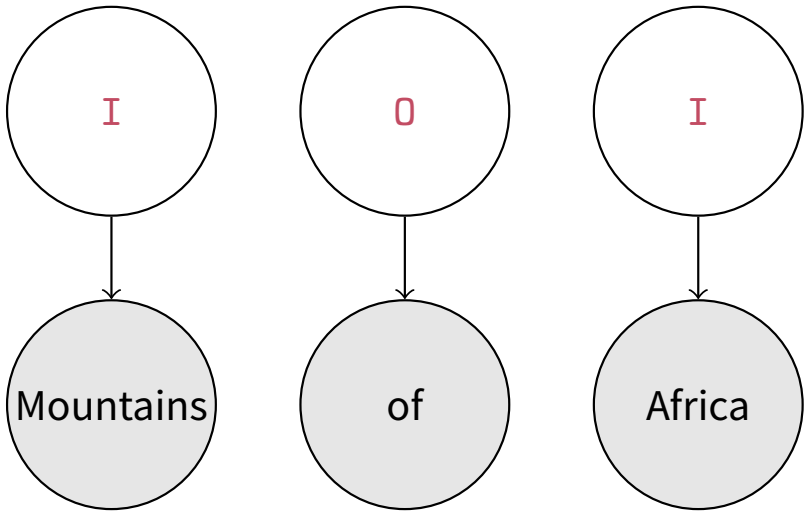
- Observe $\boldsymbol{x}$ and use Bayes' rule:
  $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

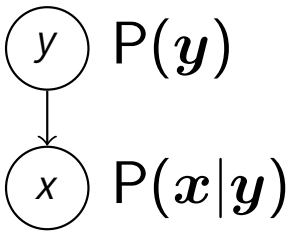
Why so bad?

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	.0
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	.0
$\mathrm{P}(\text{under}\|y) =$	.0	.0	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	.0	.0002	.0
$\mathrm{P}(\text{of}\|y) =$	.0	.0127	.0212
...			



Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

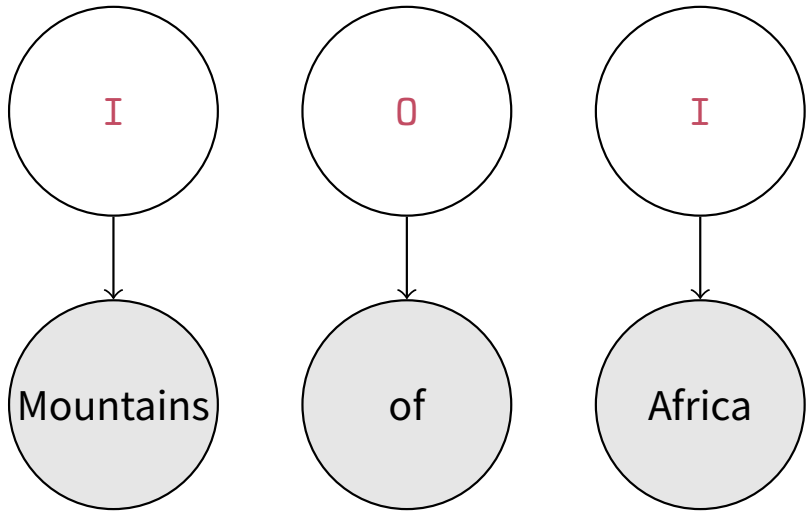
$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:
  $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

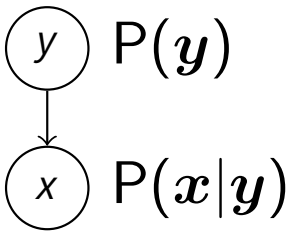
	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	ϵ
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	ϵ
$\mathrm{P}(\text{under}\|y) =$	ϵ	ϵ	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$\mathrm{P}(\text{of}\|y) =$	ϵ	.0127	.0212
…			



Why so bad? Smoothing doesn't fix it.

Remember Naïve Bayes

- We want $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$ for prediction.
- We build two simpler models:



$\mathrm{P}(\boldsymbol{y})$

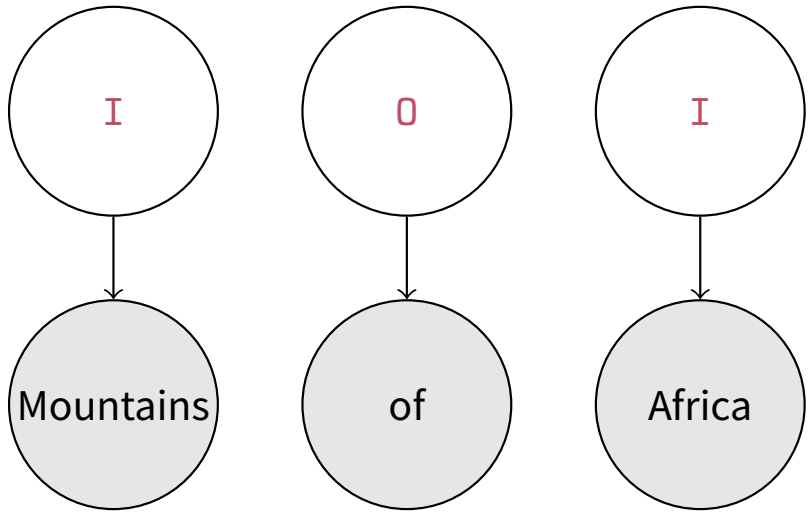
$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$

- Observe $\boldsymbol{x}$ and use Bayes' rule:
  $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \mathrm{P}(\boldsymbol{y})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
- For NER, a form of NB:

$$\mathrm{P}(\boldsymbol{y}) = \prod_i \mathrm{P}(y_i)$$

$$\mathrm{P}(\boldsymbol{x}|\boldsymbol{y}) = \prod_i \mathrm{P}(x_i|y_i)$$

	B	I	O
$\mathrm{P}(y) =$	.115	.052	.833
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	ϵ
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	ϵ
$\mathrm{P}(\text{under}\|y) =$	ϵ	ϵ	.0009
$\mathrm{P}(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$\mathrm{P}(\text{of}\|y) =$	ϵ	.0127	.0212
…			

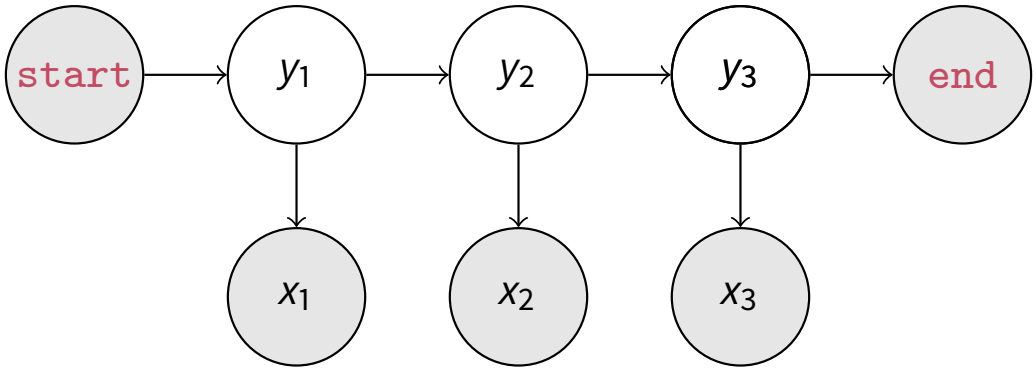


Why so bad? Smoothing doesn't fix it.

$\mathrm{P}([\texttt{I O I}]) = .225 \gg 0!$

Hidden Markov Models

- Jointly model a sequence of observations x_i and **hidden states** y_i.
- States modeled by a first-order Markov model.
- Each observation is conditioned only on the corresponding state.



$$\boldsymbol{x} = [x_1, \ldots, x_L]; \quad \boldsymbol{y} = [\texttt{start}, y_1, \ldots, y_L, \texttt{end}].$$

$$\mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) = \underbrace{\prod_{i=1}^{L+1} \mathrm{P}(y_i|y_{i-1})}_{\mathrm{P}(\boldsymbol{y})} \cdot \underbrace{\prod_{i=1}^{L} \mathrm{P}(x_i|y_i)}_{\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})}$$

Estimating HMMs: Maximum Likelihood

$$\text{maximize} \left(\prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y}) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \prod_{i=1}^{L(\boldsymbol{x})+1} \mathrm{P}(y_i|y_{i-1}) \cdot \prod_{i=1}^{L(\boldsymbol{x})} \mathrm{P}(x_i|y_i) \right)$$

A HMM is “just” a Markov model and an emission model :P

Estimating HMMs: Maximum Likelihood

$$\text{maximize} \left(\prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y}) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \prod_{i=1}^{L(\boldsymbol{x})+1} \mathrm{P}(y_i|y_{i-1}) \cdot \prod_{i=1}^{L(\boldsymbol{x})} \mathrm{P}(x_i|y_i) \right)$$

A HMM is "just" a Markov model and an emission model :P

Transition probabilities: $\mathrm{P}(y_i = \texttt{b}|y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a},\texttt{b}]}{\#\texttt{a}}$ *just like the MM!*

$\mathrm{P}(\texttt{B}\|\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\|\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\|\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\|\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\|\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\|\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\|\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\|\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\|\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\|\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\|\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\|\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\|\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\|\texttt{O}) = .084$

Estimating HMMs: Maximum Likelihood

$$\text{maximize} \left(\prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y}) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \prod_{i=1}^{L(\boldsymbol{x})+1} \mathrm{P}(y_i|y_{i-1}) \cdot \prod_{i=1}^{L(\boldsymbol{x})} \mathrm{P}(x_i|y_i) \right)$$

A HMM is "just" a Markov model and an emission model :P

Transition probabilities: $\mathrm{P}(y_i = \texttt{b}|y_{i-1} = \texttt{a}) = \frac{\#[\texttt{a},\texttt{b}]}{\#\texttt{a}}$ *just like the MM!*

$\mathrm{P}(\texttt{B}\|\texttt{start}) = .393$	$\mathrm{P}(\texttt{B}\|\texttt{B}) = .009$	$\mathrm{P}(\texttt{B}\|\texttt{I}) = .003$	$\mathrm{P}(\texttt{B}\|\texttt{O}) = .102$
$\mathrm{P}(\texttt{I}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{I}\|\texttt{B}) = .369$	$\mathrm{P}(\texttt{I}\|\texttt{I}) = .178$	$\mathrm{P}(\texttt{I}\|\texttt{O}) = .0$
$\mathrm{P}(\texttt{O}\|\texttt{start}) = .607$	$\mathrm{P}(\texttt{O}\|\texttt{B}) = .610$	$\mathrm{P}(\texttt{O}\|\texttt{I}) = .779$	$\mathrm{P}(\texttt{O}\|\texttt{O}) = .815$
$\mathrm{P}(\texttt{end}\|\texttt{start}) = .0$	$\mathrm{P}(\texttt{end}\|\texttt{B}) = .013$	$\mathrm{P}(\texttt{end}\|\texttt{I}) = .040$	$\mathrm{P}(\texttt{end}\|\texttt{O}) = .084$

Emission probabilities: $\mathrm{P}(x_i = w|y_i = \texttt{a}) = \frac{\#\,\mathrm{tag}(w)=\texttt{a}}{\#\texttt{a}}$ *just like NB!*

	B	I	O
$\mathrm{P}(\text{South}\|y) =$	.0051	.0002	ϵ
$\mathrm{P}(\text{Africa}\|y) =$	.0003	.0056	ϵ
$\mathrm{P}(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$\mathrm{P}(\text{of}\|y) =$	ϵ	.0127	.0212

...

What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\ldots$			

What can we do with an HMM?

	start	B	I	O
$P(B\|y) =$	.393	.009	.003	.102
$P(I\|y) =$	.0	.369	.178	.0
$P(O\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

What can we do with an HMM?

	start	B	I	O
$P(B\|y) =$	.393	.009	.003	.102
$P(I\|y) =$	.0	.369	.178	.0
$P(O\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\ldots$			

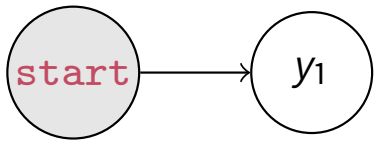
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(B\|y) =$	.393	.009	.003	.102
$P(I\|y) =$	.0	.369	.178	.0
$P(O\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\ldots$			

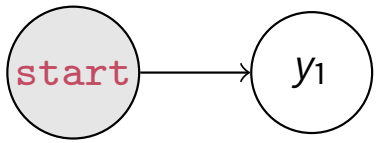
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

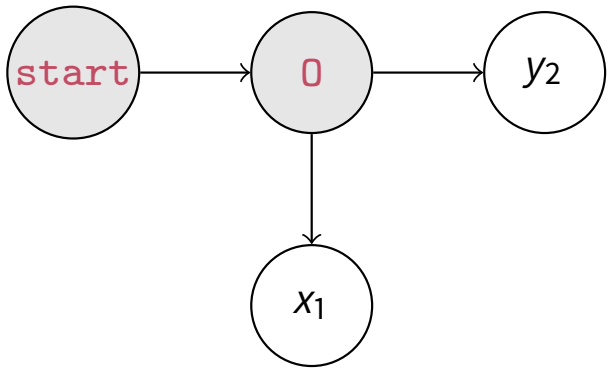
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

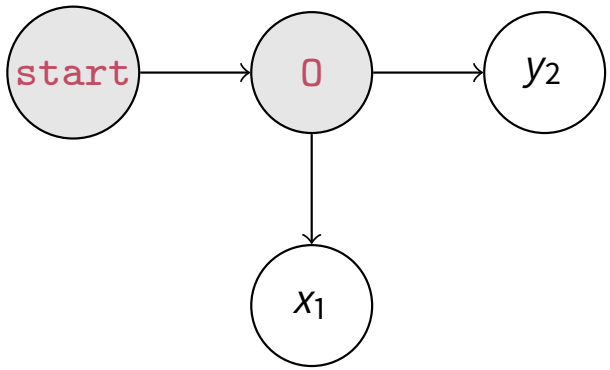
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

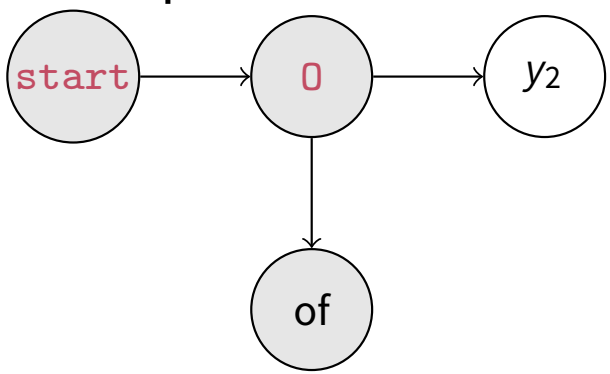
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

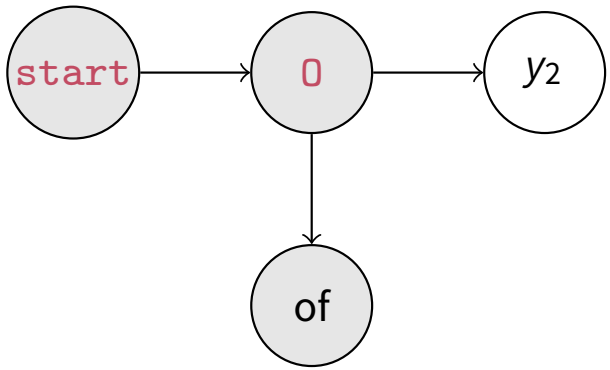
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
...			

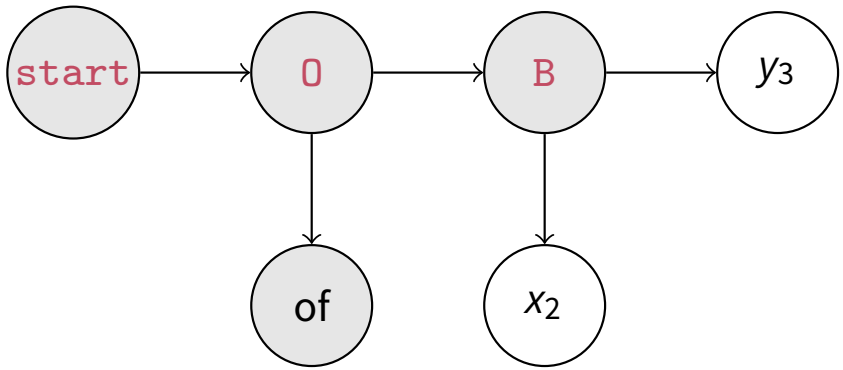
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

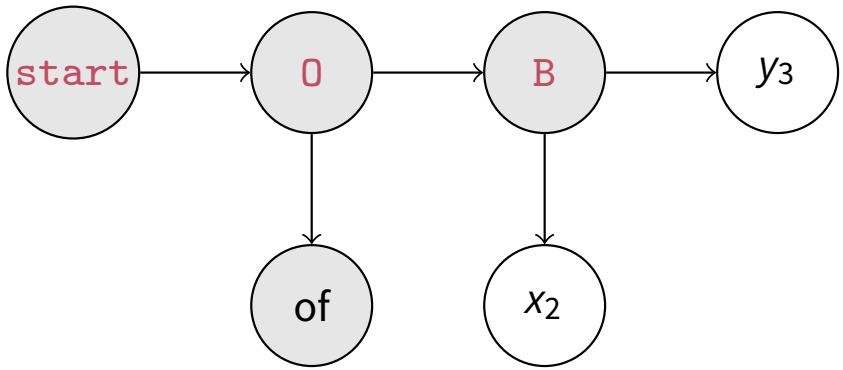
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains, of, Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
P(B|y) =	.393	.009	.003	.102
P(I|y) =	.0	.369	.178	.0
P(O|y) =	.607	.610	.779	.815
P(end|y) =	.0	.013	.040	.084

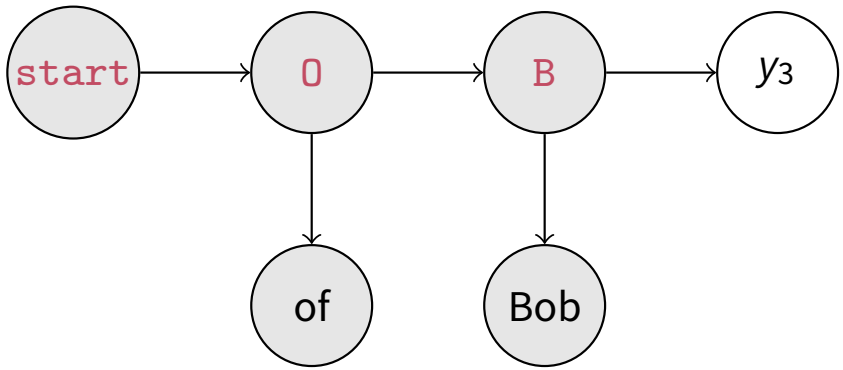
	B	I	O
P(South|y) =	.0051	.0002	ϵ
P(Africa|y) =	.0003	.0056	ϵ
P(Mountains|y) =	ϵ	.0002	ϵ
P(of|y) =	ϵ	.0127	.0212
...			

- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains, of, Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$\begin{aligned} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) = & \mathrm{P}(\texttt{B}|\texttt{start})\, \mathrm{P}(\texttt{O}|\texttt{B})\, \mathrm{P}(\texttt{B}|\texttt{O})\, \mathrm{P}(\texttt{end}|\texttt{B}) \\ & \cdot \mathrm{P}(\text{Mountains}|\texttt{B})\, \mathrm{P}(\text{of}|\texttt{O})\, \mathrm{P}(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17} \end{aligned}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(B\|y) =$	.393	.009	.003	.102
$P(I\|y) =$	.0	.369	.178	.0
$P(O\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

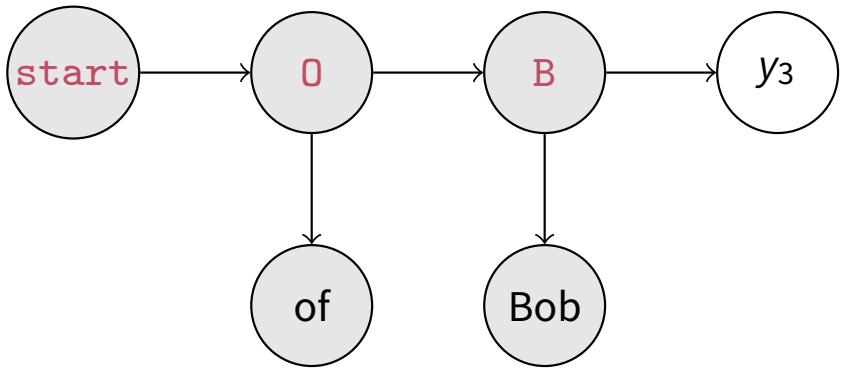
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains, of, Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

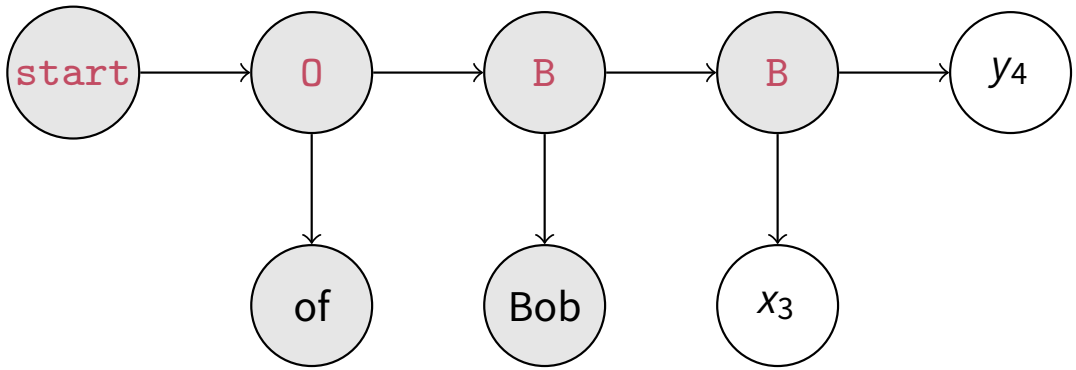
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(B\|y) =$	.393	.009	.003	.102
$P(I\|y) =$	.0	.369	.178	.0
$P(O\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

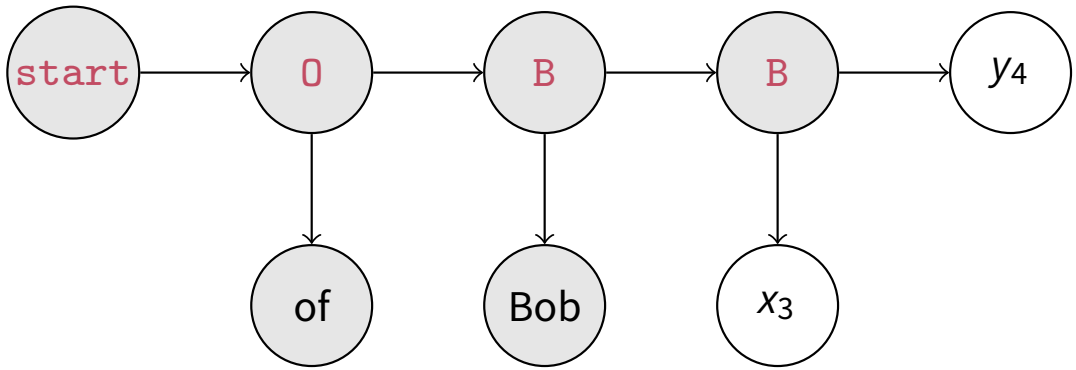
	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains, of, Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$
$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

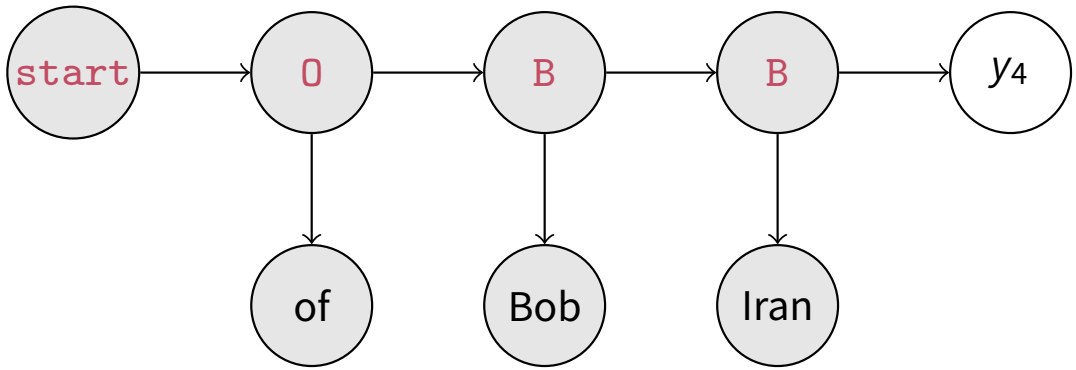
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

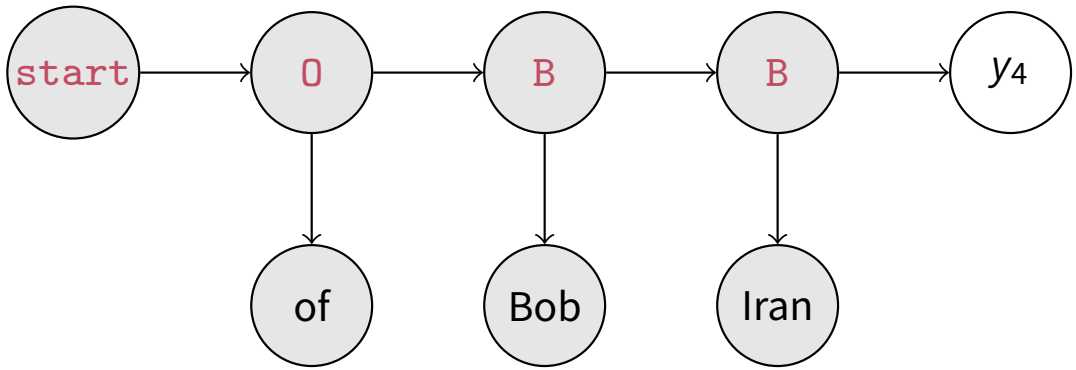
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

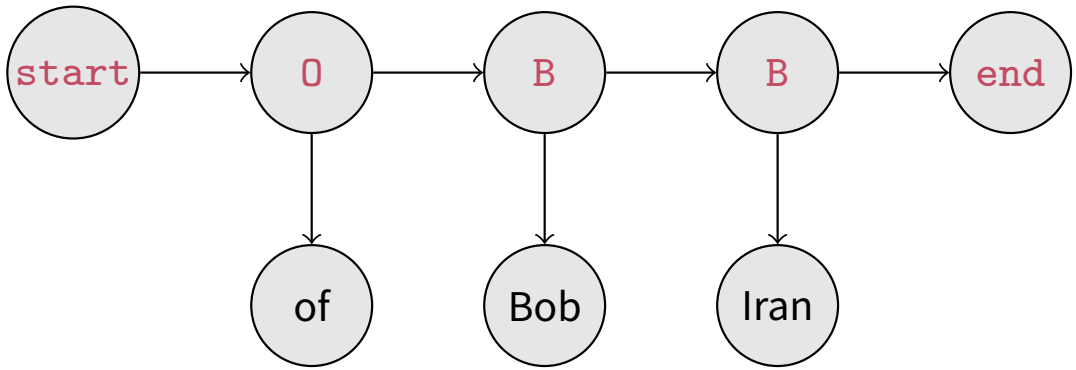
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



What can we do with an HMM?

	start	B	I	O
$P(\texttt{B}\|y) =$	.393	.009	.003	.102
$P(\texttt{I}\|y) =$	.0	.369	.178	.0
$P(\texttt{O}\|y) =$	.607	.610	.779	.815
$P(\texttt{end}\|y) =$	.0	.013	.040	.084

	B	I	O
$P(\text{South}\|y) =$	.0051	.0002	ϵ
$P(\text{Africa}\|y) =$	.0003	.0056	ϵ
$P(\text{Mountains}\|y) =$	ϵ	.0002	ϵ
$P(\text{of}\|y) =$	ϵ	.0127	.0212
$\dots$			

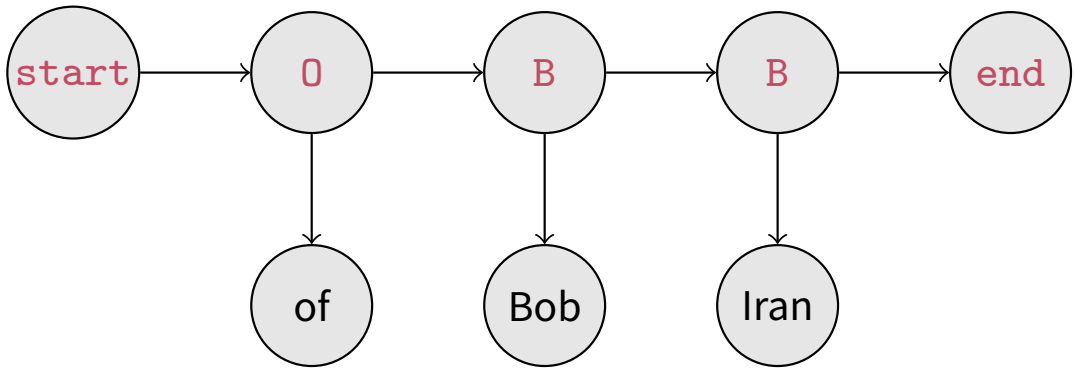
- Given a labeled sequence $(\boldsymbol{x}, \boldsymbol{y})$, assess its likelihood under the model.

$$\boldsymbol{x} = [\text{Mountains}, \text{of}, \text{Africa}]; \qquad \boldsymbol{y} = [\texttt{start}, \texttt{B}, \texttt{O}, \texttt{B}, \texttt{end}]$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(\texttt{B}|\texttt{start})\, P(\texttt{O}|\texttt{B})\, P(\texttt{B}|\texttt{O})\, P(\texttt{end}|\texttt{B})$$

$$\cdot\, P(\text{Mountains}|\texttt{B})\, P(\text{of}|\texttt{O})\, P(\text{Africa}|\texttt{B}) = 8 \cdot 10^{-17}$$

- Sample labelled sequences.



- Auto-complete equivalent: predict the most likely x_{k+1} given x_k, y_k.

Big HMM questions

1. What is the most likely label sequence $\boldsymbol{y}$, given $\boldsymbol{x}$?
2. What is the probability of $\boldsymbol{x}$?
3. What is the probability of each assignment y_i, given $\boldsymbol{x}$?
4. What is the probability of each transition $y \rightarrow y'$, given $\boldsymbol{x}$?
5. What sequence $\boldsymbol{y}$ minimizes the *Hamming cost?*

Big HMM questions: Finding the most likely y

	start	B	I	O
P(B|y) =	.393	.009	.003	.102
P(I|y) =	.0	.369	.178	.0
P(O|y) =	.607	.610	.779	.815
P(end|y) =	.0	.013	.040	.084

	B	I	O
P(South|y) =	.0051	.0002	ϵ
P(Africa|y) =	.0003	.0056	ϵ
P(Mountains|y) =	ϵ	.0002	ϵ
P(of|y) =	ϵ	.0127	.0212
...			

- Can we predict the most likely label sequence $\boldsymbol{y}$ for a given $\boldsymbol{x}$?

  $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

Big HMM questions: Finding the most likely y

	start	B	I	O
P(B|y) =	.393	.009	.003	.102
P(I|y) =	.0	.369	.178	.0
P(O|y) =	.607	.610	.779	.815
P(end|y) =	.0	.013	.040	.084

	B	I	O
P(South|y) =	.0051	.0002	ϵ
P(Africa|y) =	.0003	.0056	ϵ
P(Mountains|y) =	ϵ	.0002	ϵ
P(of|y) =	ϵ	.0127	.0212
...			

- Can we predict the most likely label sequence $\boldsymbol{y}$ for a given $\boldsymbol{x}$?

  $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \frac{\mathrm{P}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{P}(\boldsymbol{x})} = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y})$

Big HMM questions: Finding the most likely y

	start	B	I	O
P(B|y) =	.393	.009	.003	.102
P(I|y) =	.0	.369	.178	.0
P(O|y) =	.607	.610	.779	.815
P(end|y) =	.0	.013	.040	.084

	B	I	O
P(South|y) =	.0051	.0002	ϵ
P(Africa|y) =	.0003	.0056	ϵ
P(Mountains|y) =	ϵ	.0002	ϵ
P(of|y) =	ϵ	.0127	.0212
...			

- Can we predict the most likely label sequence $\boldsymbol{y}$ for a given $\boldsymbol{x}$?

  $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \frac{\mathrm{P}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{P}(\boldsymbol{x})} = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y})$

  One algo: enumeration (correct, but prohibitive):

  for $\boldsymbol{y} \in \Sigma^L$

  compute $P(\boldsymbol{x},\boldsymbol{y})$

  Return the highest found.

Big HMM questions: Finding the most likely y

	start	B	I	O
P(B|y) =	.393	.009	.003	.102
P(I|y) =	.0	.369	.178	.0
P(O|y) =	.607	.610	.779	.815
P(end|y) =	.0	.013	.040	.084

	B	I	O
P(South|y) =	.0051	.0002	ϵ
P(Africa|y) =	.0003	.0056	ϵ
P(Mountains|y) =	ϵ	.0002	ϵ
P(of|y) =	ϵ	.0127	.0212
...			

- Can we predict the most likely label sequence $\boldsymbol{y}$ for a given $\boldsymbol{x}$?

  $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \frac{\mathrm{P}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{P}(\boldsymbol{x})} = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x},\boldsymbol{y})$

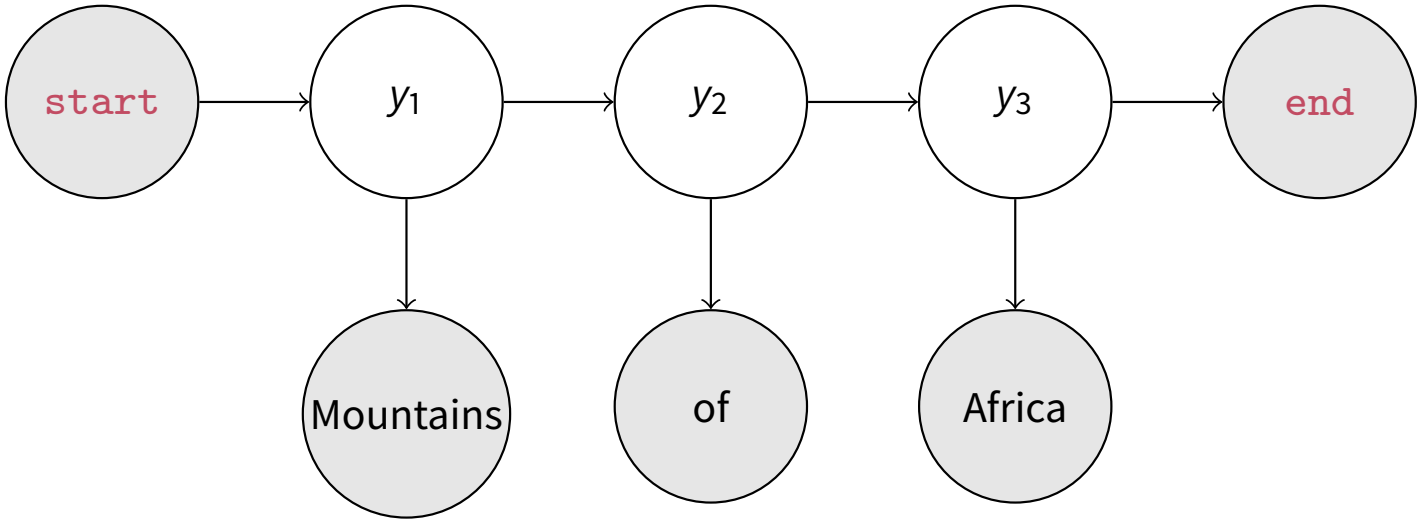
  One algo: enumeration (correct, but prohibitive):

  for $y_1 \in \Sigma$:
    for $y_2 \in \Sigma$:
      ...
      for $y_L \in \Sigma$:
        compute $P(\boldsymbol{x},\boldsymbol{y})$

  Return the highest found.

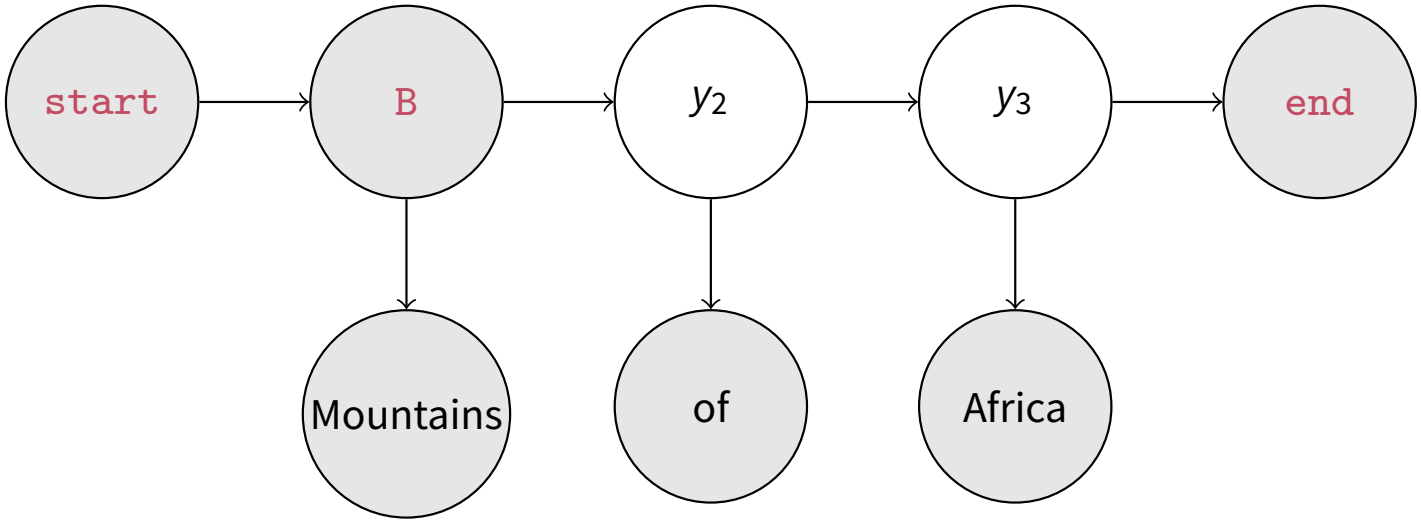
Big HMM questions: Finding the most likely y

Fast idea: **Greedy prediction!**



Big HMM questions: Finding the most likely y

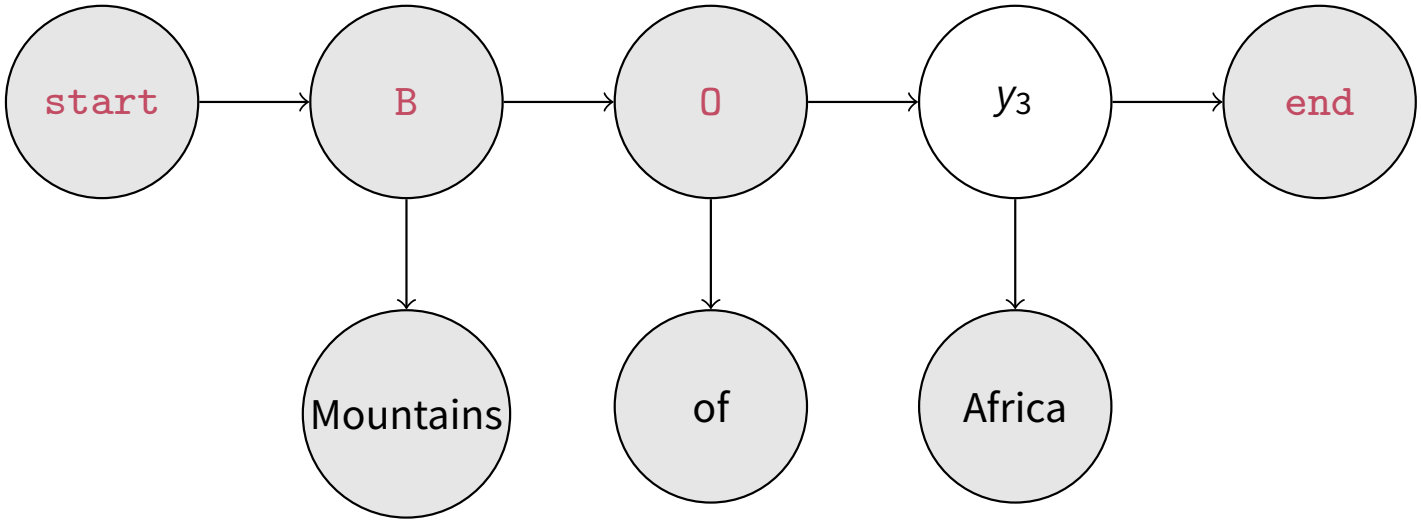
Fast idea: **Greedy prediction!**



- $y_1 = \arg\max_{y \in \Sigma} \mathrm{P}(y|\texttt{start})\,\mathrm{P}(x_1|y)$

Big HMM questions: Finding the most likely y

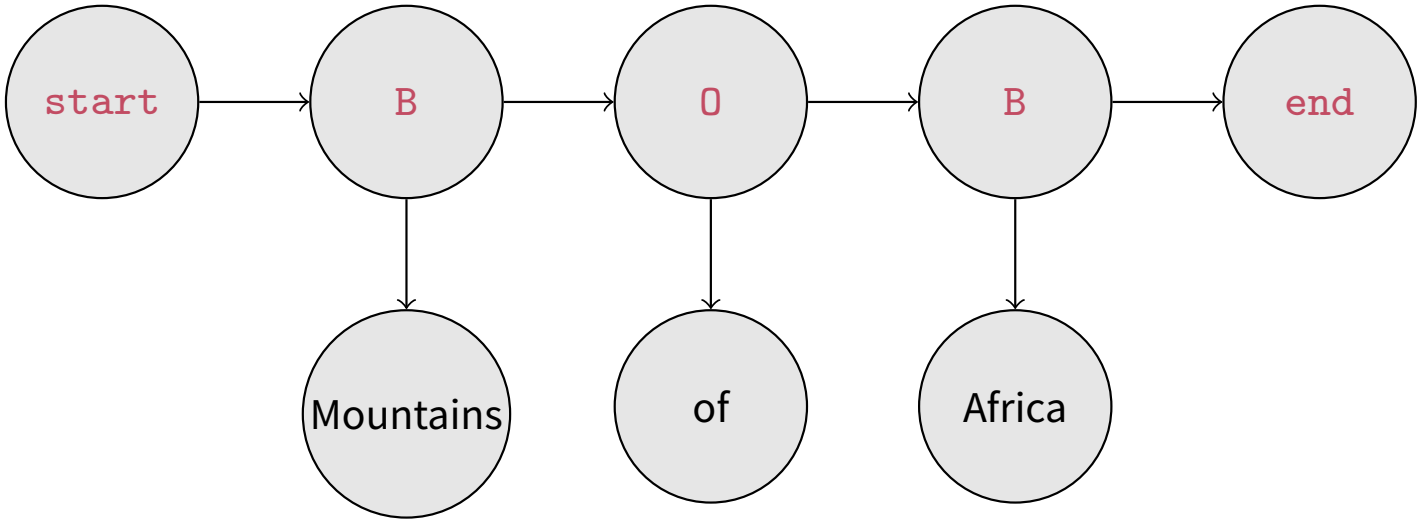
Fast idea: **Greedy prediction!**



- $y_1 = \arg\max_{y \in \Sigma} \mathrm{P}(y|\texttt{start})\,\mathrm{P}(x_1|y)$
- $y_2 = \arg\max_{y \in \Sigma} \mathrm{P}(y|y_1)\,\mathrm{P}(x_2|y)$

Big HMM questions: Finding the most likely y

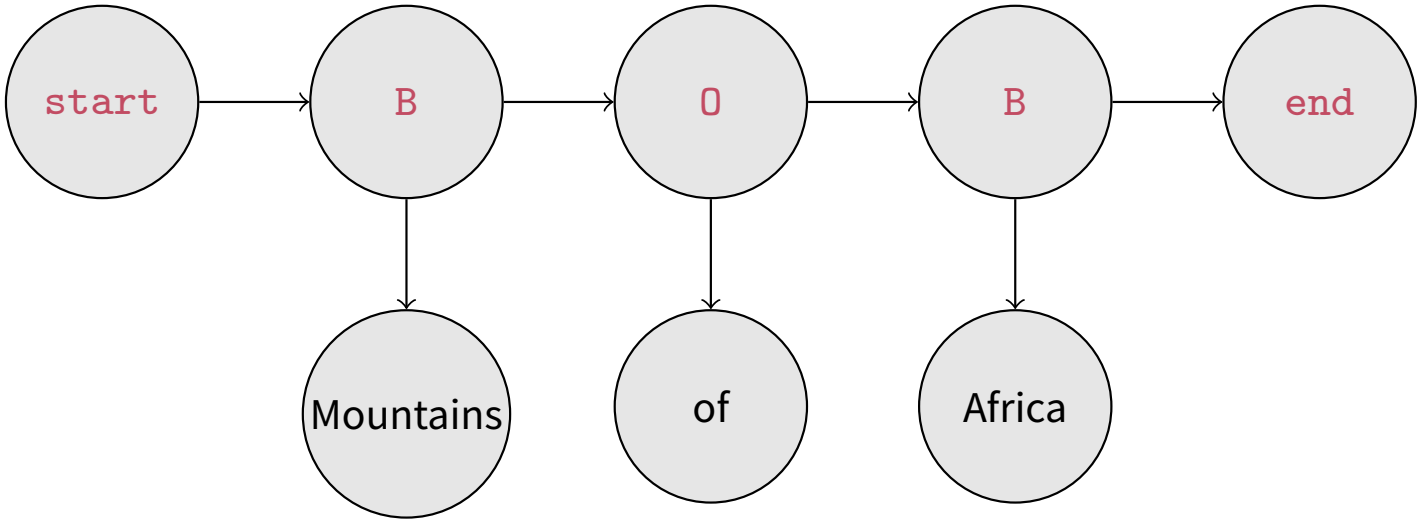
Fast idea: **Greedy prediction!**



- $y_1 = \arg\max_{y \in \Sigma} \mathrm{P}(y|\texttt{start})\,\mathrm{P}(x_1|y)$
- $y_2 = \arg\max_{y \in \Sigma} \mathrm{P}(y|y_1)\,\mathrm{P}(x_2|y)$
- $y_3 = \arg\max_{y \in \Sigma} \mathrm{P}(y|y_2)\,\mathrm{P}(x_3|y)\,\mathrm{P}(\texttt{end}|y)$

Big HMM questions: Finding the most likely y

Fast idea: **Greedy prediction!**

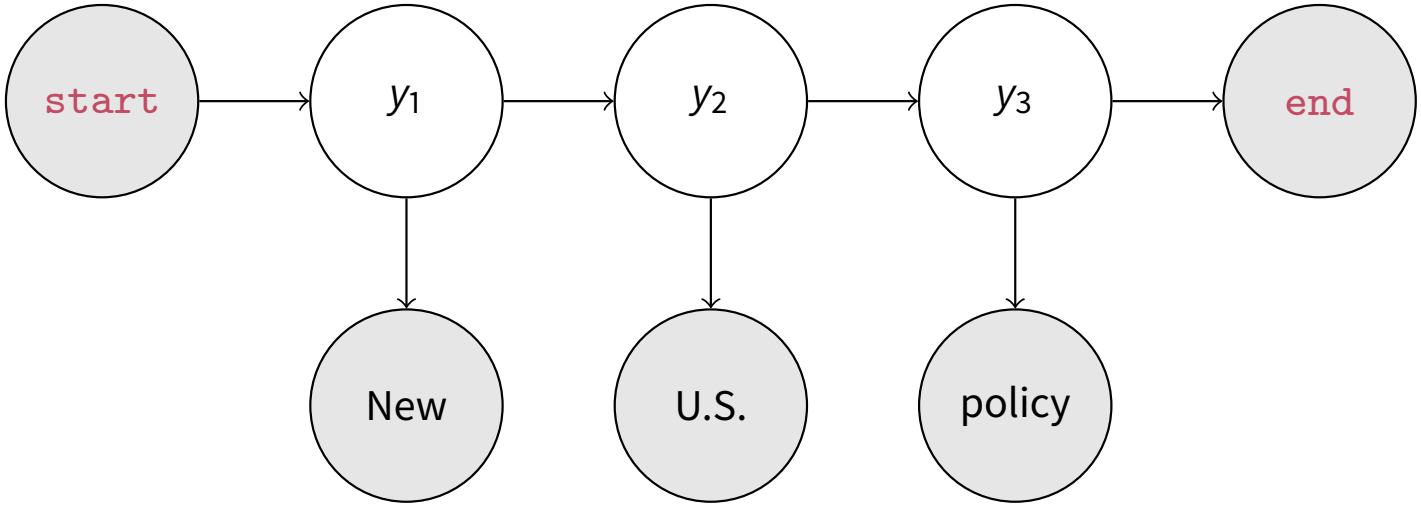


- $y_1 = \arg\max_{y \in \Sigma} \mathrm{P}(y|\texttt{start})\,\mathrm{P}(x_1|y)$
- $y_2 = \arg\max_{y \in \Sigma} \mathrm{P}(y|y_1)\,\mathrm{P}(x_2|y)$
- $y_3 = \arg\max_{y \in \Sigma} \mathrm{P}(y|y_2)\,\mathrm{P}(x_3|y)\,\mathrm{P}(\texttt{end}|y)$

  Is this algorithm correct? (Does it return $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$?)

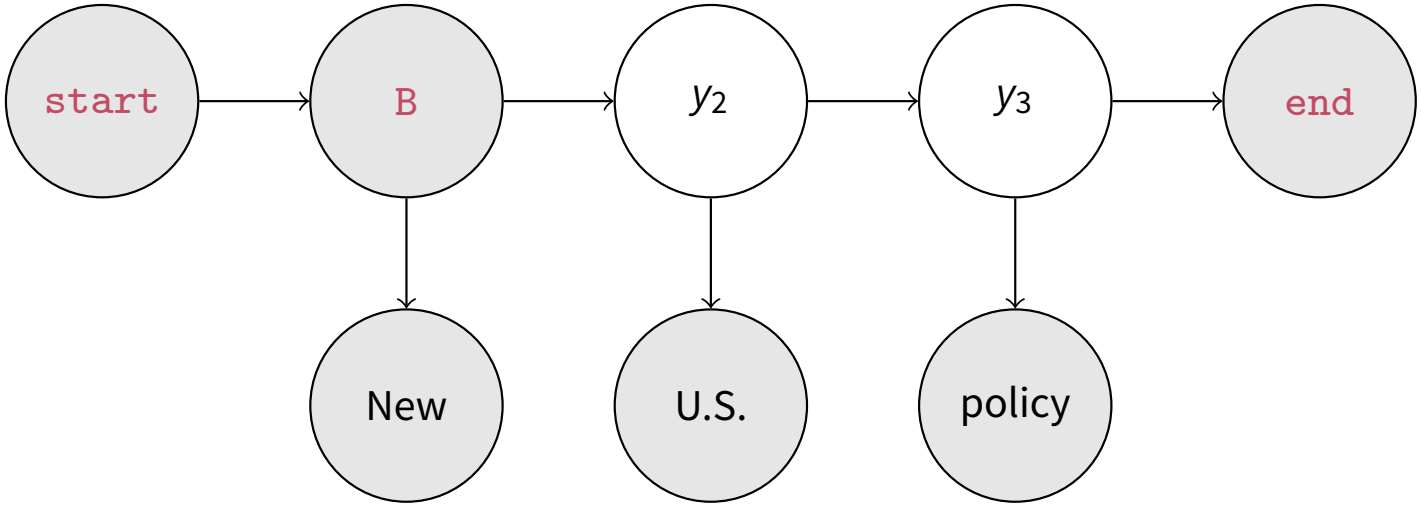
Big HMM questions: Finding the most likely y

Greedy prediction!



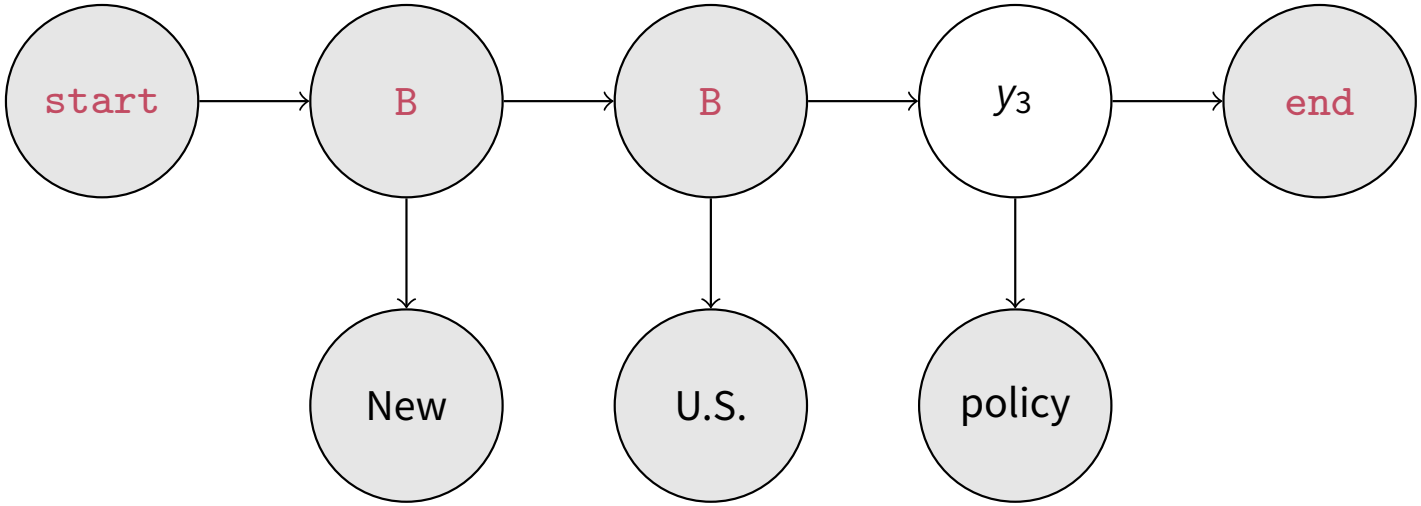
Big HMM questions: Finding the most likely y

Greedy prediction!



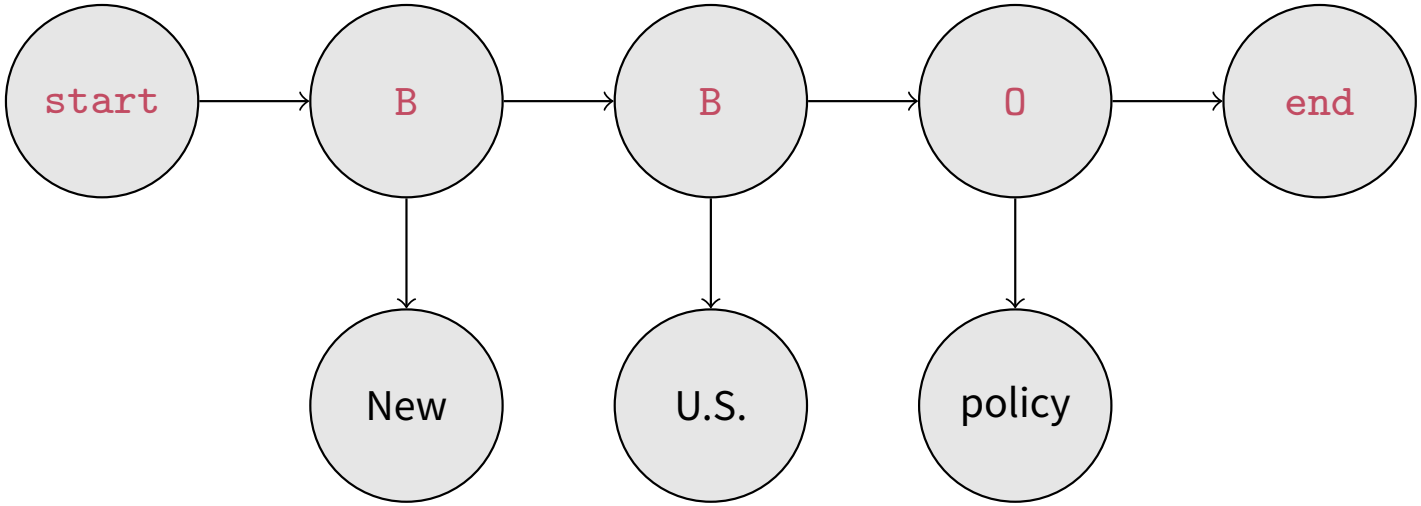
Big HMM questions: Finding the most likely y

Greedy prediction!



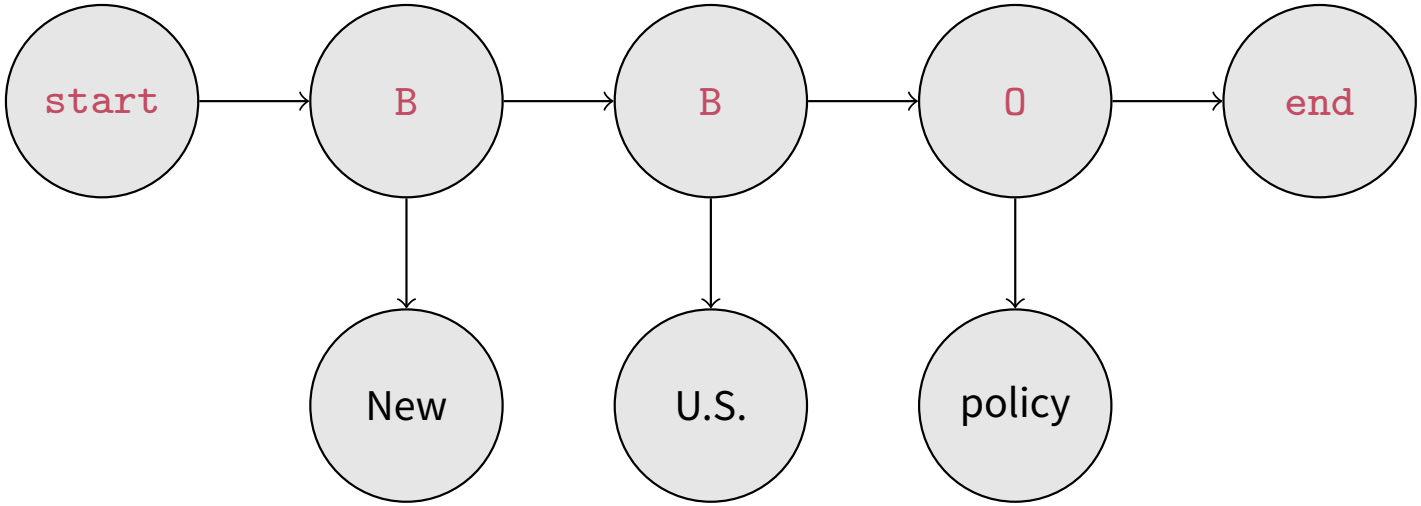
Big HMM questions: Finding the most likely y

Greedy prediction!



Big HMM questions: Finding the most likely y

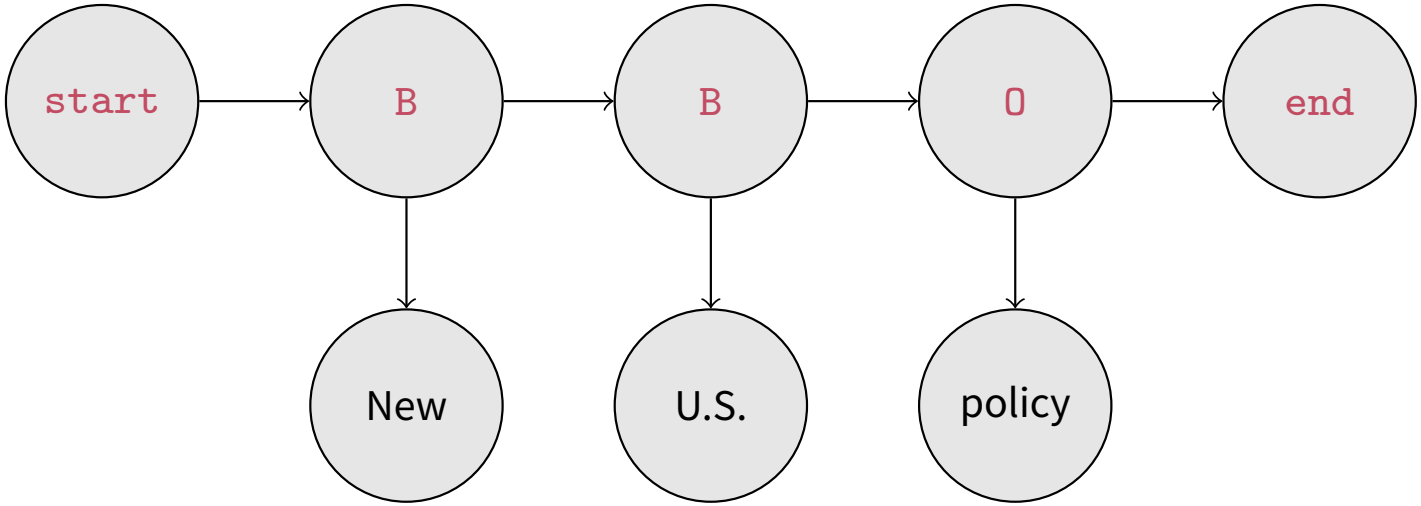
Greedy prediction!



- The true $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) = [\mathtt{O}, \mathtt{B}, \mathtt{O}]$
- We just got lucky the first time! What went wrong?

Big HMM questions: Finding the most likely y

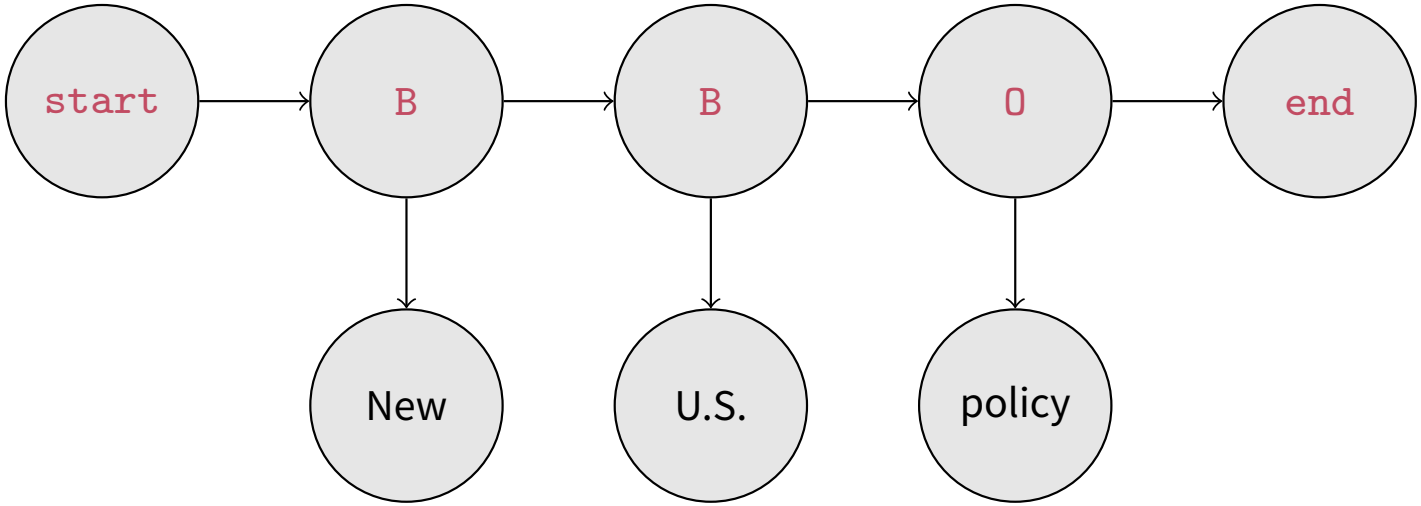
Greedy prediction!



- The true $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) = [\mathtt{O}, \mathtt{B}, \mathtt{O}]$
- We just got lucky the first time! What went wrong?
- We commit to a wrong label in the beginning, and can't go back.

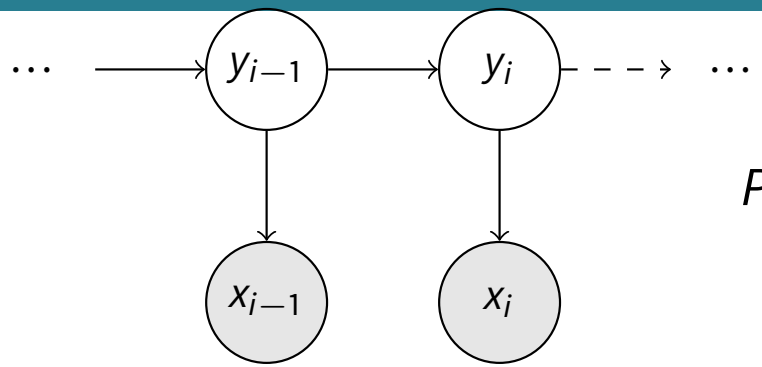
Big HMM questions: Finding the most likely y

Greedy prediction!



- The true $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) = [\mathtt{O}, \mathtt{B}, \mathtt{O}]$
- We just got lucky the first time! What went wrong?
- We commit to a wrong label in the beginning, and can't go back.
- We should keep some memory of the past, to be able to undo.

Big HMM questions: Finding the most likely y



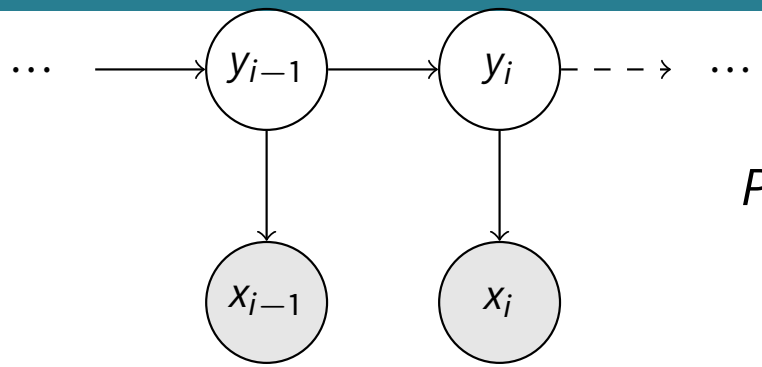
$$P(x_1, \ldots, x_i, y_1, \ldots, y_i) = \prod_{k=1}^{i} \mathrm{P}(x_k|y_k)\,\mathrm{P}(y_k|y_{k-1})$$

- At step i, let's assign a score to each possible choice for y_i. It will depend on the optimal labels $y_1, \ldots, y_{i-1}$.

$$\text{score}_i(y) = \max_{y_1,\ldots,y_{i-1}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, y_i = y)$$

But wait!

Big HMM questions: Finding the most likely y



$$P(x_1, \ldots, x_i, y_1, \ldots, y_i) = \prod_{k=1}^{i} \mathrm{P}(x_k|y_k)\,\mathrm{P}(y_k|y_{k-1})$$

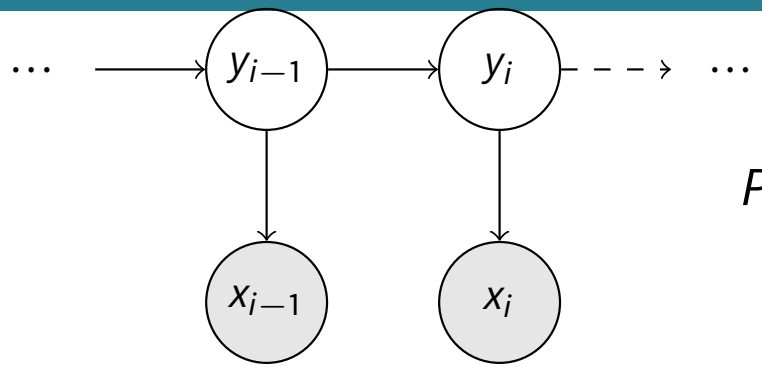
- At step i, let's assign a score to each possible choice for y_i. It will depend on the optimal labels $y_1, \ldots, y_{i-1}$.

$$\text{score}_i(y) = \max_{y_1,\ldots,y_{i-1}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, y_i = y)$$

But wait!

$$= \max_{y_1,\ldots,y_{i-1}} \Big(\mathrm{P}(x_i|y)\,\mathrm{P}(y|y_{i-1})\,\mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}) \Big)$$

Big HMM questions: Finding the most likely y



$$P(x_1, \ldots, x_i, y_1, \ldots, y_i) = \prod_{k=1}^{i} \mathrm{P}(x_k|y_k)\,\mathrm{P}(y_k|y_{k-1})$$

- At step i, let's assign a score to each possible choice for y_i. It will depend on the optimal labels $y_1, \ldots, y_{i-1}$.

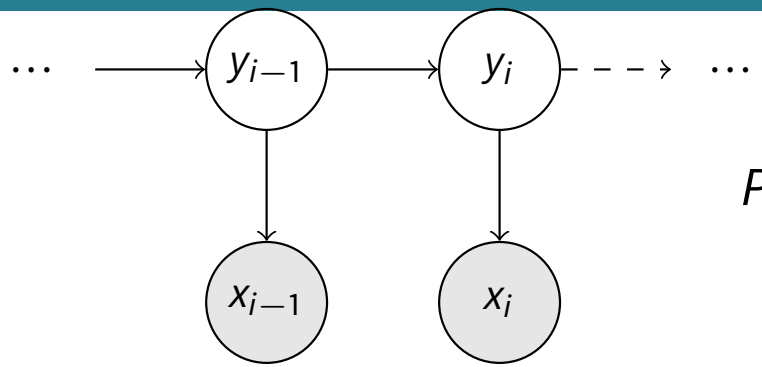
$$\text{score}_i(y) = \max_{y_1,\ldots,y_{i-1}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, y_i = y)$$

But wait!

$$= \max_{y_1,\ldots,y_{i-1}} \Big(\mathrm{P}(x_i|y)\,\mathrm{P}(y|y_{i-1})\,\mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}) \Big)$$

$$= \mathrm{P}(x_i|y) \max_{y'} \Big(\mathrm{P}(y|y') \max_{y_1,\ldots,y_{i-2}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1} = y') \Big)$$

Big HMM questions: Finding the most likely y



$$P(x_1, \ldots, x_i, y_1, \ldots, y_i) = \prod_{k=1}^{i} \mathrm{P}(x_k|y_k)\,\mathrm{P}(y_k|y_{k-1})$$

- At step i, let's assign a score to each possible choice for y_i. It will depend on the optimal labels $y_1, \ldots, y_{i-1}$.

$$\text{score}_i(y) = \max_{y_1,\ldots,y_{i-1}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, y_i = y)$$

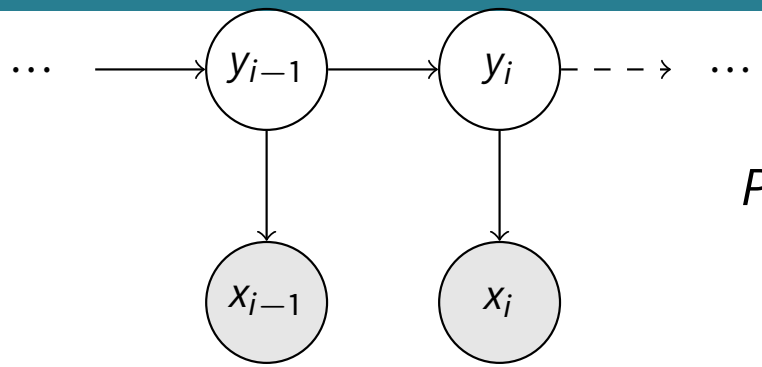
But wait!

$$= \max_{y_1,\ldots,y_{i-1}} \Big(\mathrm{P}(x_i|y)\,\mathrm{P}(y|y_{i-1})\,\mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}) \Big)$$

$$= \mathrm{P}(x_i|y) \max_{y'} \Big(\mathrm{P}(y|y') \max_{y_1,\ldots,y_{i-2}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1} = y') \Big)$$

$$= \mathrm{P}(x_i|y) \max_{y'} \Big(\mathrm{P}(y|y')\,\text{score}_{i-1}(y') \Big)$$

Big HMM questions: Finding the most likely y



$$P(x_1, \ldots, x_i, y_1, \ldots, y_i) = \prod_{k=1}^{i} \mathrm{P}(x_k|y_k)\,\mathrm{P}(y_k|y_{k-1})$$

- At step i, let's assign a score to each possible choice for y_i. It will depend on the optimal labels $y_1, \ldots, y_{i-1}$.

$$\text{score}_i(y) = \max_{y_1,\ldots,y_{i-1}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i, y_i = y)$$

But wait!

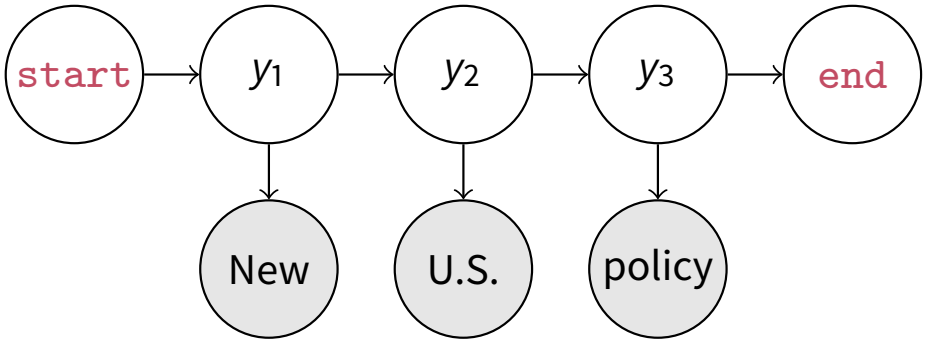
$$= \max_{y_1,\ldots,y_{i-1}} \Big(\mathrm{P}(x_i|y)\,\mathrm{P}(y|y_{i-1})\,\mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}) \Big)$$

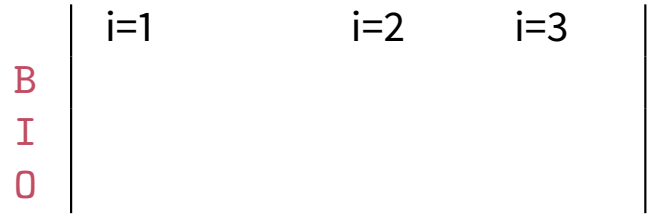
$$= \mathrm{P}(x_i|y) \max_{y'} \Big(\mathrm{P}(y|y') \max_{y_1,\ldots,y_{i-2}} \mathrm{P}(x_1, y_1, \ldots, x_{i-1}, y_{i-1} = y') \Big)$$

$$= \mathrm{P}(x_i|y) \max_{y'} \Big(\mathrm{P}(y|y')\,\text{score}_{i-1}(y') \Big)$$

- We can compute the scores recursively!

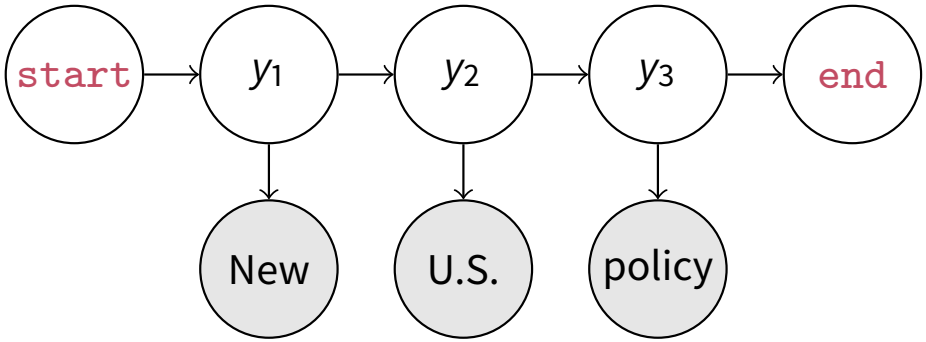
Big HMM questions: Finding the most likely y





$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	$\text{score}_1(\texttt{B})$		
I	$\text{score}_1(\texttt{I})$		
O	$\text{score}_1(\texttt{O})$		

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

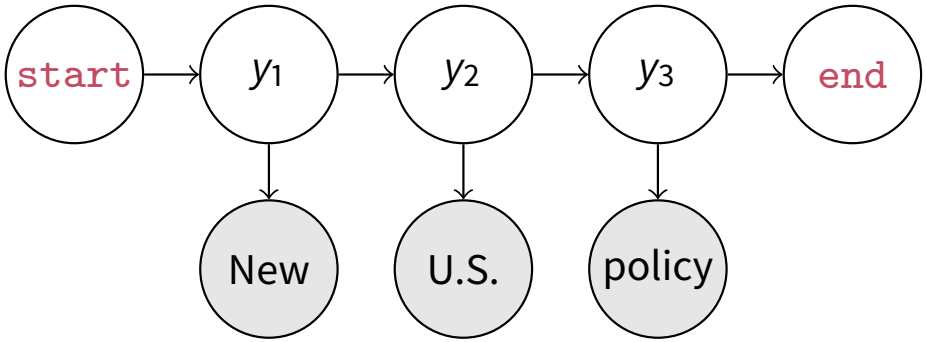
- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start})$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start})$

  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start})$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002		
I	.0		
O	.0003		

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

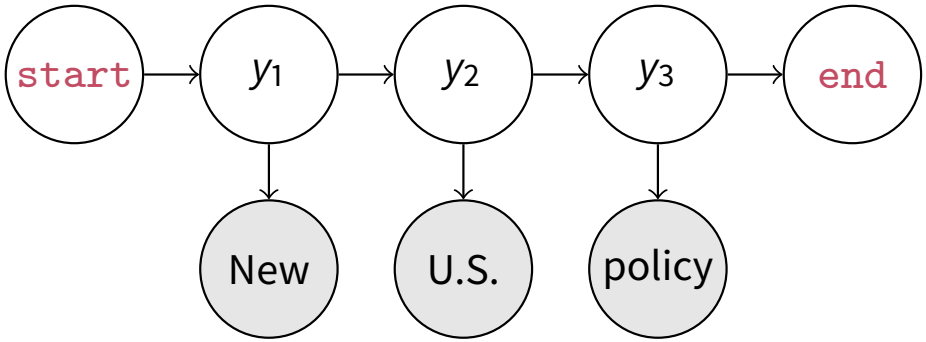
- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start}) = .006 \cdot .393$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start}) = .001 \cdot .0$

  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start}) = .0006 \cdot .607$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002		
I	.0		
O	.0003		

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start}) = .006 \cdot .393$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start}) = .001 \cdot .0$

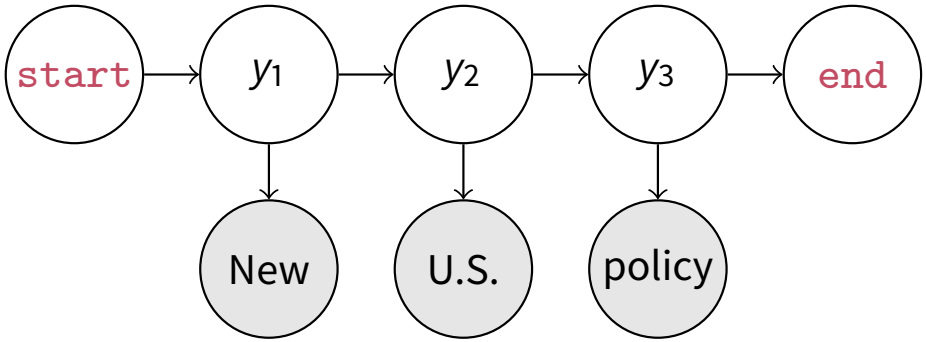
  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\text{score}_2(\texttt{B}) = \text{P}(\text{U.S.}|\texttt{B}) \max_{y'} \text{P}(\texttt{B}|y') \, \text{score}_1(y')$$

$$= \text{P}(\text{U.S.}|\texttt{B}) \max \begin{cases} \text{P}(\texttt{B}|\texttt{B}) \, \text{score}_1(\texttt{B}) = .009 \cdot .002 & = .00002 \\ \text{P}(\texttt{B}|\texttt{I}) \, \text{score}_1(\texttt{I}) = .003 \cdot .0 & = .0 \\ \text{P}(\texttt{B}|\texttt{O}) \, \text{score}_1(\texttt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00003$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	
I	.0		
O	.0003		

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start}) = .006 \cdot .393$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start}) = .001 \cdot .0$

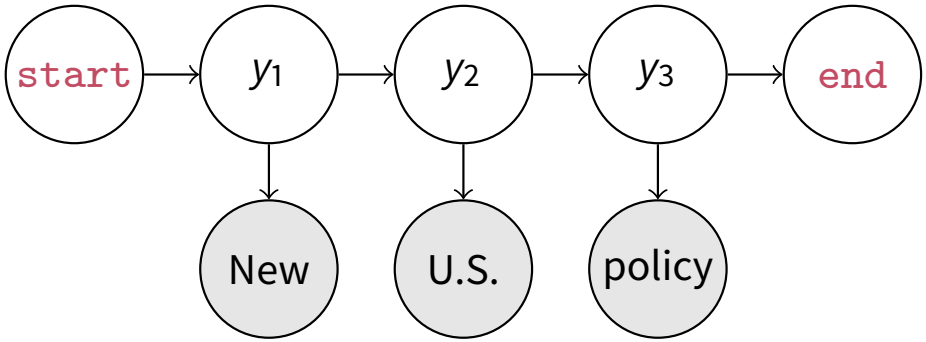
  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\text{score}_2(\texttt{B}) = \text{P}(\text{U.S.}|\texttt{B}) \max_{y'} \text{P}(\texttt{B}|y') \, \text{score}_1(y')$$

$$= \text{P}(\text{U.S.}|\texttt{B}) \max \begin{cases} \text{P}(\texttt{B}|\texttt{B}) \, \text{score}_1(\texttt{B}) = .009 \cdot .002 & = .00002 \\ \text{P}(\texttt{B}|\texttt{I}) \, \text{score}_1(\texttt{I}) = .003 \cdot .0 & = .0 \\ \text{P}(\texttt{B}|\texttt{O}) \, \text{score}_1(\texttt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00003$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	
I	.0	7e-11	
O	.0003	6e-12	

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start}) = .006 \cdot .393$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start}) = .001 \cdot .0$

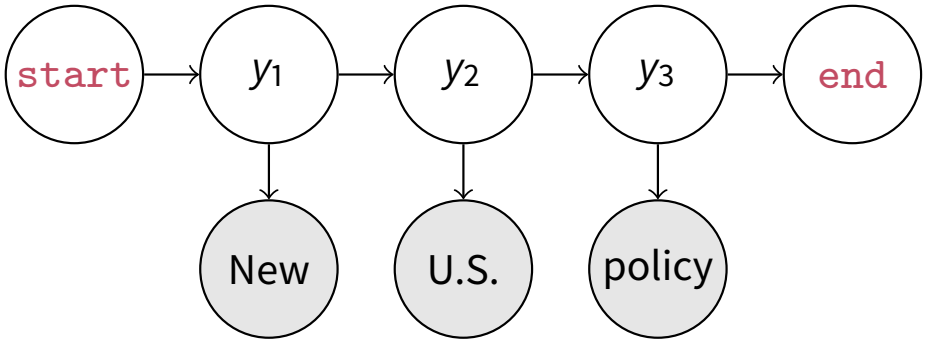
  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\text{score}_2(\texttt{B}) = \text{P}(\text{U.S.}|\texttt{B}) \max_{y'} \text{P}(\texttt{B}|y') \, \text{score}_1(y')$$

$$= \text{P}(\text{U.S.}|\texttt{B}) \max \begin{cases} \text{P}(\texttt{B}|\texttt{B}) \, \text{score}_1(\texttt{B}) = .009 \cdot .002 & = .00002 \\ \text{P}(\texttt{B}|\texttt{I}) \, \text{score}_1(\texttt{I}) = .003 \cdot .0 & = .0 \\ \text{P}(\texttt{B}|\texttt{O}) \, \text{score}_1(\texttt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00003$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	2e-16
I	.0	7e-11	2e-14
O	.0003	6e-12	5e-11

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At $i = 1$, the previous label can only be `start`:

  $\text{score}_1(\texttt{B}) = \text{P}(\text{New}|\texttt{B}) \, \text{P}(\texttt{B}|\texttt{start}) = .006 \cdot .393$

  $\text{score}_1(\texttt{I}) = \text{P}(\text{New}|\texttt{I}) \, \text{P}(\texttt{I}|\texttt{start}) = .001 \cdot .0$

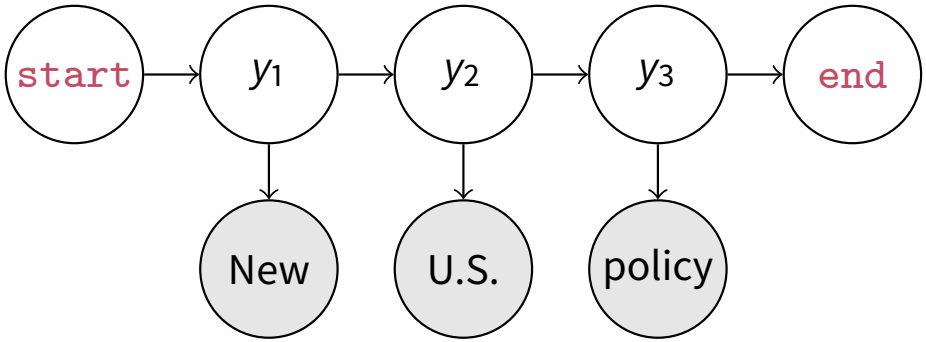
  $\text{score}_1(\texttt{O}) = \text{P}(\text{New}|\texttt{O}) \, \text{P}(\texttt{O}|\texttt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\text{score}_2(\texttt{B}) = \text{P}(\text{U.S.}|\texttt{B}) \max_{y'} \text{P}(\texttt{B}|y') \, \text{score}_1(y')$$

$$= \text{P}(\text{U.S.}|\texttt{B}) \max \begin{cases} \text{P}(\texttt{B}|\texttt{B}) \, \text{score}_1(\texttt{B}) = .009 \cdot .002 & = .00002 \\ \text{P}(\texttt{B}|\texttt{I}) \, \text{score}_1(\texttt{I}) = .003 \cdot .0 & = .0 \\ \text{P}(\texttt{B}|\texttt{O}) \, \text{score}_1(\texttt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00003$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	2e-16
I	.0	7e-11	2e-14
O	.0003	6e-12	5e-11

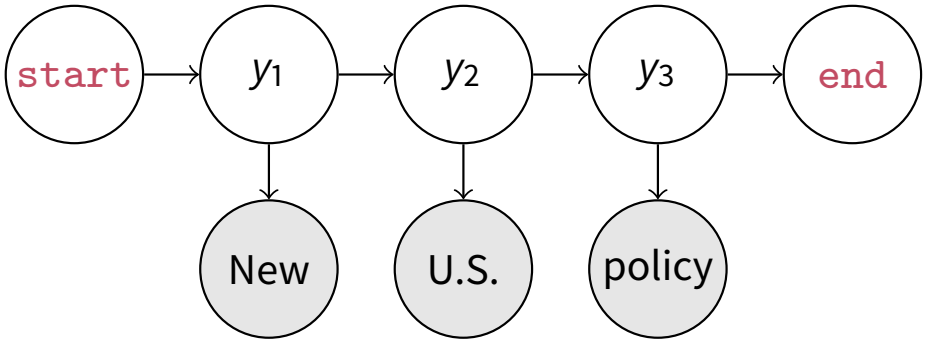
$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At the end we observe:

$$\max_y P(\texttt{end}|y) \, \text{score}_L(y) = \max_y P(\texttt{end}|y) \max_{y_1,\ldots,y_{L-1}} P(x_1, y_1, \ldots, x_L, y_L = y)$$

$$= \max_{\boldsymbol{y}} \text{P}(\boldsymbol{x}, \boldsymbol{y})$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	2e-16
I	.0	7e-11	2e-14
O	.0003	6e-12	5e-11

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At the end we observe:

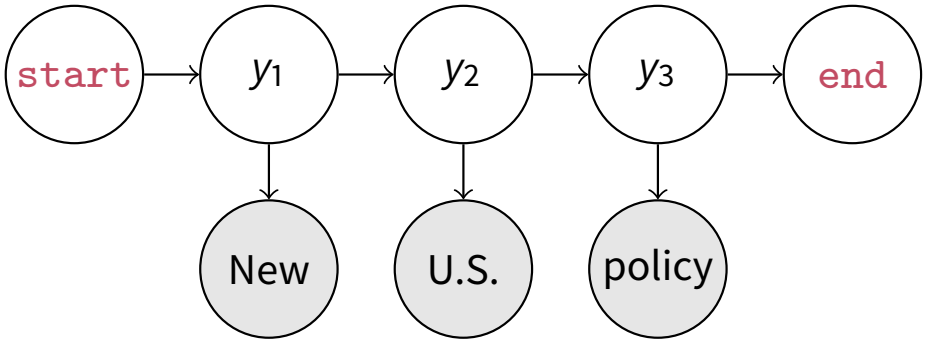
$$\max_y P(\texttt{end}|y) \, \text{score}_L(y) = \max_y P(\texttt{end}|y) \max_{y_1,\ldots,y_{L-1}} P(x_1, y_1, \ldots, x_L, y_L = y)$$

$$= \max_{\boldsymbol{y}} \text{P}(\boldsymbol{x}, \boldsymbol{y})$$

- In this case,

$$\max \begin{cases} \text{P}(\texttt{end}|\texttt{B}) \, \text{score}_3(\texttt{B}) = .013 \cdot \text{2e-16} \\ \text{P}(\texttt{end}|\texttt{I}) \, \text{score}_3(\texttt{I}) = .04 \cdot \text{2e-14} \\ \text{P}(\texttt{end}|\texttt{O}) \, \text{score}_3(\texttt{O}) = .084 \cdot \text{5e-11} \end{cases} = \text{4e-12}$$

Big HMM questions: Finding the most likely y



	i=1	i=2	i=3
B	.002	5e-7	2e-16
I	.0	7e-11	2e-14
O	.0003	6e-12	5e-11

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At the end we observe:

$$\max_y P(\texttt{end}|y) \, \text{score}_L(y) = \max_y P(\texttt{end}|y) \max_{y_1,\ldots,y_{L-1}} P(x_1, y_1, \ldots, x_L, y_L = y)$$

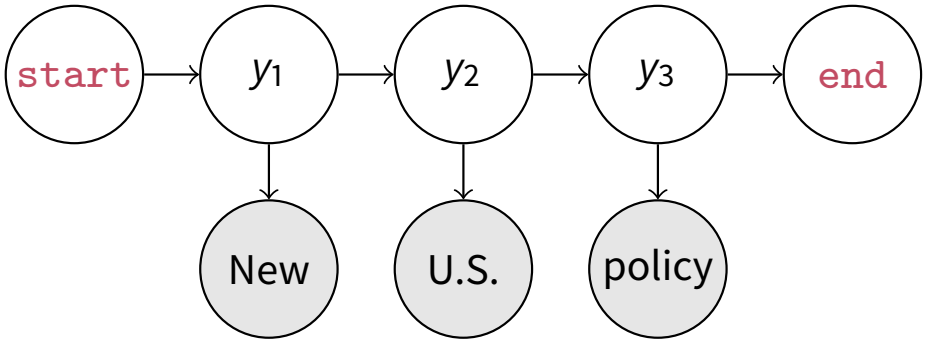
$$= \max_{\boldsymbol{y}} \text{P}(\boldsymbol{x}, \boldsymbol{y})$$

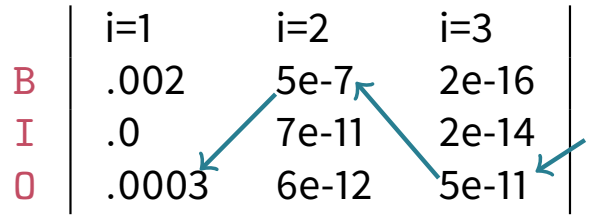
- In this case,

$$\max \begin{cases} \text{P}(\texttt{end}|\texttt{B}) \, \text{score}_3(\texttt{B}) = .013 \cdot \text{2e-16} \\ \text{P}(\texttt{end}|\texttt{I}) \, \text{score}_3(\texttt{I}) = .04 \cdot \text{2e-14} \\ \text{P}(\texttt{end}|\texttt{O}) \, \text{score}_3(\texttt{O}) = .084 \cdot \text{5e-11} \end{cases} = \text{4e-12}$$

- What is the $\boldsymbol{y}$ that gives this value?

Big HMM questions: Finding the most likely y





$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_i(y')$$

- At the end we observe:

$$\max_y P(\texttt{end}|y) \, \text{score}_L(y) = \max_y P(\texttt{end}|y) \max_{y_1,\ldots,y_{L-1}} P(x_1, y_1, \ldots, x_L, y_L = y)$$

$$= \max_{\boldsymbol{y}} \text{P}(\boldsymbol{x}, \boldsymbol{y})$$

- In this case,

$$\max \begin{cases} \text{P}(\texttt{end}|\texttt{B}) \, \text{score}_3(\texttt{B}) = .013 \cdot \text{2e-16} \\ \text{P}(\texttt{end}|\texttt{I}) \, \text{score}_3(\texttt{I}) = .04 \cdot \text{2e-14} \\ \text{P}(\texttt{end}|\texttt{O}) \, \text{score}_3(\texttt{O}) = .084 \cdot \text{5e-11} \end{cases} = \text{4e-12}$$

- What is the y that gives this value? **Backtrace** remembering each max!

The Viterbi algorithm

input: sequence $x_1, \ldots, x_L$, emission and transition probabilities

Forward: compute scores recursively

$\text{score}_1(y) = \text{P}(y|\texttt{start}) \cdot \text{P}(x_1|y) \quad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

  for $y \in \Sigma$ **do**

$$\text{score}_i(y) = \text{P}(x_i|y) \cdot \max_{y'} \Big(\text{P}(y|y') \cdot \text{score}_{i-1}(y') \Big)$$

$$\text{backptr}_i(y) = \quad \arg\max_{y'} \Big(\text{P}(y|y') \cdot \text{score}_{i-1}(y') \Big)$$

Backward: follow backpointers

$$p = \max_{y'} \Big(\text{P}(\texttt{end}|y') \cdot \text{score}_L(y') \Big)$$

$$\widehat{y}_L = \arg\max_{y'} \Big(\text{P}(\texttt{end}|y') \cdot \text{score}_L(y') \Big)$$

for $i = L - 1$ **down to** 1 **do**

  $\widehat{y}_i = \text{backptr}_{i+1}(\widehat{y}_{i+1})$

output: the most likely sequence $\widehat{\boldsymbol{y}} = [\widehat{y}_1, \ldots, \widehat{y}_L]$
and the joint likelihood $p = \text{P}(\boldsymbol{x}, \widehat{\boldsymbol{y}})$

Viterbi with log-probabilities

- Notice how probabilities get tiny (1e-14) even for short sequences.
- Multiplying small numbers is not numerically robust. Fortunately,

$$u < v \iff \log u < \log v \quad \text{and} \quad \log ab = \log a + \log b$$

- We can equivalently find $\hat{\boldsymbol{y}}$ as:

$$\begin{aligned}
\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) &= \arg\max_{\boldsymbol{y}} \log \mathrm{P}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{y}} \log \left(\prod_{i=1}^{L+1} \mathrm{P}(y_i|y_{i-1}) \cdot \prod_{i=1}^{L} \mathrm{P}(x_i|y_i) \right) \\
&= \arg\max_{\boldsymbol{y}} \left(\sum_{i=1}^{L+1} \log \mathrm{P}(y_i|y_{i-1}) + \sum_{i=1}^{L} \log \mathrm{P}(x_i|y_i) \right)
\end{aligned}$$

- Mathematically equivalent, but what good is mathematics when computers can't represent your numbers?

Viterbi with log-probabilities (USE THIS ONE)

input: sequence $x_1, \ldots, x_L$, emission and transition **log-**probabilities

Forward: compute scores recursively

$\widetilde{\text{score}}_1(y) = \log \text{P}(y|\texttt{start}) + \log \text{P}(x_1|y) \quad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

  for $y \in \Sigma$ **do**

$$\widetilde{\text{score}}_i(y) = \log \text{P}(x_i|y) + \max_{y'} \left(\log \text{P}(y|y') + \widetilde{\text{score}}_{i-1}(y') \right)$$

$$\text{backptr}_i(y) = \arg\max_{y'} \left(\log \text{P}(y|y') + \widetilde{\text{score}}_{i-1}(y') \right)$$

Backward: follow backpointers

$$l = \max_{y'} \left(\log \text{P}(\texttt{end}|y') + \widetilde{\text{score}}_L(y') \right)$$

$$\widehat{y}_L = \arg\max_{y'} \left(\log \text{P}(\texttt{end}|y') + \widetilde{\text{score}}_L(y') \right)$$

for $i = L - 1$ **down to** 1 **do**

  $\widehat{y}_i = \text{backptr}_{i+1}(\widehat{y}_{i+1})$

output: the most likely sequence $\widehat{\boldsymbol{y}} = [\widehat{y}_1, \ldots, \widehat{y}_L]$
and the joint log-likelihood $l = \log \text{P}(\boldsymbol{x}, \widehat{\boldsymbol{y}})$

note: $\widetilde{\text{score}}_i(y) = \log \text{score}_i(y)$

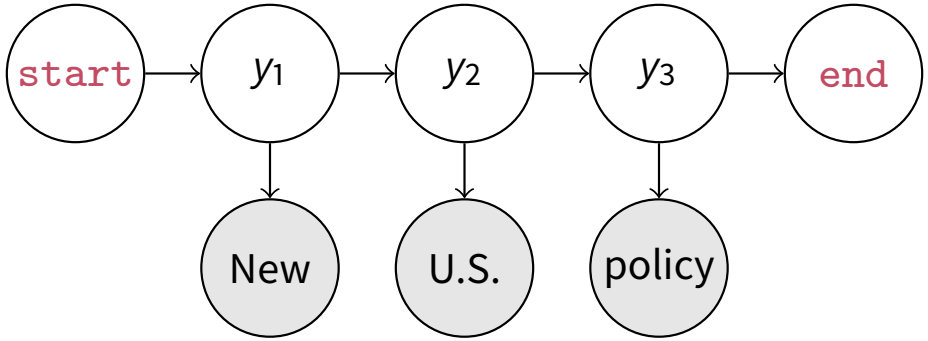
Summing Up: Viterbi

- Computes the most likely sequence of tags: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$
- This is called MAP (*maximum a posteriori*) decoding
- An instance of a dynamic programming algorithm: makes use of recurrence to reuse *partial solutions*.

Big HMM questions

1. What is the most likely label sequence $\boldsymbol{y}$, given $\boldsymbol{x}$?
2. What is the probability of $\boldsymbol{x}$?
3. What is the probability of each assignment y_i, given $\boldsymbol{x}$?
4. What is the probability of each transition $y \rightarrow y'$, given $\boldsymbol{x}$?
5. What sequence $\boldsymbol{y}$ minimizes the *Hamming cost?*

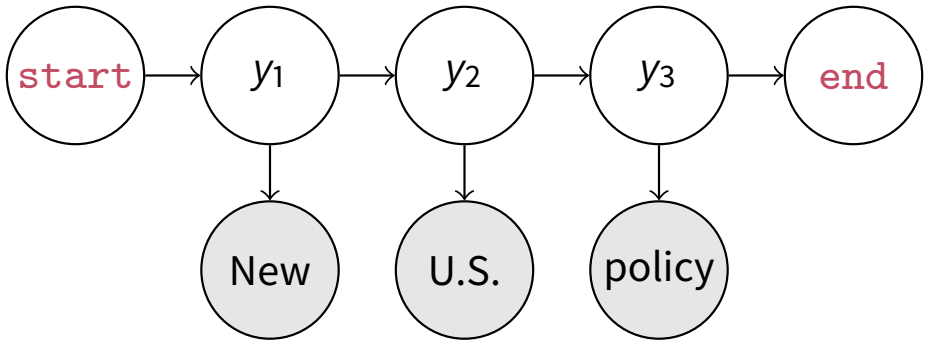
Big HMM questions: Marginal probability of x



- Treat $\boldsymbol{y}$ as unknown (missing). Marginal probability of $\boldsymbol{x}$:

$$\mathsf{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathsf{P}(\boldsymbol{x}, \boldsymbol{y})$$

Big HMM questions: Marginal probability of x

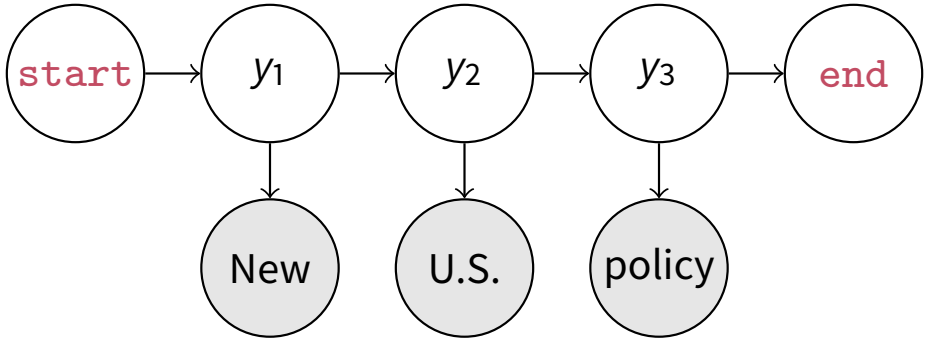


- Treat $\boldsymbol{y}$ as unknown (missing). Marginal probability of $\boldsymbol{x}$:

$$\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$$

- Why?
  - Compare likelihood of observations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$,
    e.g. language model, outlier detection...
  - Maximize this to learn HMM without supervision.
  - Assess $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{P}(\boldsymbol{x})}$.

Big HMM questions: Marginal probability of x



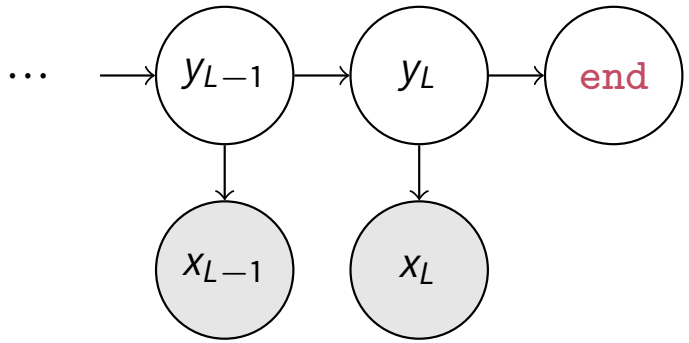
- Treat $\boldsymbol{y}$ as unknown (missing). Marginal probability of $\boldsymbol{x}$:

$$\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$$

- Why?
  - Compare likelihood of observations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$, e.g. language model, outlier detection...
  - Maximize this to learn HMM without supervision.
  - Assess $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{P}(\boldsymbol{x})}$.
- One algo: enumeration (correct, but prohibitive):

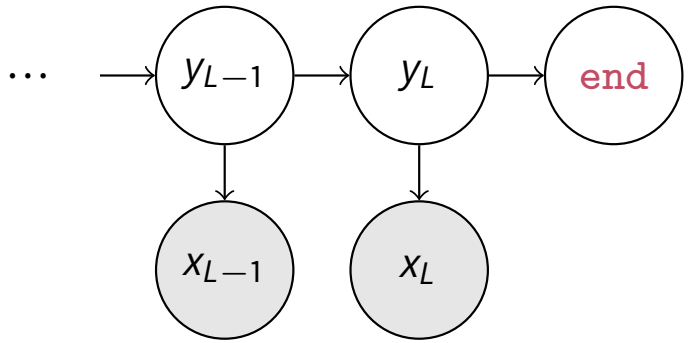
  $p \leftarrow 0$
  for $\boldsymbol{y} \in \Sigma^L$
  $\quad p \leftarrow p + P(\boldsymbol{x}, \boldsymbol{y})$
  Return p

Big HMM questions: Marginal probability of x



- Treat $\boldsymbol{y}$ as unknown. Marginal probability of $\boldsymbol{x}$: $\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Remember "what's the probability that the 3rd label is B?"

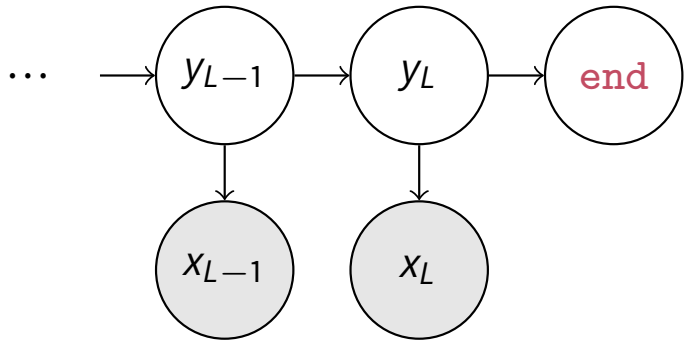
Big HMM questions: Marginal probability of x



- Treat $\boldsymbol{y}$ as unknown. Marginal probability of $\boldsymbol{x}$: $\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Remember "what's the probability that the 3rd label is B?"

$$\mathrm{P}(\boldsymbol{x}) = \sum_{y} \mathrm{P}(\texttt{end}|y) \underbrace{\sum_{y_1,\ldots,y_{L-1}} \mathrm{P}(x_1,\ldots,x_L,y_1,\ldots,y_{L-1},y_L = y)}_{\alpha_L(y)}$$

Big HMM questions: Marginal probability of x

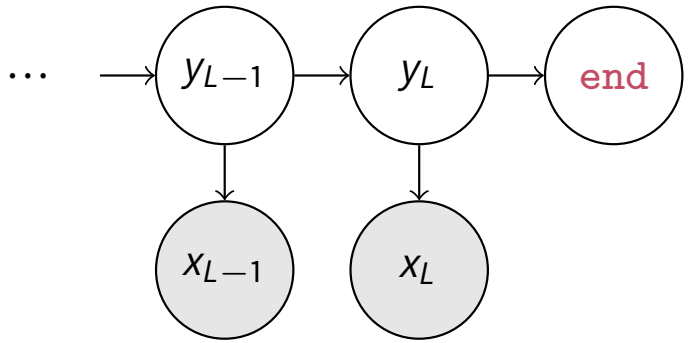


- Treat $\boldsymbol{y}$ as unknown. Marginal probability of $\boldsymbol{x}$: $\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Remember "what's the probability that the 3rd label is B?"

$$\mathrm{P}(\boldsymbol{x}) = \sum_{y} \mathrm{P}(\texttt{end}|y) \underbrace{\sum_{y_1,\ldots,y_{L-1}} \mathrm{P}(x_1,\ldots,x_L,y_1,\ldots,y_{L-1},y_L = y)}_{\alpha_L(y)}$$

$$= \sum_{y} \mathrm{P}(\texttt{end}|y)\, \mathrm{P}(x_L|y) \sum_{y'} P(y|y') \underbrace{\sum_{y_1,\ldots,y_{L-2}} \mathrm{P}(x_1,\ldots,x_{L-1},y_1,\ldots,y_{L-1} = y')}_{\alpha_{L-1}(y')}$$

Big HMM questions: Marginal probability of x



- Treat $\boldsymbol{y}$ as unknown. Marginal probability of $\boldsymbol{x}$: $\mathrm{P}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Remember "what's the probability that the 3rd label is B?"

$$\mathrm{P}(\boldsymbol{x}) = \sum_{y} \mathrm{P}(\texttt{end}|y) \underbrace{\sum_{y_1,\ldots,y_{L-1}} \mathrm{P}(x_1,\ldots,x_L,y_1,\ldots,y_{L-1},y_L = y)}_{\alpha_L(y)}$$

$$= \sum_{y} \mathrm{P}(\texttt{end}|y)\, \mathrm{P}(x_L|y) \sum_{y'} P(y|y') \underbrace{\sum_{y_1,\ldots,y_{L-2}} \mathrm{P}(x_1,\ldots,x_{L-1},y_1,\ldots,y_{L-1} = y')}_{\alpha_{L-1}(y')}$$

- Recurrence: $\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y'} \mathrm{P}(y|y')\, \alpha_{i-1}(y')$

Notice a similarity?

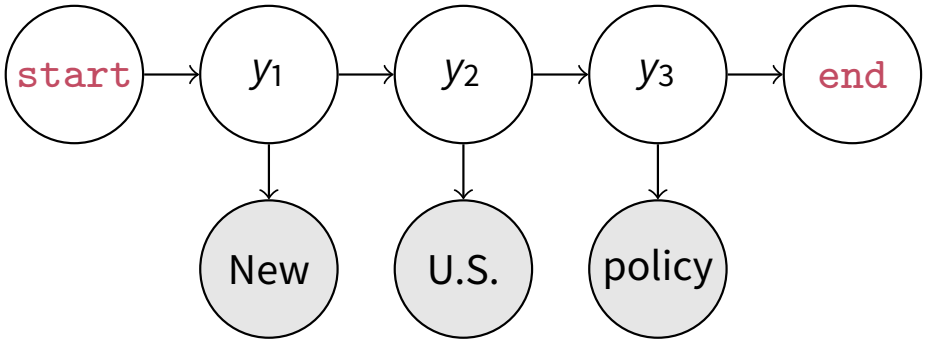
Viterbi recurrence:

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_{i-1}(y')$$

Marginalization recurrence:

$$\alpha_i(y) = \text{P}(x_i|y) \sum_{y'} \text{P}(y|y') \, \alpha_{i-1}(y')$$

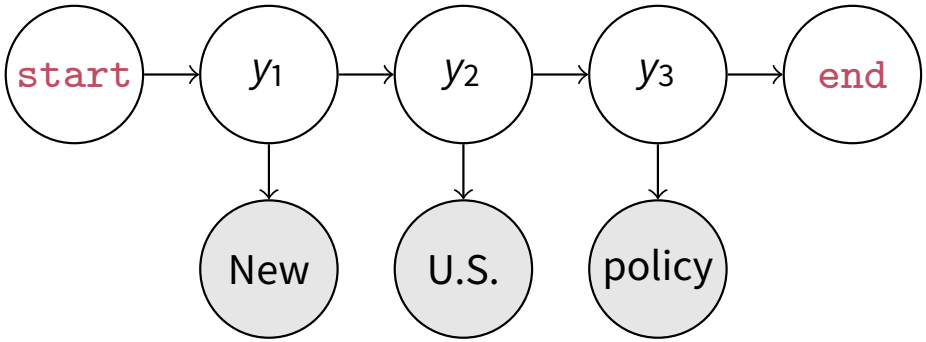
Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B			
I			
O			

$$\alpha_i(y) = \mathsf{P}(x_i|y) \sum_{y_{i'}} \mathsf{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	$\alpha_1(\texttt{B})$		
I	$\alpha_1(\texttt{I})$		
O	$\alpha_1(\texttt{O})$		

$$\alpha_i(y) = \mathsf{P}(x_i|y) \sum_{y_{i'}} \mathsf{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

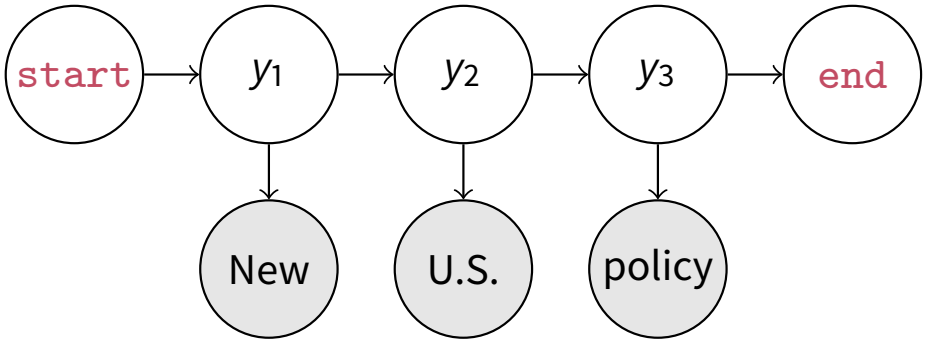
- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\texttt{B}) = \mathsf{P}(\text{New}|\texttt{B})\, \mathsf{P}(\texttt{B}|\texttt{start})$

  $\alpha_1(\texttt{I}) = \mathsf{P}(\text{New}|\texttt{I})\, \mathsf{P}(\texttt{I}|\texttt{start})$

  $\alpha_1(\texttt{O}) = \mathsf{P}(\text{New}|\texttt{O})\, \mathsf{P}(\texttt{O}|\texttt{start})$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002		
I	.0		
O	.0003		

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

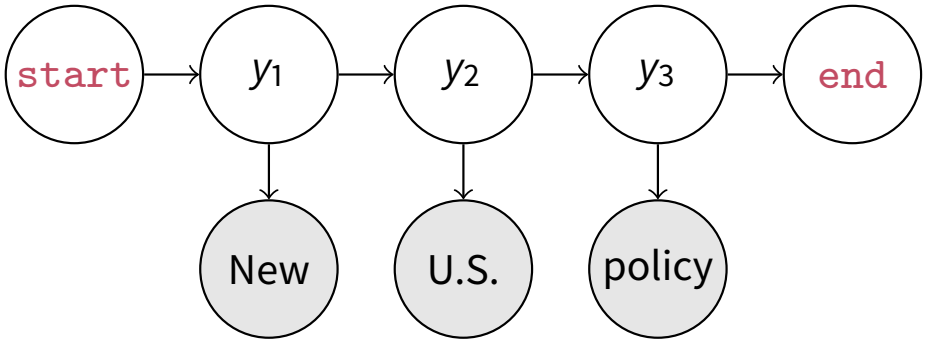
- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\mathtt{B}) = \mathrm{P}(\text{New}|\mathtt{B})\, \mathrm{P}(\mathtt{B}|\mathtt{start}) = .006 \cdot .393$

  $\alpha_1(\mathtt{I}) = \mathrm{P}(\text{New}|\mathtt{I})\, \mathrm{P}(\mathtt{I}|\mathtt{start}) = .001 \cdot .0$

  $\alpha_1(\mathtt{O}) = \mathrm{P}(\text{New}|\mathtt{O})\, \mathrm{P}(\mathtt{O}|\mathtt{start}) = .0006 \cdot .607$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002		
I	.0		
O	.0003		

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\mathtt{B}) = \mathrm{P}(\text{New}|\mathtt{B})\, \mathrm{P}(\mathtt{B}|\mathtt{start}) = .006 \cdot .393$

  $\alpha_1(\mathtt{I}) = \mathrm{P}(\text{New}|\mathtt{I})\, \mathrm{P}(\mathtt{I}|\mathtt{start}) = .001 \cdot .0$

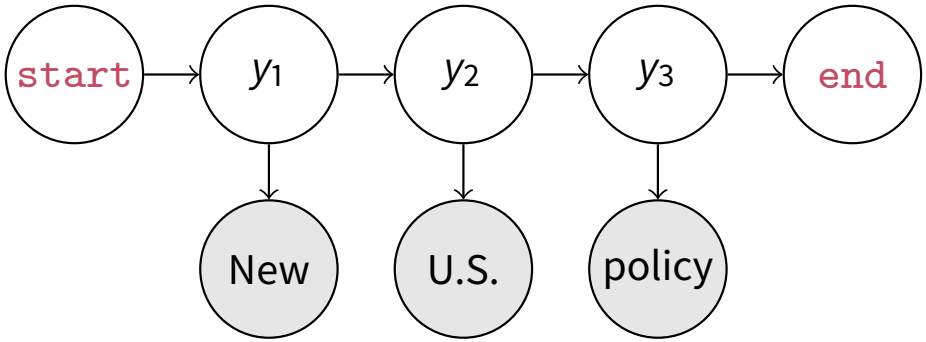
  $\alpha_1(\mathtt{O}) = \mathrm{P}(\text{New}|\mathtt{O})\, \mathrm{P}(\mathtt{O}|\mathtt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\alpha_2(\mathtt{B}) = \mathrm{P}(\text{U.S.}|\mathtt{B}) \sum_{y'} \mathrm{P}(\mathtt{B}|y')\, \alpha_1(y')$$

$$= \mathrm{P}(\text{U.S.}|\mathtt{B})\text{sum} \begin{cases} \mathrm{P}(\mathtt{B}|\mathtt{B})\, \alpha_1(\mathtt{B}) = .009 \cdot .002 & = .00002 \\ \mathrm{P}(\mathtt{B}|\mathtt{I})\, \alpha_1(\mathtt{I}) = .003 \cdot .0 & = .0 \\ \mathrm{P}(\mathtt{B}|\mathtt{O})\, \alpha_1(\mathtt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00005$$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002	8e-6	
I	.0		
O	.0003		

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\mathtt{B}) = \mathrm{P}(\text{New}|\mathtt{B})\, \mathrm{P}(\mathtt{B}|\mathtt{start}) = .006 \cdot .393$

  $\alpha_1(\mathtt{I}) = \mathrm{P}(\text{New}|\mathtt{I})\, \mathrm{P}(\mathtt{I}|\mathtt{start}) = .001 \cdot .0$

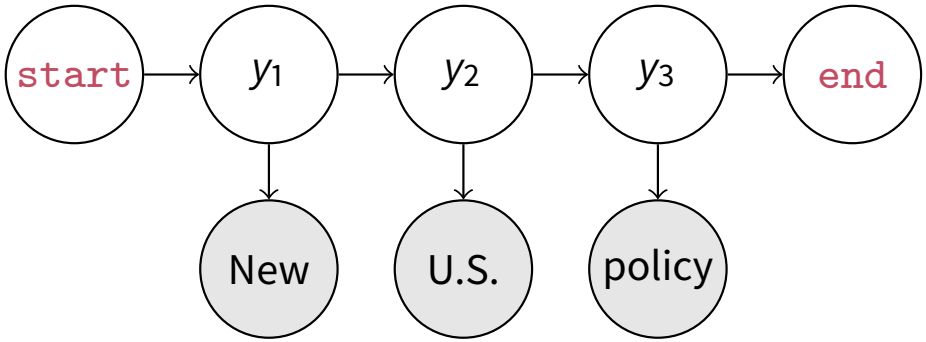
  $\alpha_1(\mathtt{O}) = \mathrm{P}(\text{New}|\mathtt{O})\, \mathrm{P}(\mathtt{O}|\mathtt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\alpha_2(\mathtt{B}) = \mathrm{P}(\text{U.S.}|\mathtt{B}) \sum_{y'} \mathrm{P}(\mathtt{B}|y')\, \alpha_1(y')$$

$$= \mathrm{P}(\text{U.S.}|\mathtt{B})\text{sum} \begin{cases} \mathrm{P}(\mathtt{B}|\mathtt{B})\, \alpha_1(\mathtt{B}) = .009 \cdot .002 & = .00002 \\ \mathrm{P}(\mathtt{B}|\mathtt{I})\, \alpha_1(\mathtt{I}) = .003 \cdot .0 & = .0 \\ \mathrm{P}(\mathtt{B}|\mathtt{O})\, \alpha_1(\mathtt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00005$$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002	8e-6	
I	.0	7e-11	
O	.0003	7e-12	

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\mathtt{B}) = \mathrm{P}(\text{New}|\mathtt{B})\, \mathrm{P}(\mathtt{B}|\mathtt{start}) = .006 \cdot .393$

  $\alpha_1(\mathtt{I}) = \mathrm{P}(\text{New}|\mathtt{I})\, \mathrm{P}(\mathtt{I}|\mathtt{start}) = .001 \cdot .0$

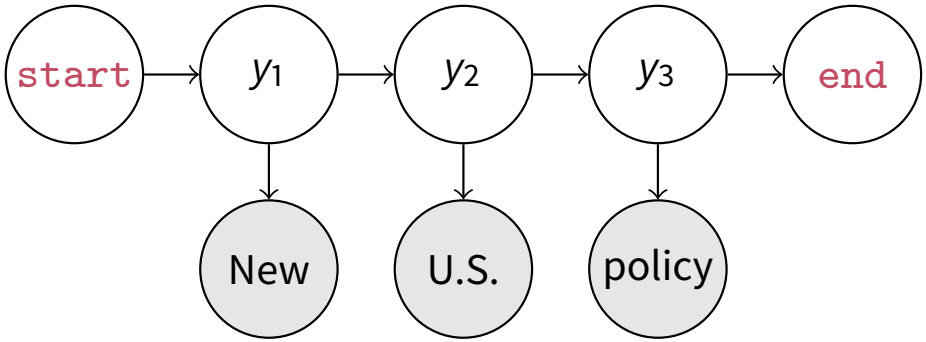
  $\alpha_1(\mathtt{O}) = \mathrm{P}(\text{New}|\mathtt{O})\, \mathrm{P}(\mathtt{O}|\mathtt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\alpha_2(\mathtt{B}) = \mathrm{P}(\text{U.S.}|\mathtt{B}) \sum_{y'} \mathrm{P}(\mathtt{B}|y')\, \alpha_1(y')$$

$$= \mathrm{P}(\text{U.S.}|\mathtt{B})\text{sum} \begin{cases} \mathrm{P}(\mathtt{B}|\mathtt{B})\, \alpha_1(\mathtt{B}) = .009 \cdot .002 & = .00002 \\ \mathrm{P}(\mathtt{B}|\mathtt{I})\, \alpha_1(\mathtt{I}) = .003 \cdot .0 & = .0 \\ \mathrm{P}(\mathtt{B}|\mathtt{O})\, \alpha_1(\mathtt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00005$$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002	8e-6	2e-16
I	.0	7e-11	2e-14
O	.0003	7e-12	5e-11

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At $i = 1$, the previous label can only be `start`:

  $\alpha_1(\mathtt{B}) = \mathrm{P}(\text{New}|\mathtt{B})\, \mathrm{P}(\mathtt{B}|\mathtt{start}) = .006 \cdot .393$

  $\alpha_1(\mathtt{I}) = \mathrm{P}(\text{New}|\mathtt{I})\, \mathrm{P}(\mathtt{I}|\mathtt{start}) = .001 \cdot .0$

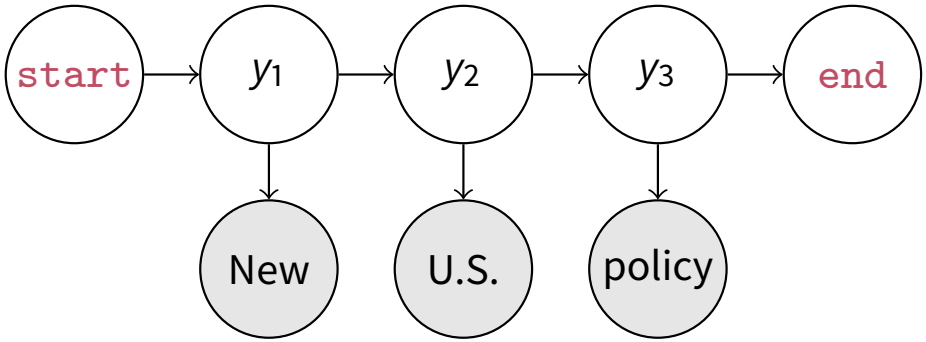
  $\alpha_1(\mathtt{O}) = \mathrm{P}(\text{New}|\mathtt{O})\, \mathrm{P}(\mathtt{O}|\mathtt{start}) = .0006 \cdot .607$

- At $i = 2$:

$$\alpha_2(\mathtt{B}) = \mathrm{P}(\text{U.S.}|\mathtt{B}) \sum_{y'} \mathrm{P}(\mathtt{B}|y')\, \alpha_1(y')$$

$$= \mathrm{P}(\text{U.S.}|\mathtt{B})\text{sum} \begin{cases} \mathrm{P}(\mathtt{B}|\mathtt{B})\, \alpha_1(\mathtt{B}) = .009 \cdot .002 & = .00002 \\ \mathrm{P}(\mathtt{B}|\mathtt{I})\, \alpha_1(\mathtt{I}) = .003 \cdot .0 & = .0 \\ \mathrm{P}(\mathtt{B}|\mathtt{O})\, \alpha_1(\mathtt{O}) = .102 \cdot .0003 & = .00003 \end{cases} = .016 \cdot .00005$$

Big HMM questions: Marginal probability of x



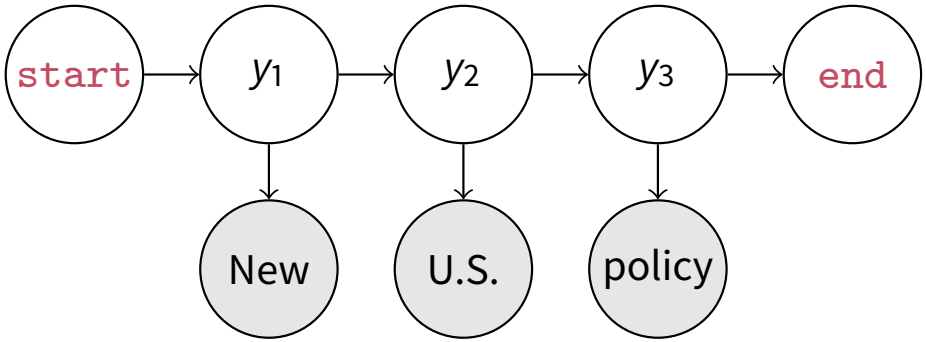
	i=1	i=2	i=3
B	.002	8e-6	2e-16
I	.0	7e-11	2e-14
O	.0003	7e-12	5e-11

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At the end we can consider:

$$P(\boldsymbol{x}) = \sum_{y} P(\texttt{end}|y)\, \alpha_L(y)$$

Big HMM questions: Marginal probability of x



	i=1	i=2	i=3
B	.002	8e-6	2e-16
I	.0	7e-11	2e-14
O	.0003	7e-12	5e-11

$$\alpha_i(y) = \mathrm{P}(x_i|y) \sum_{y_{i'}} \mathrm{P}(y|y_{i'})\, \alpha_i(y_{i'})$$

- At the end we can consider:

$$P(\boldsymbol{x}) = \sum_{y} P(\texttt{end}|y)\, \alpha_L(y)$$

- In this case,

$$\text{sum} \begin{cases} \mathrm{P}(\texttt{end}|\texttt{B})\, \alpha_3(\texttt{B}) = .013 \cdot 2\text{e-}16 \\ \mathrm{P}(\texttt{end}|\texttt{I})\, \alpha_3(\texttt{I}) = .04 \cdot 2\text{e-}14 \\ \mathrm{P}(\texttt{end}|\texttt{O})\, \alpha_3(\texttt{O}) = .084 \cdot 5\text{e-}11 \end{cases} = 4\text{e-}12$$

The Forward algorithm

input: sequence $x_1, \ldots, x_L$, emission and transition probabilities

Forward: compute α recursively

$\alpha_1(y) = \mathrm{P}(y|\texttt{start}) \cdot \mathrm{P}(x_1|y) \quad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

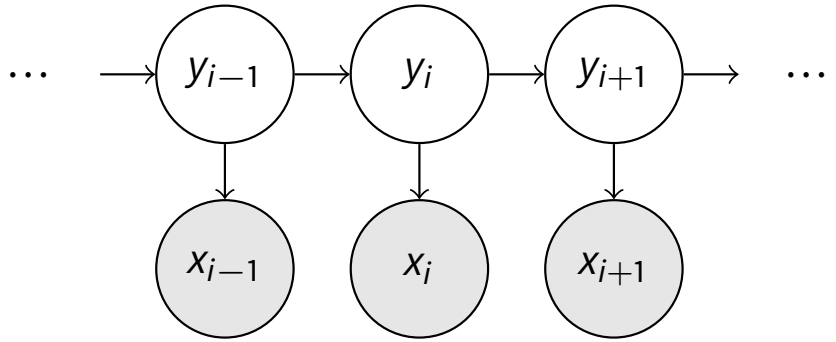
  for $y \in \Sigma$ **do**

    $\alpha_i(y) = \mathrm{P}(x_i|y) \cdot \sum_{y'} \Big(\mathrm{P}(y|y') \cdot \alpha_{i-1}(y') \Big)$

$p = \sum_{y'} \Big(\mathrm{P}(\texttt{end}|y') \cdot \alpha_L(y') \Big)$

output: the marginal likelihood $p = \mathrm{P}(\boldsymbol{x})$

Big HMM questions: Probability of an assignment

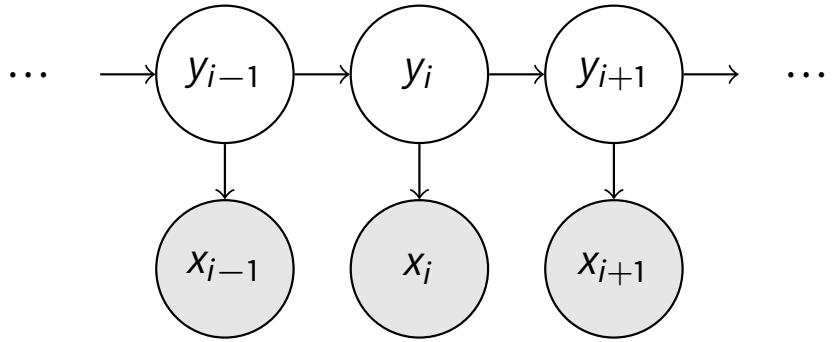


$$\mathrm{P}(y_i = y|\boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, \boldsymbol{x})}{\sum_{y'} \mathrm{P}(y_i = y', \boldsymbol{x})} = ?$$

The answer must be influenced by

- $y_1, \ldots, y_{i-1}$ (in turn depending only on $x_1, \ldots, x_{i-1}$)
- $y_{i+1}, \ldots, y_L$ (in turn depending only on $x_{i+1}, \ldots, x_L$)

Big HMM questions: Probability of an assignment



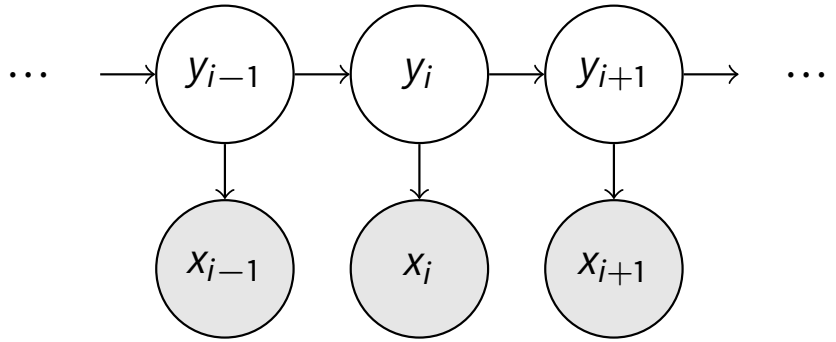
$$\mathrm{P}(y_i = y|\boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, \boldsymbol{x})}{\sum_{y'} \mathrm{P}(y_i = y', \boldsymbol{x})} = ?$$

The answer must be influenced by

- $y_1, \ldots, y_{i-1}$ (in turn depending only on $x_1, \ldots, x_{i-1}$)
- $y_{i+1}, \ldots, y_L$ (in turn depending only on $x_{i+1}, \ldots, x_L$)

$$\mathrm{P}(y_i = y, \boldsymbol{x}) = \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, y_{i+1}, \ldots, y_L, \boldsymbol{x})$$

Big HMM questions: Probability of an assignment



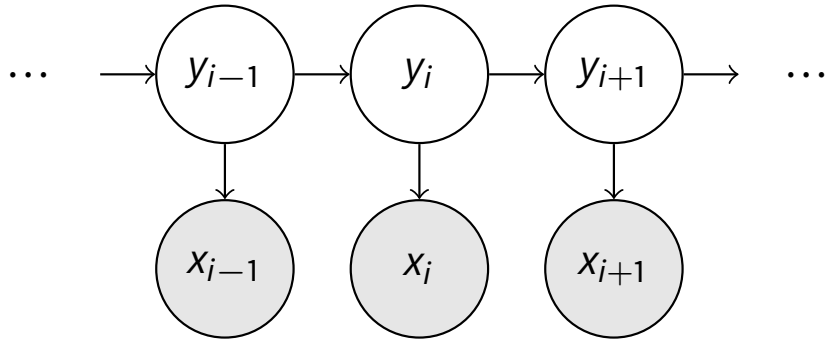
$$\mathrm{P}(y_i = y|\boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, \boldsymbol{x})}{\sum_{y'} \mathrm{P}(y_i = y', \boldsymbol{x})} = ?$$

The answer must be influenced by

- $y_1, \ldots, y_{i-1}$ (in turn depending only on $x_1, \ldots, x_{i-1}$)
- $y_{i+1}, \ldots, y_L$ (in turn depending only on $x_{i+1}, \ldots, x_L$)

$$\begin{aligned}
\mathrm{P}(y_i = y, \boldsymbol{x}) &= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, y_{i+1}, \ldots, y_L, \boldsymbol{x}) \\
&= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i) \cdot \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)
\end{aligned}$$

Big HMM questions: Probability of an assignment



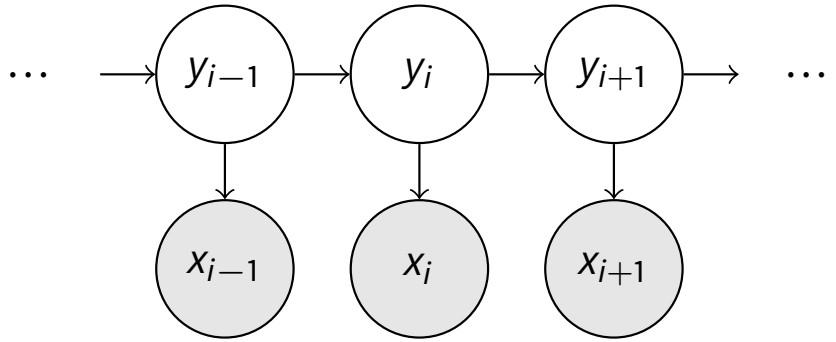
$$\mathrm{P}(y_i = y|\boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, \boldsymbol{x})}{\sum_{y'} \mathrm{P}(y_i = y', \boldsymbol{x})} = ?$$

The answer must be influenced by

- $y_1, \ldots, y_{i-1}$ (in turn depending only on $x_1, \ldots, x_{i-1}$)
- $y_{i+1}, \ldots, y_L$ (in turn depending only on $x_{i+1}, \ldots, x_L$)

$$\begin{aligned}
\mathrm{P}(y_i = y, \boldsymbol{x}) &= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, y_{i+1}, \ldots, y_L, \boldsymbol{x}) \\
&= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i) \cdot \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y) \\
&= \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1}, \ldots, y_L} \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{?}
\end{aligned}$$

Big HMM questions: Probability of an assignment



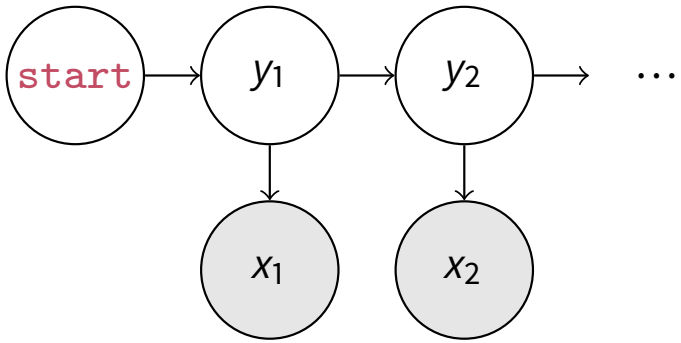
$$\mathrm{P}(y_i = y|\boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, \boldsymbol{x})}{\sum_{y'} \mathrm{P}(y_i = y', \boldsymbol{x})} = ?$$

The answer must be influenced by

- $y_1, \ldots, y_{i-1}$ (in turn depending only on $x_1, \ldots, x_{i-1}$)
- $y_{i+1}, \ldots, y_L$ (in turn depending only on $x_{i+1}, \ldots, x_L$)

$$\begin{aligned}
\mathrm{P}(y_i = y, \boldsymbol{x}) &= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, y_{i+1}, \ldots, y_L, \boldsymbol{x}) \\
&= \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_L} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i) \cdot \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y) \\
&= \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1}, \ldots, y_L} \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{\beta_i(y)}
\end{aligned}$$

Forward in reverse?!



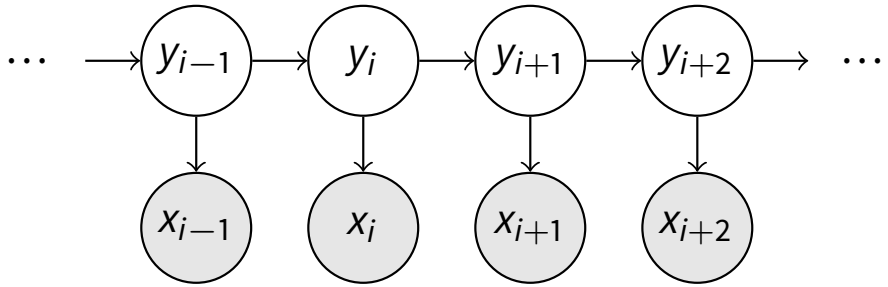
- β looks a lot like running forward in the other direction...

$$\mathrm{P}(\boldsymbol{x}) = \sum_y P(y|\texttt{start})\,\mathrm{P}(x_1|y) \underbrace{\sum_{y_2,\ldots,y_L} \mathrm{P}(x_2,\ldots,x_L,y_2,\ldots,y_L|y_1=y)}_{\beta_1(y)}$$

$$= \sum_y P(y|\texttt{start})\,\mathrm{P}(x_1|y) \sum_{y'} \mathrm{P}(y'|y)\,\mathrm{P}(x_2|y') \underbrace{\sum_{y_3,\ldots,y_L} \mathrm{P}(x_3,\ldots,x_L,y_3,\ldots,y_L|y_2=y')}_{\beta_2(y')}$$

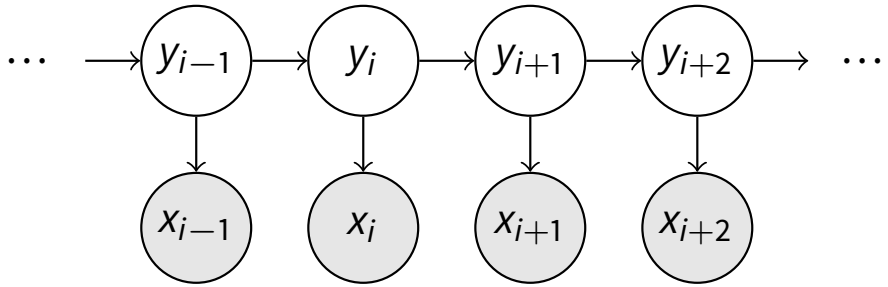
- Recurrence: $\beta_i(y) = \sum_{y'} \mathrm{P}(y'|y)\,\mathrm{P}(x_{i+1}|y')\,\beta_{i+1}(y')$

Big HMM questions: Probability of a transition



$$\mathrm{P}(y_i = y, y_{i+1} = y' | \boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}$$

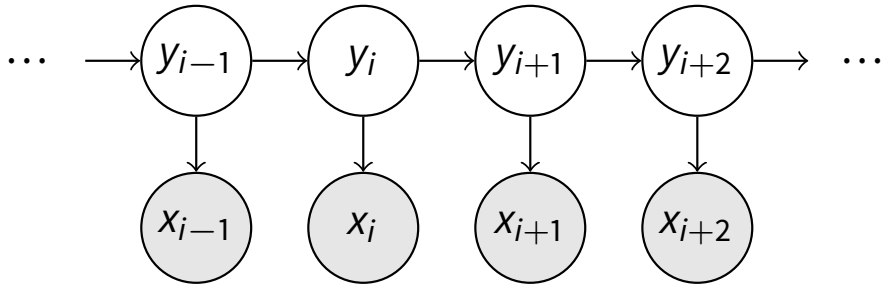
Big HMM questions: Probability of a transition



$$\mathrm{P}(y_i = y, y_{i+1} = y' | \boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}$$

$$\mathrm{P}(y_i = y, \boldsymbol{x}) = \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1}, \ldots, y_L} \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{\beta_i(y)}$$

Big HMM questions: Probability of a transition



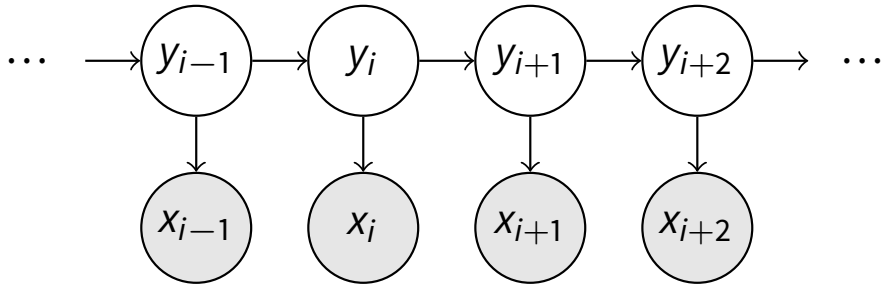
$$\mathrm{P}(y_i = y, y_{i+1} = y' | \boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}$$

$$\mathrm{P}(y_i = y, \boldsymbol{x}) = \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1}, \ldots, y_L} \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{\beta_i(y)}$$

$$\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x}) = \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \mathrm{P}(y' | y) \cdot \mathrm{P}(x_{i+1} | y')$$

$$\cdot \underbrace{\sum_{y_{i+2}, \ldots, y_L} \mathrm{P}(y_{i+2}, \ldots, y_L, x_{i+2}, \ldots, x_L | y_{i+1} = y')}_{\beta_{i+1}(y')}$$

Big HMM questions: Probability of a transition



$$\mathrm{P}(y_i = y, y_{i+1} = y' | \boldsymbol{x}) = \frac{\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}$$

$$\mathrm{P}(y_i = y, \boldsymbol{x}) = \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1}, \ldots, y_L} \mathrm{P}(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{\beta_i(y)}$$

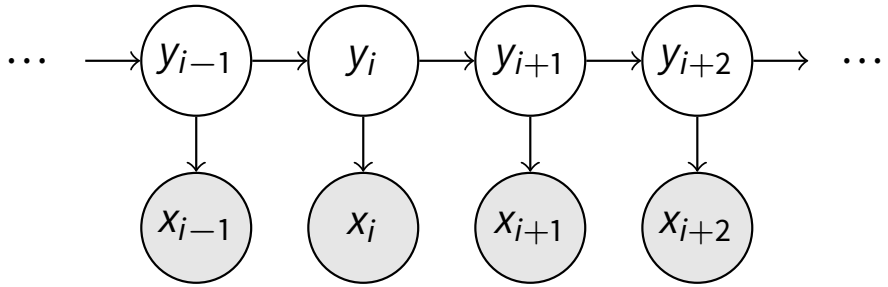
$$\mathrm{P}(y_i = y, y_{i+1} = y' \boldsymbol{x}) = \underbrace{\sum_{y_1, \ldots, y_{i-1}} \mathrm{P}(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \mathrm{P}(y' | y) \cdot \mathrm{P}(x_{i+1} | y')$$

$$\cdot \underbrace{\sum_{y_{i+2}, \ldots, y_L} \mathrm{P}(y_{i+2}, \ldots, y_L, x_{i+2}, \ldots, x_L | y_{i+1} = y')}_{\beta_{i+1}(y')}$$

Independent of position $(i, i+1)$:

$$\mathrm{P}([y, y'], \boldsymbol{x}) =$$

Big HMM questions: Probability of a transition



$$P(y_i = y, y_{i+1} = y' | \boldsymbol{x}) = \frac{P(y_i = y, y_{i+1} = y' \boldsymbol{x})}{P(\boldsymbol{x})}$$

$$P(y_i = y, \boldsymbol{x}) = \underbrace{\sum_{y_1,\ldots,y_{i-1}} P(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot \underbrace{\sum_{y_{i+1},\ldots,y_L} P(y_{i+1}, \ldots, y_L, x_{i+1}, \ldots, x_L | y_i = y)}_{\beta_i(y)}$$

$$P(y_i = y, y_{i+1} = y' \boldsymbol{x}) = \underbrace{\sum_{y_1,\ldots,y_{i-1}} P(y_1, \ldots, y_{i-1}, y_i = y, x_1, \ldots, x_i)}_{\alpha_i(y)} \cdot P(y'|y) \cdot P(x_{i+1}|y')$$

$$\underbrace{\cdot \sum_{y_{i+2},\ldots,y_L} P(y_{i+2}, \ldots, y_L, x_{i+2}, \ldots, x_L | y_{i+1} = y')}_{\beta_{i+1}(y')}$$

Independent of position $(i, i+1)$:

$$P([y, y'], \boldsymbol{x}) = \sum_{i=1}^{L-1} P(y_i = y, y_{i+1} = y', \boldsymbol{x}) = P(y'|y) \sum_{i=1}^{L-1} \alpha_i(y)\, P(x_{i+1}|y')\, \beta_{i+1}(y')$$

The Forward-Backward algorithm

input: sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$, emission and transition probabilities

Forward: compute α recursively

$\alpha_1(y) = \mathsf{P}(y|\texttt{start}) \cdot \mathsf{P}(\boldsymbol{x}_1|y) \quad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

  for $y \in \Sigma$ **do**

$$\alpha_i(y) = \mathsf{P}(x_i|y) \sum_{y'} \Big(\mathsf{P}(y|y') \cdot \alpha_{i-1}(y') \Big)$$

Forward: compute β recursively

$\beta_L(y) = \mathsf{P}(\texttt{end}|y) \quad \forall y \in \Sigma$

for $i = L - 1$ **down to** 1 **do**

  for $y \in \Sigma$ **do**

$$\beta_i(y) = \sum_{y'} \Big(\mathsf{P}(y'|y) \cdot \mathsf{P}(x_{i+1}|y') \cdot \beta_{i+1}(y') \Big)$$

output: The marginal likelihood $\mathsf{P}(\boldsymbol{x}) = \sum_{y'} \alpha_i(y')\, \beta_i(y')$ for any i;

Posterior unigram marginal probas $\mathsf{P}(y_i = y|\boldsymbol{x}) = \frac{\alpha_i(y)\, \beta_i(y)}{\mathsf{P}(\boldsymbol{x})}$;

Posterior transition marginal probas $\mathsf{P}([y, y']|\boldsymbol{x}) = \frac{\mathsf{P}(y'|y)}{\mathsf{P}(\boldsymbol{x})} \sum_{i=1}^{L-1} \alpha_i(y)\, \mathsf{P}(x_{i+1}|y')\, \beta_{i+1}(y')$.

Viterbi recurrence:

$$\text{score}_i(y) = \text{P}(x_i|y) \max_{y'} \text{P}(y|y') \, \text{score}_{i-1}(y')$$

Forward recurrence:

$$\alpha_i(y) = \text{P}(x_i|y) \sum_{y'} \text{P}(y|y') \, \alpha_{i-1}(y')$$

Backward recurrence:

$$\beta_i(y) = \sum_{y'} \text{P}(y'|y) \, \text{P}(x_{i+1}|y') \, \beta_{i+1}(y')$$

In log-domain:

Viterbi recurrence:

$$\widetilde{\text{score}}_i(y) = \log \mathrm{P}(x_i|y) + \max_{y'} \Big(\log \mathrm{P}(y|y') + \widetilde{\text{score}}_{i-1}(y') \Big)$$

Forward recurrence:

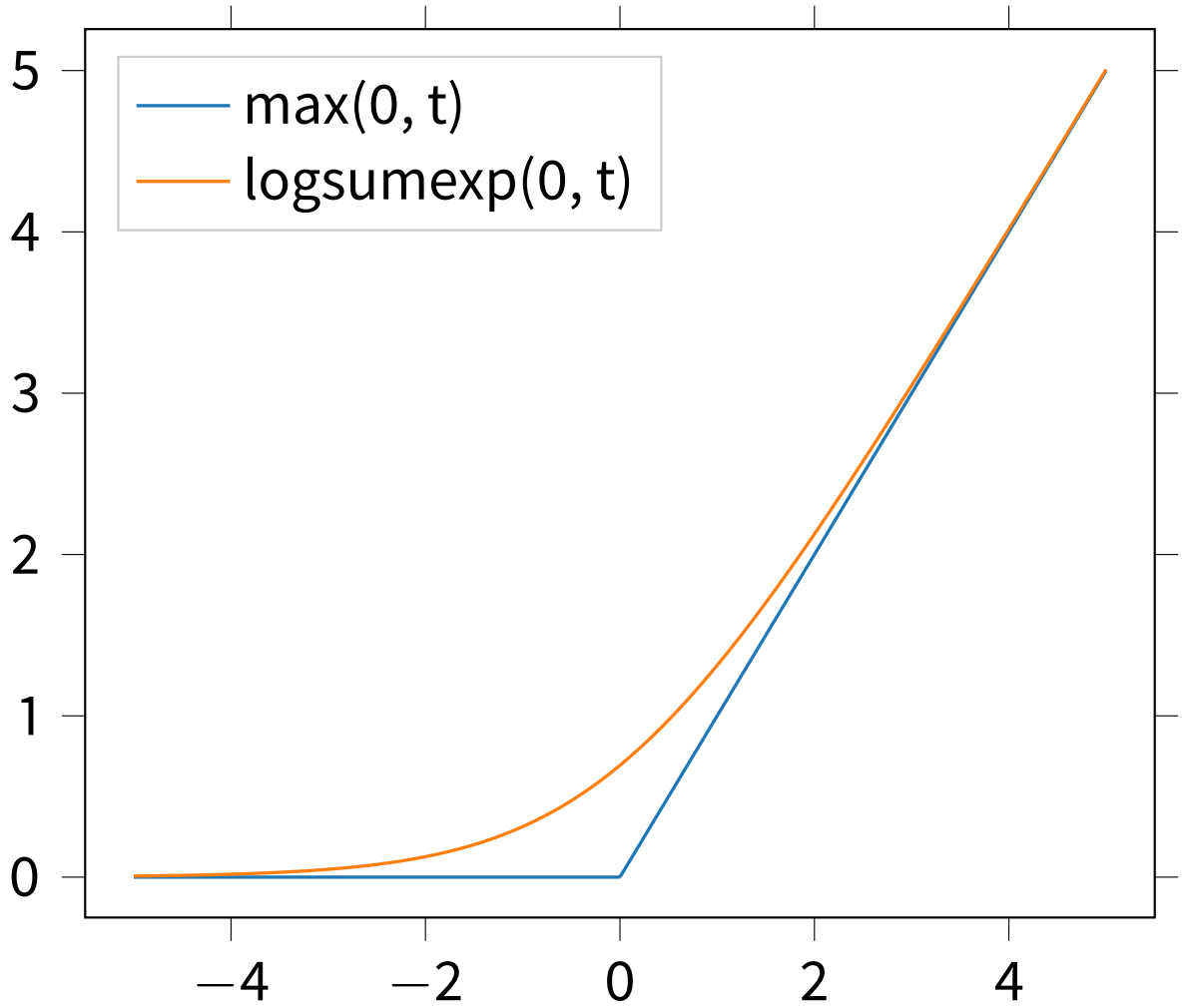
$$\widetilde{\alpha}_i(y) = \log \mathrm{P}(x_i|y) + \log \sum_{y'} \exp \Big(\log \mathrm{P}(y|y') + \widetilde{\alpha}_{i-1}(y') \Big)$$

Backward recurrence:

$$\widetilde{\beta}_i(y) = \log \sum_{y'} \exp \Big(\log \mathrm{P}(y'|y) + \log \mathrm{P}(x_{i+1}|y') + \widetilde{\beta}_{i+1}(y') \Big)$$

note: $\widetilde{\alpha}_i(y) = \log \alpha_i(y)$, etc.

Log-sum-exp



The Forward-Backward algorithm in log-domain

input: sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$, emission and transition **log**-probabilities

Forward: compute $\widetilde{\alpha}$ recursively

$\widetilde{\alpha}_1(y) = \log \mathrm{P}(y|\texttt{start}) + \log \mathrm{P}(\boldsymbol{x}_1|y) \quad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

  for $y \in \Sigma$ **do**

$$\widetilde{\alpha}_i(y) = \log \mathrm{P}(x_i|y) + \log \sum_{y'} \exp\Big(\log \mathrm{P}(y|y') + \widetilde{\alpha}_{i-1}(y')\Big)$$

Backward: compute $\widetilde{\beta}$ recursively

$\widetilde{\beta}_L(y) = \log \mathrm{P}(\texttt{end}|y) \quad \forall y \in \Sigma$

for $i = L - 1$ **down to** 1 **do**

  for $y \in \Sigma$ **do**

$$\widetilde{\beta}_i(y) = \log \sum_{y'} \exp\Big(\log \mathrm{P}(y'|y) + \log \mathrm{P}(x_{i+1}|y') + \widetilde{\beta}_{i+1}(y')\Big)$$

output: The log-marginal $\log \mathrm{P}(\boldsymbol{x}) = \log \sum_{y'} \exp\Big(\widetilde{\alpha}_i(y') + \widetilde{\beta}_i(y')\Big)$ for any i;

Posterior marginal log-probas: unigram

$\log \mathrm{P}(y_i = y|\boldsymbol{x}) = \widetilde{\alpha}_i(y) + \widetilde{\beta}_i(y) - \log \mathrm{P}(\boldsymbol{x})$;

and transition:

$$\log \mathrm{P}([y, y']|\boldsymbol{x}) = \log \mathrm{P}(y'|y) + \log \sum_{i=1}^{L-1} \exp\Big(\widetilde{\alpha}_i(y) + \log \mathrm{P}(x_{i+1}|y') + \widetilde{\beta}_{i+1}(y')\Big) - \log \mathrm{P}(\boldsymbol{x}).$$

note: $\widetilde{\alpha}_i(y) = \log \alpha_i(y), \widetilde{\beta}_i(y) = \log \beta_i(y)$

Big HMM questions

1. What is the most likely label sequence $\boldsymbol{y}$, given $\boldsymbol{x}$?
2. What is the probability of $\boldsymbol{x}$?
3. What is the probability of each assignment y_i, given $\boldsymbol{x}$?
4. What is the probability of each transition $y \rightarrow y'$, given $\boldsymbol{x}$?
5. What sequence $\boldsymbol{y}$ minimizes the *Hamming cost?*

Minimizing costs, a.k.a., risks

- Our HMM defines a distribution over labelings $\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})$.
- The HMM will be given a new sequence $\boldsymbol{x}$ to label, producing $\hat{\boldsymbol{y}}$.
- We then observe the true $\boldsymbol{y}^\star$, and incur a cost

$$\text{cost}(\hat{\boldsymbol{y}}, \boldsymbol{y}^\star).$$

- How do we predict so as to minimize the expected cost

$$\hat{\boldsymbol{y}} = \arg\min_{\boldsymbol{y}} \mathbb{E}_{\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})} \Big[\text{cost}(\boldsymbol{y}, \boldsymbol{y}^\star) \Big]$$

Minimizing costs, a.k.a., risks

- Consider the sequence zero-one cost:

$$\text{cost}_{01}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \begin{cases} 1, & \boldsymbol{y} \neq \hat{\boldsymbol{y}} \\ 0, & \boldsymbol{y} = \hat{\boldsymbol{y}} \end{cases}$$

- The cost we may expect to pay is

$$\begin{aligned} \mathbb{E}_{\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})}\Big[\text{cost}_{01}(\boldsymbol{y}, \boldsymbol{y}^\star)\Big] &= \sum_{\boldsymbol{y} \neq \boldsymbol{y}^\star} \mathsf{P}(\boldsymbol{y}|\boldsymbol{x}) \\ &= 1 - \mathsf{P}(\boldsymbol{y}^\star|\boldsymbol{x}) \end{aligned}$$

- We should return the sequence it assigns most probability to:

$$\hat{\boldsymbol{y}} = \arg\min_{\boldsymbol{y}} \big(1 - \mathsf{P}(\boldsymbol{y}|\boldsymbol{x})\big) = \arg\max_{\boldsymbol{y}} \mathsf{P}(\boldsymbol{y}|\boldsymbol{x})$$

- Viterbi computes this sequence!

Minimizing costs, a.k.a., risks

- Now consider the Hamming (word-wise) cost:

$$\text{cost}_H(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_i \text{cost}_{01}(y_i, \hat{y}_i)$$

- The expected cost we pay is

$$\begin{aligned}\mathbb{E}_{\text{P}(\boldsymbol{y}|\boldsymbol{x})}\Big[\text{cost}_H(\boldsymbol{y}, \boldsymbol{y}^\star)\Big] &= \sum_i \sum_y \text{cost}_{01}(y, \hat{y}_i)\,\text{P}(y_i = y|\boldsymbol{x}) \\ &= \sum_i \big(1 - \text{P}(y_i = y_i^\star|\boldsymbol{x})\big)\end{aligned}$$

- **Posterior decoding:** Get α, β (Forward-Backward); pick $\hat{\boldsymbol{y}}$ such that

$$\hat{y}_i = \arg\max_y \text{P}(y_i = y|\boldsymbol{x}) = \arg\max_y \big(\widetilde{\alpha}_i(y) + \widetilde{\beta}_i(y)\big)$$

- Exercise: This can be extended for any cost of the form $\text{cost}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{L} c_u(y_i, \hat{y}_i) + \sum_{i=1}^{L-1} c_t(y_i, y_{i+1}, \hat{y}_i, \hat{y}_{i+1})$

Big HMM questions

1. What is the most likely label sequence $\boldsymbol{y}$, given $\boldsymbol{x}$?
2. What is the probability of $\boldsymbol{x}$?
3. What is the probability of each assignment y_i, given $\boldsymbol{x}$?
4. What is the probability of each transition $y \rightarrow y'$, given $\boldsymbol{x}$?
5. What sequence $\boldsymbol{y}$ minimizes the *Hamming cost?*

Two important algorithms for sequential models

All of these big questions are solved by these two similar algorithms.

Viterbi
(find the max sequence)

Forward-Backward
(find the *average sequence*)

Dealing with other types of data

- So far $\boldsymbol{y}$ was a sequence of labels, and $\boldsymbol{x}$ a sequence of **words**, so emission probabilities $\mathrm{P}(x_i|y_i)$ are just tables.
- Everything works with other choices for $\mathrm{P}(x_i|y_i)$. Examples:

Dealing with other types of data

- So far $\boldsymbol{y}$ was a sequence of labels, and $\boldsymbol{x}$ a sequence of **words**, so emission probabilities $\mathrm{P}(x_i|y_i)$ are just tables.
- Everything works with other choices for $\mathrm{P}(x_i|y_i)$. Examples:
  - x_i are sentences. Example: sentence-level review sentiment.

$\boldsymbol{x}$	$\boldsymbol{y}$
x_1: `I bought this knife set last year.`	y_1: `neutral`
x_2: `I was pleasantly surprised with it.`	y_2: `positive`
x_3: `They're still sharp.`	y_3: `positive`

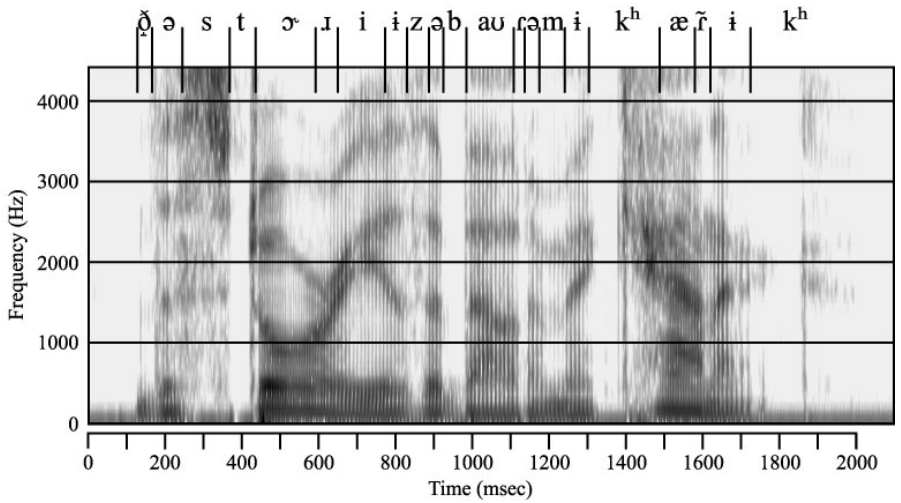
    Naïve Bayes assumption for emissions: $\mathrm{P}(x_i|y_i) = \prod_{w \in x_i} \mathrm{P}(w|y_i)$

Dealing with other types of data

- So far $\boldsymbol{y}$ was a sequence of labels, and $\boldsymbol{x}$ a sequence of **words**, so emission probabilities $\mathrm{P}(x_i|y_i)$ are just tables.
- Everything works with other choices for $\mathrm{P}(x_i|y_i)$. Examples:
  - x_i are sentences. Example: sentence-level review sentiment.

$\boldsymbol{x}$	$\boldsymbol{y}$
x_1: `I bought this knife set last year.`	y_1: `neutral`
x_2: `I was pleasantly surprised with it.`	y_2: `positive`
x_3: `They're still sharp.`	y_3: `positive`

    Naïve Bayes assumption for emissions: $\mathrm{P}(x_i|y_i) = \prod_{w \in x_i} \mathrm{P}(w|y_i)$
  - $x_i \in \mathbb{R}^d$ are continuous vectors. Example: speech-to-phoneme



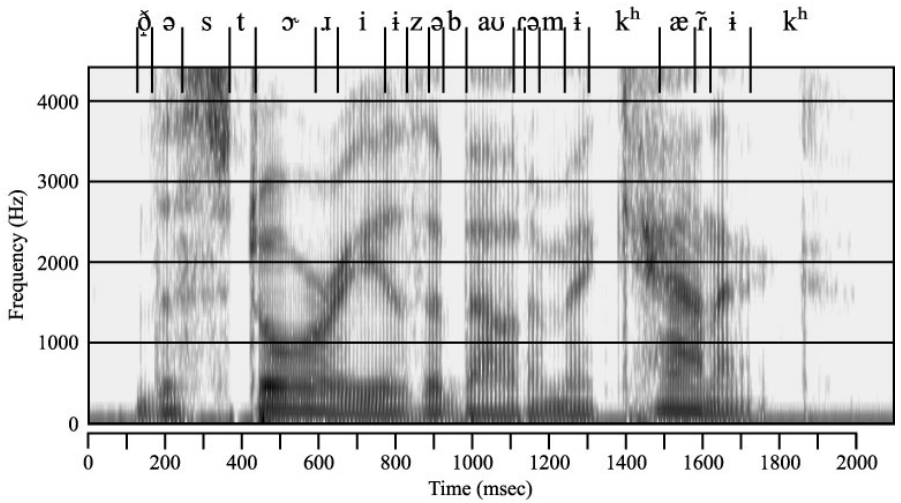
Gaussian emissions:
$\mathrm{P}(x_i|y_i) = \mathcal{N}(\mu_{y_i}, S_{y_i})$

Dealing with other types of data

- So far $\boldsymbol{y}$ was a sequence of labels, and $\boldsymbol{x}$ a sequence of **words**, so emission probabilities $\mathrm{P}(x_i|y_i)$ are just tables.
- Everything works with other choices for $\mathrm{P}(x_i|y_i)$. Examples:
  - x_i are sentences. Example: sentence-level review sentiment.

$\boldsymbol{x}$	$\boldsymbol{y}$
x_1: `I bought this knife set last year.`	y_1: `neutral`
x_2: `I was pleasantly surprised with it.`	y_2: `positive`
x_3: `They're still sharp.`	y_3: `positive`

    Naïve Bayes assumption for emissions: $\mathrm{P}(x_i|y_i) = \prod_{w \in x_i} \mathrm{P}(w|y_i)$
  - $x_i \in \mathbb{R}^d$ are continuous vectors. Example: speech-to-phoneme



Gaussian emissions:
$\mathrm{P}(x_i|y_i) = \mathcal{N}(\mu_{y_i}, S_{y_i})$

- Or, turn to feature-based discriminative models (later.)

Outline

Unsupervised learning

How to estimate the emission probabilities $\mathrm{P}(\boldsymbol{x}_i|\boldsymbol{y}_i)$ and the transition probabilities $\mathrm{P}(\boldsymbol{y}_i|\boldsymbol{y}_{i-1})$?

1. **Supervised learning:** assumes we have *labeled* training data $\{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})\}$ – we've seen this!

2. **Unsupervised learning:** assumes all we have is *unlabeled* training data $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$

HMM Unsupervised Learning

Assumes all we have is *unlabeled* training data $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$

Maximum Likelihood Estimation with incomplete data!

$$\text{maximize} \prod_{n=1}^{N} \mathrm{P}(\boldsymbol{x}^{(n)}) = \prod_{n=1}^{N} \sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$$

Algorithm: **Expectation-Maximization (EM)**.

Expectation-Maximization (Baum-Welch)

If we knew $\boldsymbol{y}$, could fit HMM by counting and normalizing

If we knew the model parameters, we could estimate the posterior marginal probabilities (soft counts) $\mathsf{P}(\boldsymbol{y}_i \mid \boldsymbol{x})$ and $\mathsf{P}(\boldsymbol{y}_{i-1}, \boldsymbol{y}_i \mid \boldsymbol{x})$

This is a chicken-and-egg problem!

Expectation-Maximization (Baum-Welch)

Initialize HMM at random.

Alternate:

- **E-step:** Get a soft-labelling of $\boldsymbol{x}$ from current model, keep track of soft counts $\mathsf{P}(\boldsymbol{y}_i \mid \boldsymbol{x})$ and $\mathsf{P}(\boldsymbol{y}_{i-1}, \boldsymbol{y}_i \mid \boldsymbol{x})$
  - Forward-Backward for each data point.
- **M-step:** Do a "supervised" update of the HMM
  - Count & normalize the soft-labels!

Guarantees improvement, but the problem has multiple optima.

Summary of HMMs

- Assumptions? **Markov assumption on states; words are conditionally independent given the state.**
- Decoding algorithms: **Viterbi/forward-backward.**
- Learning? **Maximum likelihood (count and normalize)** for the supervised case, **EM** for the unsupervised case.

So far: Structured prediction models

Binary/Multi-class	Structured Prediction
Naive Bayes	HMMs
Perceptron	?
Logistic Regression	?
SVMs	?

Outline

Recall: Generative vs discriminative

Generative models

- Model the joint $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ by choosing $\mathrm{P}(\boldsymbol{y})$, $\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
  - E.g., Naïve Bayes
  - Easy to fit: just count, one pass over the data.
- Make predictions: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Use Bayes' rule to get $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

Recall: Generative vs discriminative

Generative models

- Model the joint $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ by choosing $\mathrm{P}(\boldsymbol{y})$, $\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
  - E.g., Naïve Bayes
  - Easy to fit: just count, one pass over the data.
- Make predictions: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Use Bayes' rule to get $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

Discriminative models

- Directly try to assign a **score** to each class
  - linear, feature-driven: $s_{\boldsymbol{y}} = \boldsymbol{w} \cdot \phi(\boldsymbol{x}, \boldsymbol{y})$,
  - or a neural network: $s_{\boldsymbol{y}} = f(\boldsymbol{y}; \boldsymbol{x})$

Recall: Generative vs discriminative

Generative models

- Model the joint $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ by choosing $\mathrm{P}(\boldsymbol{y})$, $\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
  - E.g., Naïve Bayes
  - Easy to fit: just count, one pass over the data.
- Make predictions: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Use Bayes' rule to get $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

Discriminative models

- Directly try to assign a **score** to each class
  - linear, feature-driven: $s_{\boldsymbol{y}} = \boldsymbol{w} \cdot \phi(\boldsymbol{x}, \boldsymbol{y})$,
  - or a neural network: $s_{\boldsymbol{y}} = f(\boldsymbol{y}; \boldsymbol{x})$
- Make predictions: $\arg\max_{\boldsymbol{y}} s_{\boldsymbol{y}}$

Recall: Generative vs discriminative

Generative models

- Model the joint $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ by choosing $\mathrm{P}(\boldsymbol{y})$, $\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
  - E.g., Naïve Bayes
  - Easy to fit: just count, one pass over the data.
- Make predictions: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Use Bayes' rule to get $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

Discriminative models

- Directly try to assign a **score** to each class
  - linear, feature-driven: $s_{\boldsymbol{y}} = \boldsymbol{w} \cdot \phi(\boldsymbol{x}, \boldsymbol{y})$,
  - or a neural network: $s_{\boldsymbol{y}} = f(\boldsymbol{y}; \boldsymbol{x})$
- Make predictions: $\arg\max_{\boldsymbol{y}} s_{\boldsymbol{y}}$
- Harder to train, needs iterative optimization
  - Perceptron: Try to make $s_{\boldsymbol{y}^{\text{true}}} > s_{\boldsymbol{y}}$
  - Logistic regression: Define $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \exp(s_{\boldsymbol{y}})$, maximize $\mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x})$.
  - SVM: Try to make $s_{\boldsymbol{y}^{\text{true}}} > 1 + s_{\boldsymbol{y}}$

Recall: Generative vs discriminative

Generative models

- Model the joint $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ by choosing $\mathrm{P}(\boldsymbol{y})$, $\mathrm{P}(\boldsymbol{x}|\boldsymbol{y})$
  - E.g., Naïve Bayes
  - Easy to fit: just count, one pass over the data.
- Make predictions: $\arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$
- Use Bayes' rule to get $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x})$

In HMM, with Viterbi resp. Forward-Backward

Discriminative models

- Directly try to assign a **score** to each class
  - linear, feature-driven: $s_{\boldsymbol{y}} = \boldsymbol{w} \cdot \phi(\boldsymbol{x}, \boldsymbol{y})$,
  - or a neural network: $s_{\boldsymbol{y}} = f(\boldsymbol{y}; \boldsymbol{x})$
- Make predictions: $\arg\max_{\boldsymbol{y}} s_{\boldsymbol{y}}$
- Harder to train, needs iterative optimization
  - Perceptron: Try to make $s_{\boldsymbol{y}^{\text{true}}} > s_{\boldsymbol{y}}$
  - Logistic regression: Define $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \propto \exp(s_{\boldsymbol{y}})$, maximize $\mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x})$.
  - SVM: Try to make $s_{\boldsymbol{y}^{\text{true}}} > 1 + s_{\boldsymbol{y}}$

Does the same trick work?

Outline

Perceptron

Recall the simple & powerful Perceptron algorithm.

- Process one pair $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$ at each round
  - Take $\boldsymbol{x}$; predict a sequence $\widehat{\boldsymbol{y}}$.
  - If prediction is correct, proceed. If not, adjust.

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

Perceptron

Recall the simple & powerful Perceptron algorithm.

- Process one pair $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$ at each round
  - Take $\boldsymbol{x}$; predict a sequence $\widehat{\boldsymbol{y}}$.
  - If prediction is correct, proceed. If not, adjust.

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

Perceptron

Recall the simple & powerful Perceptron algorithm.

- Process one pair $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$ at each round
  - Take $\boldsymbol{x}$; predict a sequence $\widehat{\boldsymbol{y}}$.
  - If prediction is correct, proceed. If not, adjust.

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

Why couldn't y be an entire sequence here?

Perceptron

Recall the simple & powerful Perceptron algorithm.

- Process one pair $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$ at each round
  - Take $\boldsymbol{x}$; predict a sequence $\widehat{\boldsymbol{y}}$.
  - If prediction is correct, proceed. If not, adjust.

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

Why couldn't y be an entire sequence here?
Mathematically, all is cool. Algorithmically...

Perceptron

Recall the simple & powerful Perceptron algorithm.

- Process one pair $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$ at each round
  - Take $\boldsymbol{x}$; predict a sequence $\widehat{\boldsymbol{y}}$.
  - If prediction is correct, proceed. If not, adjust.

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

Why couldn't y be an entire sequence here?
Mathematically, all is cool. Algorithmically...

Perceptron

Everything would be fine if we had a way to calculate

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$$

This would work, but prohibitive:.

for $\boldsymbol{y} = [\texttt{start}, y_1, \ldots, y_L, \texttt{end}] \in \mathcal{Y}$ **do**
  compute $f(\boldsymbol{y}; \boldsymbol{x})$ # e.g. $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$, or nnet forward pass
return the highest scoring $\boldsymbol{y}$

Perceptron

Everything would be fine if we had a way to calculate

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$$

This would work, but prohibitive:.

for $\boldsymbol{y} = [\texttt{start}, y_1, \ldots, y_L, \texttt{end}] \in \mathcal{Y}$ **do**
  compute $f(\boldsymbol{y}; \boldsymbol{x})$ # e.g. $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$, or nnet forward pass
return the highest scoring $\boldsymbol{y}$

This looks similar to the problem Viterbi solved in HMMs!

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} \log \mathsf{P}(\boldsymbol{y}, \boldsymbol{x})$$

Perceptron

Everything would be fine if we had a way to calculate

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$$

This would work, but prohibitive:.

for $\boldsymbol{y} = [\texttt{start}, y_1, \ldots, y_L, \texttt{end}] \in \mathcal{Y}$ **do**
  compute $f(\boldsymbol{y}; \boldsymbol{x})$ # e.g. $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$, or nnet forward pass
return the highest scoring $\boldsymbol{y}$

This looks similar to the problem Viterbi solved in HMMs!

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} \log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x})$$

What magic made Viterbi work there? – can we replicate it?

The HMM magic formula: decomposition into parts

Viterbi was able to efficiently compute

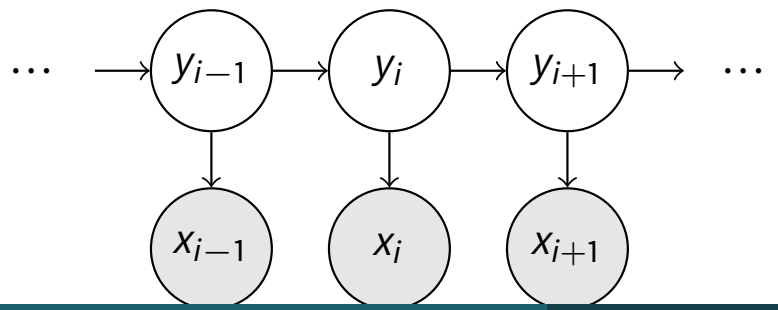
$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} \log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x})$$

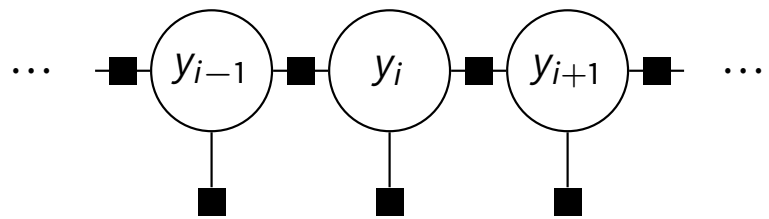
because of the **decomposition into parts**:

$$\log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{\log \mathrm{P}(x_i | y_i)}_{\text{emission log-proba}} + \sum_{i=1}^{L+1} \underbrace{\log \mathrm{P}(y_i | y_{i-1})}_{\text{transition log-proba}}$$

We should try to design our scorer f such that

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{f_i^{(u)}(y_i; \boldsymbol{x})}_{\text{unary score}} + \sum_{i=2}^{L} \underbrace{f_i^{(t)}(y_i, y_{i-1}; \boldsymbol{x})}_{\text{transition score}}$$





The HMM magic formula: decomposition into parts

Viterbi was able to efficiently compute

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} \log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x})$$

because of the **decomposition into parts**:

$$\log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{\log \mathrm{P}(x_i|y_i)}_{\text{emission log-proba}} + \sum_{i=1}^{L+1} \underbrace{\log \mathrm{P}(y_i|y_{i-1})}_{\text{transition log-proba}}$$

We should try to design our scorer f such that

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{f_i^{(u)}(y_i; \boldsymbol{x})}_{\text{unary score}} + \sum_{i=2}^{L} \underbrace{f_i^{(t)}(y_i, y_{i-1}; \boldsymbol{x})}_{\text{transition score}}$$

Given $\boldsymbol{x}$, unary scores form a $|\Sigma| \times L$ array $s^{(u)}_{y,i}$
and transition scores form a $|\Sigma| \times |\Sigma| \times L$ array $s^{(t)}_{y',y,i}$

Transitions from `start` and to `end`?

The HMM magic formula: decomposition into parts

Viterbi was able to efficiently compute

$$\arg\max_{\boldsymbol{y} \in \mathcal{Y}} \log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x})$$

because of the **decomposition into parts**:

$$\log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{\log \mathrm{P}(x_i|y_i)}_{\text{emission log-proba}} + \sum_{i=1}^{L+1} \underbrace{\log \mathrm{P}(y_i|y_{i-1})}_{\text{transition log-proba}}$$

We should try to design our scorer f such that

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} \underbrace{f_i^{(u)}(y_i; \boldsymbol{x})}_{\text{unary score}} + \sum_{i=2}^{L} \underbrace{f_i^{(t)}(y_i, y_{i-1}; \boldsymbol{x})}_{\text{transition score}}$$

Given $\boldsymbol{x}$, unary scores form a $|\Sigma| \times L$ array $s^{(u)}_{y,i}$
and transition scores form a $|\Sigma| \times |\Sigma| \times L$ array $s^{(t)}_{y',y,i}$

Transitions from `start` and to `end`? Can be added to $s^{(u)}_{y,1}$, $s^{(u)}_{y,L}$

Score-based Viterbi

input: Unary scores $s^{(u)}$ ($|\Sigma| \times L$ array)
Transition scores $s^{(t)}$ ($|\Sigma| \times |\Sigma| \times (L-1)$ array)

Forward: compute scores recursively

$\widetilde{\text{score}}_1(y) = s^{(u)}_{y,1} \qquad \forall y \in \Sigma$

for $i = 2$ **to** L **do**

  for $y \in \Sigma$ **do**

$$\widetilde{\text{score}}_i(y) = s^{(u)}_{y,i} + \max_{y'} \left(s^{(t)}_{y',y,i} + \widetilde{\text{score}}_{i-1}(y') \right)$$

$$\text{backptr}_i(y) = \arg\max_{y'} \left(s^{(t)}_{y',y,i} + \widetilde{\text{score}}_{i-1}(y') \right)$$

Backward: follow backpointers

$f^\star = \max_{y'} \widetilde{\text{score}}_L(y')$

$\widehat{y}_L = \arg\max_{y'} \widetilde{\text{score}}_L(y')$

for $i = L-1$ **down to** 1 **do**

  $\widehat{y}_i = \text{backptr}_{i+1}(\widehat{y}_{i+1})$

output: The highest-scoring $\widehat{\boldsymbol{y}} = [\widehat{y}_1, \ldots, \widehat{y}_L]$ and its total score $f^\star$.

Decomposable scorers

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i,i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1},y_i,i}$$

- More expressive than an HMM: can look at entire $\boldsymbol{x}$. Useful cases:
- We can simulate an HMM: $\boldsymbol{W}^{(t)} \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$, $\boldsymbol{W}^{(u)} \in \mathbb{R}^{|\Sigma|\times|V|}$;
  - $s^{(t)}_{y',y,i} = w^{(t)}_{y',y}$ (ignore $\boldsymbol{x}$ and i)
  - $s^{(u)}_{y,i} = w^{(u)}_{y,x_i}$

Decomposable scorers

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i,i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1},y_i,i}$$

- More expressive than an HMM: can look at entire $\boldsymbol{x}$. Useful cases:
- We can simulate an HMM: $\boldsymbol{W}^{(t)} \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$, $\boldsymbol{W}^{(u)} \in \mathbb{R}^{|\Sigma|\times|V|}$;
  - $s^{(t)}_{y',y,i} = w^{(t)}_{y',y}$ (ignore $\boldsymbol{x}$ and i)
  - $s^{(u)}_{y,i} = w^{(u)}_{y,x_i}$
- Transitions as above; linear / neural model for unary scores
  - A word-level classifier augmented with transition scores.

Decomposable scorers

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i,i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1},y_i,i}$$

- More expressive than an HMM: can look at entire $\boldsymbol{x}$. Useful cases:
- We can simulate an HMM: $\boldsymbol{W}^{(t)} \in \mathbb{R}^{|\Sigma| \times |\Sigma|}$, $\boldsymbol{W}^{(u)} \in \mathbb{R}^{|\Sigma| \times |V|}$;
  - $s^{(t)}_{y',y,i} = w^{(t)}_{y',y}$ (ignore $\boldsymbol{x}$ and i)
  - $s^{(u)}_{y,i} = w^{(u)}_{y,x_i}$
- Transitions as above; linear / neural model for unary scores
  - A word-level classifier augmented with transition scores.
  - Linear scores: $s^{(u)}_{y,i} = \boldsymbol{w}^{(u)} \cdot \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y)$ Unary features:
    - $y = \mathrm{B}, x_i$ capitalized
    - $y = \mathrm{B}, x_{i-1}$ indefinite article (a/an)

Decomposable scorers

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i,i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1},y_i,i}$$

- More expressive than an HMM: can look at entire $\boldsymbol{x}$. Useful cases:
- We can simulate an HMM: $\boldsymbol{W}^{(t)} \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$, $\boldsymbol{W}^{(u)} \in \mathbb{R}^{|\Sigma|\times|V|}$;
  - $s^{(t)}_{y',y,i} = w^{(t)}_{y',y}$ (ignore $\boldsymbol{x}$ and i)
  - $s^{(u)}_{y,i} = w^{(u)}_{y,x_i}$
- Transitions as above; linear / neural model for unary scores
  - A word-level classifier augmented with transition scores.
  - Linear scores: $s^{(u)}_{y,i} = \boldsymbol{w}^{(u)} \cdot \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y)$ Unary features:
    - $y = \mathrm{B}, x_i$ capitalized
    - $y = \mathrm{B}, x_{i-1}$ indefinite article (a/an)
  - Neural unary scores (example)
    1. Encode each word into a vector $\boldsymbol{h}_i = g(\boldsymbol{x}, i)$.
    2. apply "output layer": $s^{(u)}_{y,i} = (\boldsymbol{W}^{(out)}\boldsymbol{h}_i + \boldsymbol{b}^{(out)})_y$

Decomposable feature-based scorers

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1}, y_i, i}$$

- Lots of NLP literature uses feature representations for everything.

$$s^{(u)}_{y,i} = \boldsymbol{w}^{(u)} \cdot \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y)$$

$$s^{(t)}_{y',y,i} = \boldsymbol{w}^{(t)} \cdot \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), (y', y))$$

or, compactly, $$f(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

where $$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) = \text{cat}\left[\sum_i \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i),\ \sum_i \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y_i, y_{i+1})\right]$$

$$\boldsymbol{w} = \text{cat}\left[\boldsymbol{w}^{(u)},\ \boldsymbol{w}^{(t)}\right]$$

- Unary features: just like in multi-class. Transition features:
  - $y = \texttt{B}, y' = \texttt{O}$ (equivalent to learning transition scores)
  - $y = \texttt{B}, y' = \texttt{O}, i = 3$
  - $y = \texttt{B}, y' = \texttt{O}, x_i$ capitalized, $x_{i-1} =$ "from"
  - $y = \texttt{B}, y' = \texttt{O}, x_{\text{last}} =$ "?"

Structured Perceptron

input: labeled data $\mathcal{D}$
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  compute unary and transition scores, $s^{(u)}$, $s^{(t)}$.
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x})$ **using Viterbi**.
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}^{\text{true}}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})$
until max. epochs
output: model weights $\boldsymbol{w}$

- If linear $f(\boldsymbol{y}; \boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$, then $\nabla_{\boldsymbol{w}} f(\boldsymbol{y}; \boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$.
- If neural, $\nabla_{\boldsymbol{w}} f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} \nabla_{\boldsymbol{w}} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} \nabla_{\boldsymbol{w}} s^{(t)}_{y_{i-1}, y_i, i}$ from autodiff.

Outline

Logistic Regression

What if we want a **discriminative probabilistic model**?

i.e., one that gives $\mathsf{P}(\boldsymbol{y}|\boldsymbol{x})$ and not just a prediction $\widehat{\boldsymbol{y}}$?

For multi-class, we had **logistic regression**.

$$\mathsf{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s_{\boldsymbol{y}}}{\sum'_{\boldsymbol{y}} \exp s'_{\boldsymbol{y}}}$$

Logistic Regression

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s_{\boldsymbol{y}}}{\sum'_{\boldsymbol{y}} \exp s_{\boldsymbol{y}'}}$$

To learn, we maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = s_{\boldsymbol{y}^{\text{true}}} - \log \sum_{\boldsymbol{y}} \exp s_{\boldsymbol{y}}$$

with gradient descent, noting that

$$\nabla \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \nabla s_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}_{\boldsymbol{y}} \nabla s_{\boldsymbol{y}}.$$

Logistic Regression

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s_{\boldsymbol{y}}}{\sum'_{\boldsymbol{y}} \exp s_{\boldsymbol{y}'}}$$

To learn, we maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = s_{\boldsymbol{y}^{\text{true}}} - \log \sum_{\boldsymbol{y}} \exp s_{\boldsymbol{y}}$$

with gradient descent, noting that

$$\nabla \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \nabla s_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}_{\boldsymbol{y}} \nabla s_{\boldsymbol{y}}.$$

In particular, for linear models,

$$\nabla_{\boldsymbol{w}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \mathbb{E}\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

Logistic Regression

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s_{\boldsymbol{y}}}{\sum'_{\boldsymbol{y}} \exp s_{\boldsymbol{y}'}}$$

To learn, we maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = s_{\boldsymbol{y}^{\text{true}}} - \log \sum_{\boldsymbol{y}} \exp s_{\boldsymbol{y}}$$

with gradient descent, noting that

$$\nabla \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \nabla s_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}_{\boldsymbol{y}} \nabla s_{\boldsymbol{y}}.$$

In particular, for linear models,

$$\nabla_{\boldsymbol{w}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \mathbb{E}\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

For neural nets, grad wrt. score vector

$$\begin{aligned}\nabla_{\boldsymbol{s}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) &= \boldsymbol{e}_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}\boldsymbol{e}_{\boldsymbol{y}} \\ &= \boldsymbol{e}_{\boldsymbol{y}^{\text{true}}} - \mathrm{softmax}(\boldsymbol{s})\end{aligned}$$

Logistic Regression

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s_{\boldsymbol{y}}}{\sum'_{\boldsymbol{y}} \exp s_{\boldsymbol{y}'}}$$

To learn, we maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = s_{\boldsymbol{y}^{\text{true}}} - \log \sum_{\boldsymbol{y}} \exp s_{\boldsymbol{y}}$$

with gradient descent, noting that

$$\nabla \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \nabla s_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}_{\boldsymbol{y}} \nabla s_{\boldsymbol{y}}.$$

In particular, for linear models,

$$\nabla_{\boldsymbol{w}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \mathbb{E}\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

For neural nets, grad wrt. score vector

$$\begin{aligned}\nabla_{\boldsymbol{s}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) &= \boldsymbol{e}_{\boldsymbol{y}^{\text{true}}} - \mathbb{E}\boldsymbol{e}_{\boldsymbol{y}} \\ &= \boldsymbol{e}_{\boldsymbol{y}^{\text{true}}} - \mathrm{softmax}(\boldsymbol{s})\end{aligned}$$

For sequence models, **Forward-Backward computes what we need!**

Conditional Random Fields

Discriminative (non-generative) structured models.

$$\log \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) = f(\boldsymbol{y}; \boldsymbol{x}) - \log \sum_{\boldsymbol{y}'} \exp f(\boldsymbol{y}'; \boldsymbol{x})$$

Given a **decomposable** sequence scorer,

$$f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1}, y_i, i}$$

Forward-Backward computes

- Normalizer / log-partition function $\log \sum_{\boldsymbol{y}'} \exp f(\boldsymbol{y}'; \boldsymbol{x}) = \log Z$.
- Unary and transition posterior marginals:
  $\log \mathrm{P}(y_i = y|\boldsymbol{x})$, $\log \mathrm{P}(y_i = y, y_{i+1} = y'|\boldsymbol{x})$.

Score-based Forward-Backward

input: Unary scores $s^{(u)}$ ($|\Sigma| \times L$ array)
Transition scores $s^{(t)}$ ($|\Sigma| \times |\Sigma| \times (L-1)$ array)

Forward: compute $\widetilde{\alpha}$ recursively
$\widetilde{\alpha}_1(y) = s^{(u)}_{y,1} \qquad \forall y \in \Sigma$
for $i = 2$ **to** L **do**
  for $y \in \Sigma$ **do**
    $\widetilde{\alpha}_i(y) = s^{(u)}_{y,i} + \log \sum_{y'} \exp\left(s^{(t)}_{y',y,i} + \widetilde{\alpha}_{i-1}(y')\right)$

Backward: compute $\widetilde{\beta}$ recursively
$\widetilde{\beta}_L(y) = 0$
for $i = L-1$ **down to** 1 **do**
  for $y \in \Sigma$ **do**
    $\widetilde{\beta}_i(y) = \log \sum_{y'} \exp\left(s^{(t)}_{y,y',i} + s^{(u)}_{y',i+1} + \widetilde{\beta}_{i+1}(y')\right)$

output: The log-partition function $\log Z = \log \sum_{y'} \exp\left(\widetilde{\alpha}_i(y') + \widetilde{\beta}_i(y')\right)$ for any i;
Posterior unigram and transition marginal log-probas:
$\log \mathrm{P}(y_i = y|\boldsymbol{x}) = \widetilde{\alpha}_i(y) + \widetilde{\beta}_i(y) - \log Z$;
$\log \mathrm{P}(y_i = y, y_{i+1} = y'|\boldsymbol{x}) = \widetilde{\alpha}_i(y) + \widetilde{\beta}_{i+1}(y') + s^{(u)}_{y',i+1} + s^{(t)}_{y,y',i+1} - \log Z$.

Training a linear model CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \log \sum_{\boldsymbol{y}} \exp \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

where

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) = \text{cat}\left[\sum_i \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i),\ \sum_i \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y_i, y_{i+1})\right]$$

$$\boldsymbol{w} = \text{cat}\left[\boldsymbol{w}^{(u)},\ \boldsymbol{w}^{(t)}\right].$$

Training a linear model CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \log \sum_{\boldsymbol{y}} \exp \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

where

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) = \text{cat}\left[\sum_i \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i),\ \sum_i \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y_i, y_{i+1})\right]$$

$$\boldsymbol{w} = \text{cat}\left[\boldsymbol{w}^{(u)},\ \boldsymbol{w}^{(t)}\right].$$

Similar to logistic regression, we get

$$\begin{aligned}\nabla_{\boldsymbol{w}^{(u)}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) &= \sum_i \left(\boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i^{\text{true}}) - \mathbb{E}_{\boldsymbol{y}} \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i)\right) \\ &= \sum_i \left(\boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i^{\text{true}}) - \sum_y \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y)\, \mathrm{P}(y_i = y|\boldsymbol{x})\right)\end{aligned}$$

Training a linear model CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}}) - \log \sum_{\boldsymbol{y}} \exp \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$$

where

$$\boldsymbol{\phi}(\boldsymbol{y}, \boldsymbol{x}) = \text{cat}\left[\textstyle\sum_i \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i),\ \sum_i \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y_i, y_{i+1})\right]$$

$$\boldsymbol{w} = \text{cat}\left[\boldsymbol{w}^{(u)},\ \boldsymbol{w}^{(t)}\right].$$

Similar to logistic regression, we get

$$\begin{aligned}\nabla_{\boldsymbol{w}^{(u)}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) &= \textstyle\sum_i \Big(\boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i^{\text{true}}) - \mathbb{E}_{\boldsymbol{y}} \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i)\Big) \\ &= \textstyle\sum_i \Big(\boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y_i^{\text{true}}) - \sum_y \boldsymbol{\phi}^{(u)}((\boldsymbol{x}, i), y)\, \mathrm{P}(y_i = y|\boldsymbol{x})\Big)\end{aligned}$$

and, for transitions,

$$\begin{aligned}\nabla_{\boldsymbol{w}^{(t)}} \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = \textstyle\sum_i \Big(&\boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y_{i-1}^{\text{true}}, y_i^{\text{true}}) \\ &- \textstyle\sum_{y,y'} \boldsymbol{\phi}^{(t)}((\boldsymbol{x}, i), y, y')\, \mathrm{P}(y_{i-1} = y, y_i = y'|\boldsymbol{x})\Big)\end{aligned}$$

Training a neural CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y}) - \log \sum_{\boldsymbol{y}} \exp f(\boldsymbol{x}, \boldsymbol{y})$$

where

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1}, y_i, i}.$$

Training a neural CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y}) - \log \sum_{\boldsymbol{y}} \exp f(\boldsymbol{x}, \boldsymbol{y})$$

where

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1}, y_i, i}.$$

Note that

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{y})}{\partial s^{(u)}_{y,i}} = \begin{cases} 1, & y_i = y \\ 0, & \text{otherwise} \end{cases} = [[y_i = y]], \qquad \frac{\partial f(\boldsymbol{x}, \boldsymbol{y})}{\partial s^{(t)}_{y,y',i}} = [[y_{i-1} = y, y_i = y']]$$

Training a neural CRF

We want to maximize

$$\log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y}) - \log \sum_{\boldsymbol{y}} \exp f(\boldsymbol{x}, \boldsymbol{y})$$

where

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} s^{(t)}_{y_{i-1}, y_i, i}.$$

Note that

$$\frac{\partial f(\boldsymbol{x}, \boldsymbol{y})}{\partial s^{(u)}_{y,i}} = \begin{cases} 1, & y_i = y \\ 0, & \text{otherwise} \end{cases} = [[y_i = y]], \qquad \frac{\partial f(\boldsymbol{x}, \boldsymbol{y})}{\partial s^{(t)}_{y,y',i}} = [[y_{i-1} = y, y_i = y']]$$

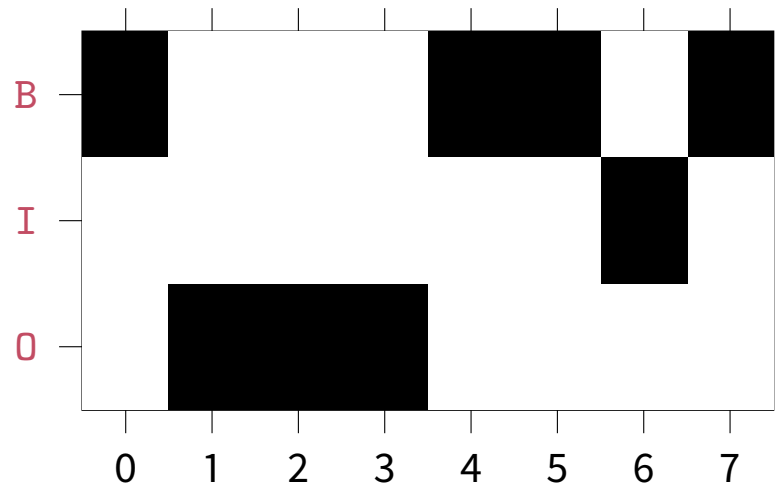
Therefore,

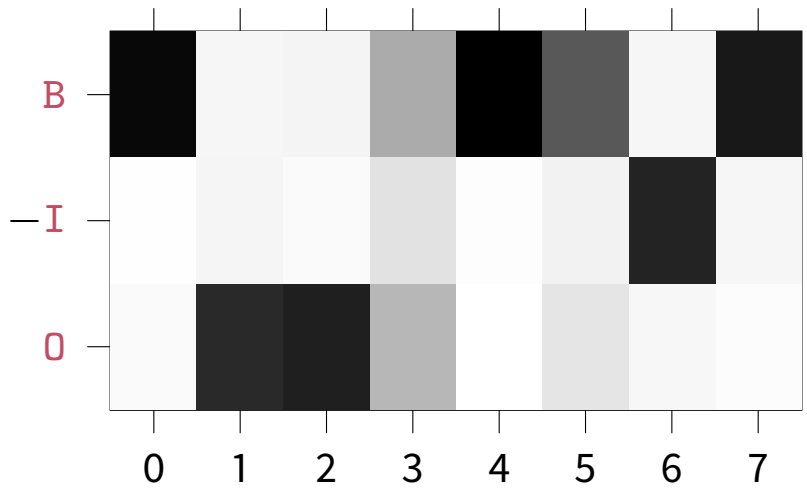
$$\frac{\partial \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x})}{\partial s^{(u)}_{y,i}} = [[y_i^{\text{true}} = y]] - \mathrm{P}(y_i = y|\boldsymbol{x})$$

$$\frac{\partial \log \mathrm{P}(\boldsymbol{y}^{\text{true}}|\boldsymbol{x})}{\partial s^{(t)}_{y,y',i}} = [[y_{i-1}^{\text{true}} = y, y_i^{\text{true}} = y']] - \mathrm{P}(y_{i-1} = y, y_i = y'|\boldsymbol{x})$$

Visually:

$$\nabla_{s(u)} \log \mathrm{P}(\boldsymbol{y}^{\text{true}} | \boldsymbol{x}) =$$





Outline

SVM and the hinge loss

SVMs are non-probabilistic max-margin classifiers.

The soft-margin perspective: minimizing a "hinge loss":

$$L(\boldsymbol{w}; (\boldsymbol{x}, \boldsymbol{y}^{\text{true}})) = \max_{\boldsymbol{y}} \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) + [[\boldsymbol{y} \neq \boldsymbol{y}^{\text{true}}]] - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$$

Intuition: the score of the correct class must be greater than the score of wrong classes by at least 1.

Structured SVM

$$L(\boldsymbol{w}; (\boldsymbol{x}, \boldsymbol{y}^{\text{true}})) = \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) + \text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}}) - f(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$$

where $\text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}}) = \sum_{i=1}^{L} [[y_i \neq y_i^{\text{true}}]]$ is the Hamming cost.

Structured SVM

$$L(\boldsymbol{w}; (\boldsymbol{x}, \boldsymbol{y}^{\text{true}})) = \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) + \text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}}) - f(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$$

where $\text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}}) = \sum_{i=1}^{L} [[y_i \neq y_i^{\text{true}}]]$ is the Hamming cost.

Cost-augmented decoding:

Finding $\arg\max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) + \text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}})$ can be done by Viterbi with adjusted unary scores

$$\tilde{s}_{y,i}^{(u)} = \begin{cases} s_{y,i}^{(u)} & y_i = y_i^{\text{true}} \\ s_{y,i}^{(u)} + 1 & y_i \neq y_i^{\text{true}} \end{cases}$$

Intuition: We give a boost to the wrong labels, and we want the global prediction to still be correct.

Similar algorithm to Structured Perceptron.

Structured SVM

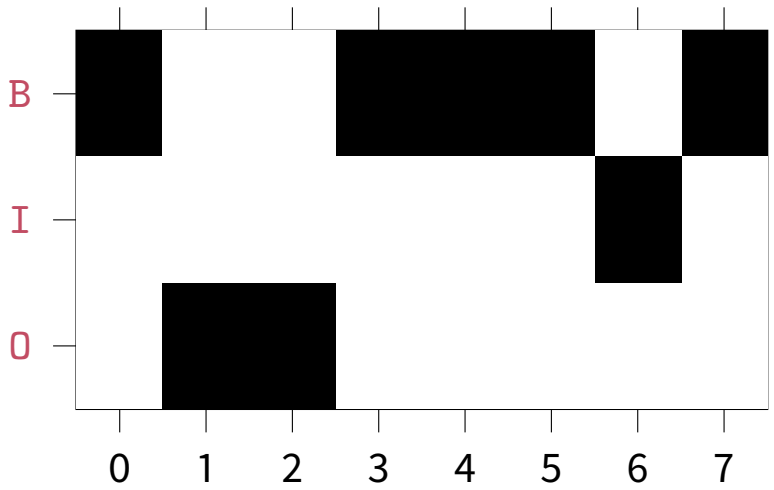
input: labeled data $\mathcal{D}$, learning rate η
initialize $\boldsymbol{w}$
repeat
  get new training example $(\boldsymbol{x}, \boldsymbol{y}^{\text{true}})$
  compute unary and transition scores, $s^{(u)}$, $s^{(t)}$
  predict $\widehat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}; \boldsymbol{x}) + \text{cost}(\boldsymbol{y}, \boldsymbol{y}^{\text{true}})$ **(cost-augmented Viterbi)**
  if $\widehat{\boldsymbol{y}} \neq \boldsymbol{y}$ **then**
    update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta\big(\nabla_{\boldsymbol{w}} f(\boldsymbol{y}^{\text{true}}; \boldsymbol{x}) - \nabla_{\boldsymbol{w}} f(\widehat{\boldsymbol{y}}; \boldsymbol{x})\big)$
until max. epochs
output: model weights $\boldsymbol{w}$

- If linear $f(\boldsymbol{y}; \boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$, then $\nabla_{\boldsymbol{w}} f(\boldsymbol{y}; \boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$.
- If neural, $\nabla_{\boldsymbol{w}} f(\boldsymbol{y}; \boldsymbol{x}) = \sum_{i=1}^{L} \nabla_{\boldsymbol{w}} s^{(u)}_{y_i, i} + \sum_{i=2}^{L} \nabla_{\boldsymbol{w}} s^{(t)}_{y_{i-1}, y_i, i}$ from autodiff.

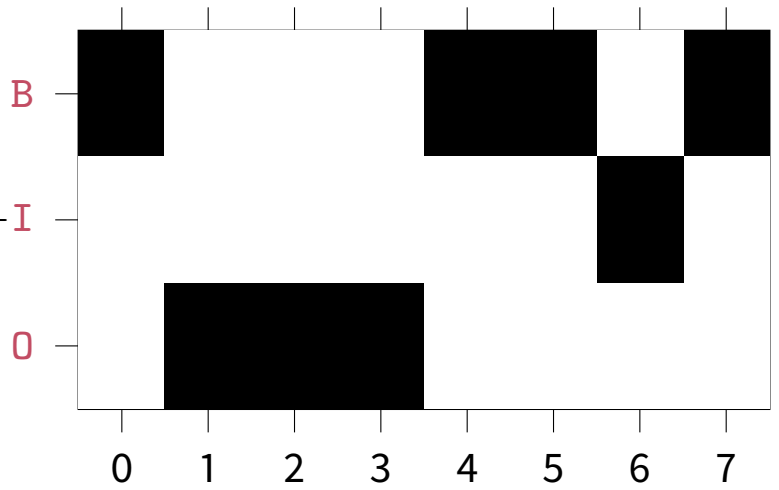
Visually:

Unlike CRF, subgradient updates for Perceptron/SSVM look like:

$$\partial_s^{(u)} L \ni$$



$-$



Structured prediction models

Binary/Multi-class	**Structured Prediction**
Naive Bayes	HMMs
Perceptron	Structured Perceptron
Logistic Regression	CRF
SVMs	Structured SVM

Discriminative vs HMM

- HMM is less expressive, but fast to train.
- Discriminative models are powerful, but require iterative training.
- HMM and CRF have probabilistic interpretations.
  (Useful when posterior analyses are desirable)
- Structured SVM is a good classifier,
  can be extended to other cost functions.
- Perceptron and Structured SVM updates are sparse;
  (may be faster for very large Σ.)
- For structures more complicated than sequences, we may not have a Forward-Backward equivalent, but we may be able to approximate $\arg\max_{\boldsymbol{y}} f(\boldsymbol{y}; \boldsymbol{x})$.