Learning Translation models

Machine Translation

$$p(English|Chinese) \sim$$

$$p(English) \times p(Chinese|English)$$

Learning an *n*-gram language model

n-gram language model

Given English word sequence $e = e_1 ... e_{|e|}$

$$
\begin{aligned}
P(e) &= \prod_{i=1}^{|e|} P(e_i | e_1, ..., e_{i-1}) \\
&\approx \prod_{i=1}^{|e|} P(e_i | e_{i-n+1}, ..., e_{i-1})
\end{aligned}
$$

Question: why approximate?

Language Models

Assume a bigram language model.

Parameters are of this form:

$$P(the|\langle START\rangle)$$

$$P(sky|the)$$

$$P(remained|sky)$$

$$P(clear|remained)$$

$$P(\langle STOP\rangle|clear)$$

Each parameter is a number in [0,1], s.t.

$$\sum_{w \in \Sigma \cup \{\langle STOP\rangle\}} P(w|v) \forall v \in \Sigma$$

Question: where do these numbers come from?

Language Models

This is just a model that we can train on data.

... in the night <u>sky</u> as it orbits earth ...

... said that the <u>sky</u> would fall if ...

... falling dollar , <u>sky</u> high interest rates ...

However , the <u>sky</u> remained clear ...

$$p(remained|sky) = ???$$

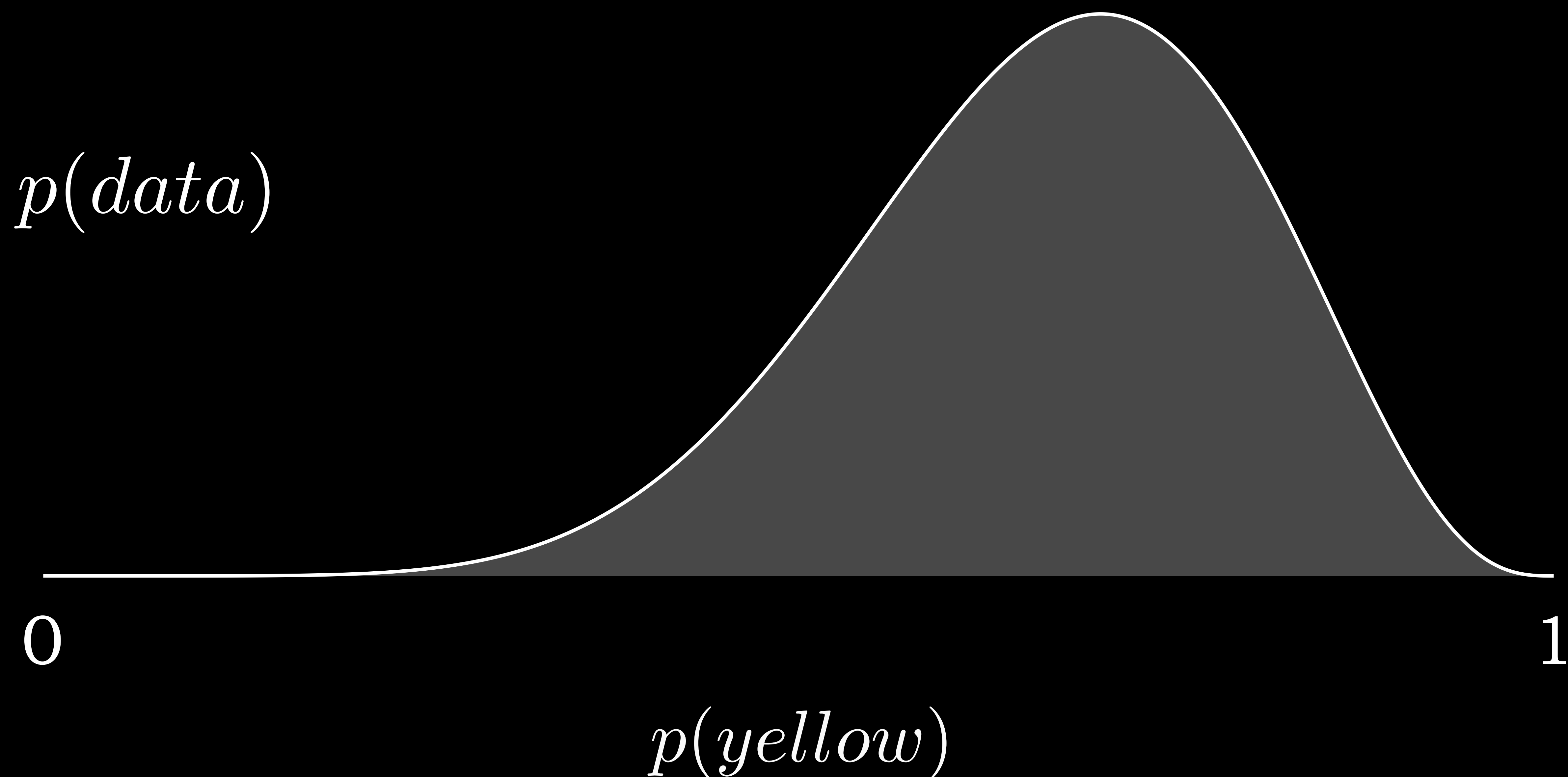
$p(yellow)$ $\quad$ $1 - p(yellow)$

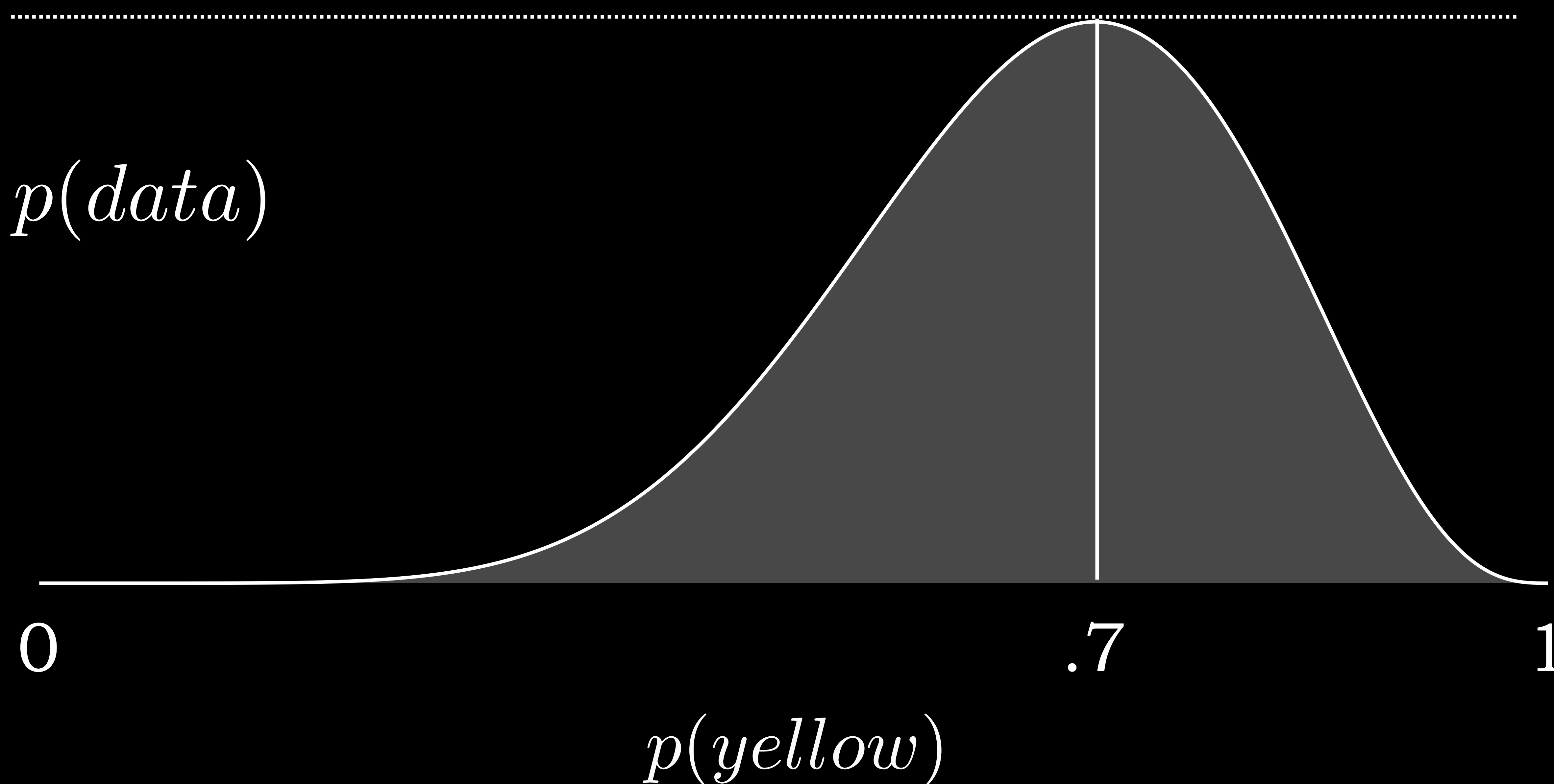
$p(yellow)$?

$p(yellow)$?

 $p(yellow)?$

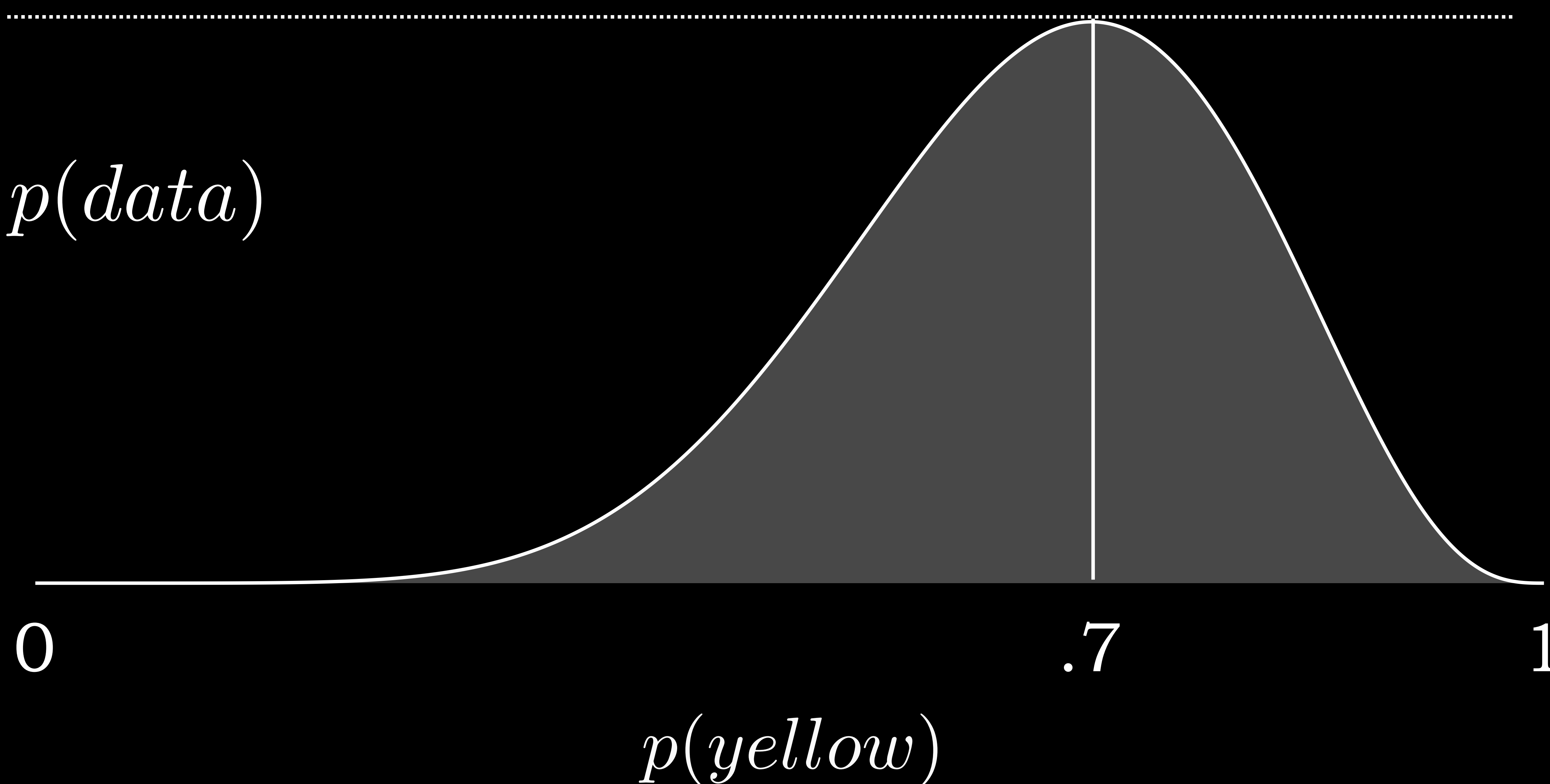
$$p(data) = p(yellow)^7 \times [1 - p(yellow)]^3$$


$p(data)$
0
1
$p(yellow)$



$p(data)$
0
.7
1
$p(yellow)$

(we'll derive this more formally shortly)




THE WØRD


THE WØRD
• Optimization

Language Models

$$p(remained|sky) = \frac{\text{\# of times I saw ``sky remained''}}{\text{\# of times I saw ``sky''}}$$

Language Models

This is a pretty old trick.

Language Models

This is a pretty old trick.

But what if we haven't seen some word sequences?

Language Models

This is a pretty old trick.

But what if we haven't seen some word sequences?

Classical approach: *smoothing*.

We will discuss smoothing until later, in a different form.

Machine Translation

$$p(English|Chinese) \sim$$

$$p(English) \times p(Chinese|English)$$

translation model

Machine Translation

$$p(English|Chinese) \sim$$

$$p(English) \times p(Chinese|English)$$

translation model

This is just a **conditional** language model.
It generates Chinese, conditioned on English.

Machine Translation

$$p(English|Chinese) \sim$$

$$p(English) \times p(Chinese|English)$$

translation model

This is just a **conditional** language model.
It generates Chinese, conditioned on English.

Question: Could we use *n*-gram models here?

Conditional LMs

Given English word sequence $e = e_1 ... e_{|e|}$
and Chinese word sequence $f = f_1 ... f_{|f|}$

Let $w = e_1 ... e_{|e|} f_1 ... f_{|f|}$

Will this work?

$$
\begin{aligned}
P(w) &= \prod_{i=1}^{|w|} P(w_i | w_1, ..., w_{i-1}) \\
&\approx \prod_{i=1}^{|w|} P(w_i | w_{i-n+1}, ..., w_{i-1})
\end{aligned}
$$

Conditional LMs

Given English word sequence $e = e_1 ... e_{|e|}$
and Chinese word sequence $f = f_1 ... f_{|f|}$

What about this?

$$w = e_1 f_1 e_2 f_2 ... e_{|e|} f_{|e|} ... f_{|f|}$$

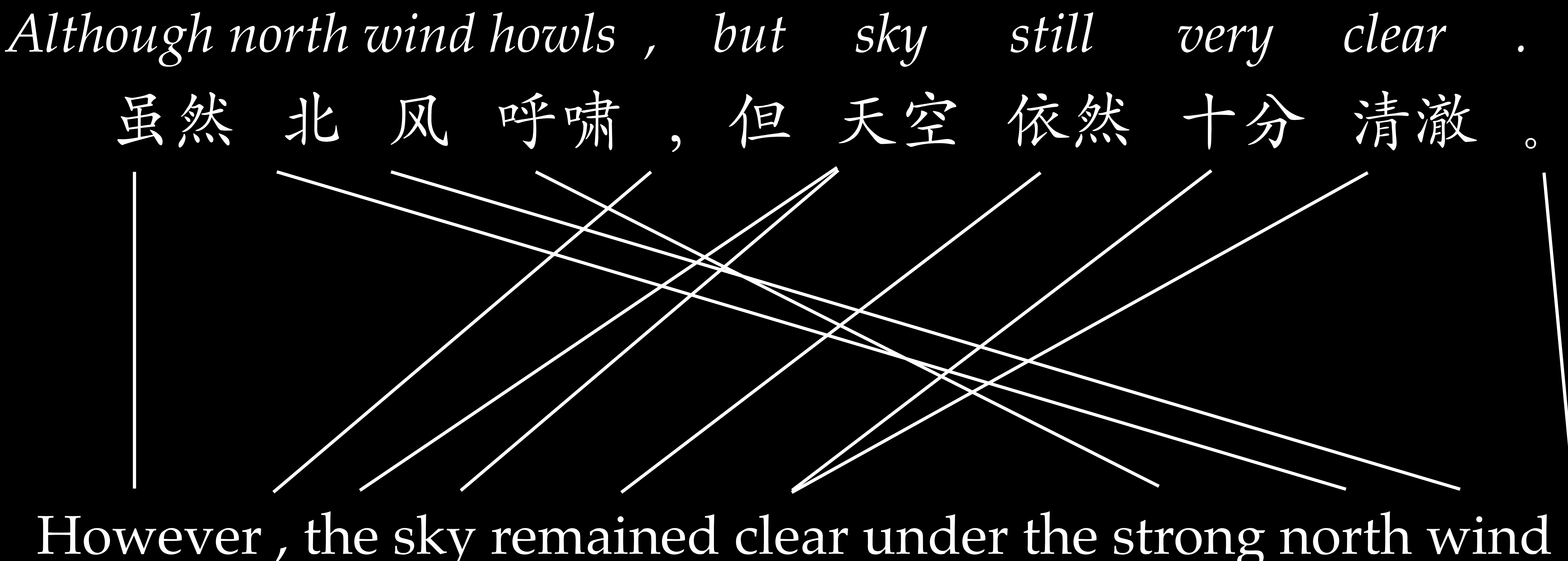
$$\begin{aligned} P(w) &= \prod_{i=1}^{|w|} P(w_i | w_1, ..., w_{i-1}) \\ &\approx \prod_{i=1}^{|w|} P(w_i | w_{i-n+1}, ..., w_{i-1}) \end{aligned}$$

Conditional LMs

Fundamental problem:

words are not in the same order!

How do we decide which words to condition on?



Conditional LMs

Word alignments!

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

Let's write a simple model in terms of word-to-word alignments:

$$p(\mathbf{f}, \mathbf{a} | \mathbf{e})$$

(Note that for this example, we have reversed the conditioning)

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

$$p(English\ length|Chinese\ length)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

$$p(Chinese\ word\ position)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However

$$p(English\ word | Chinese\ word)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

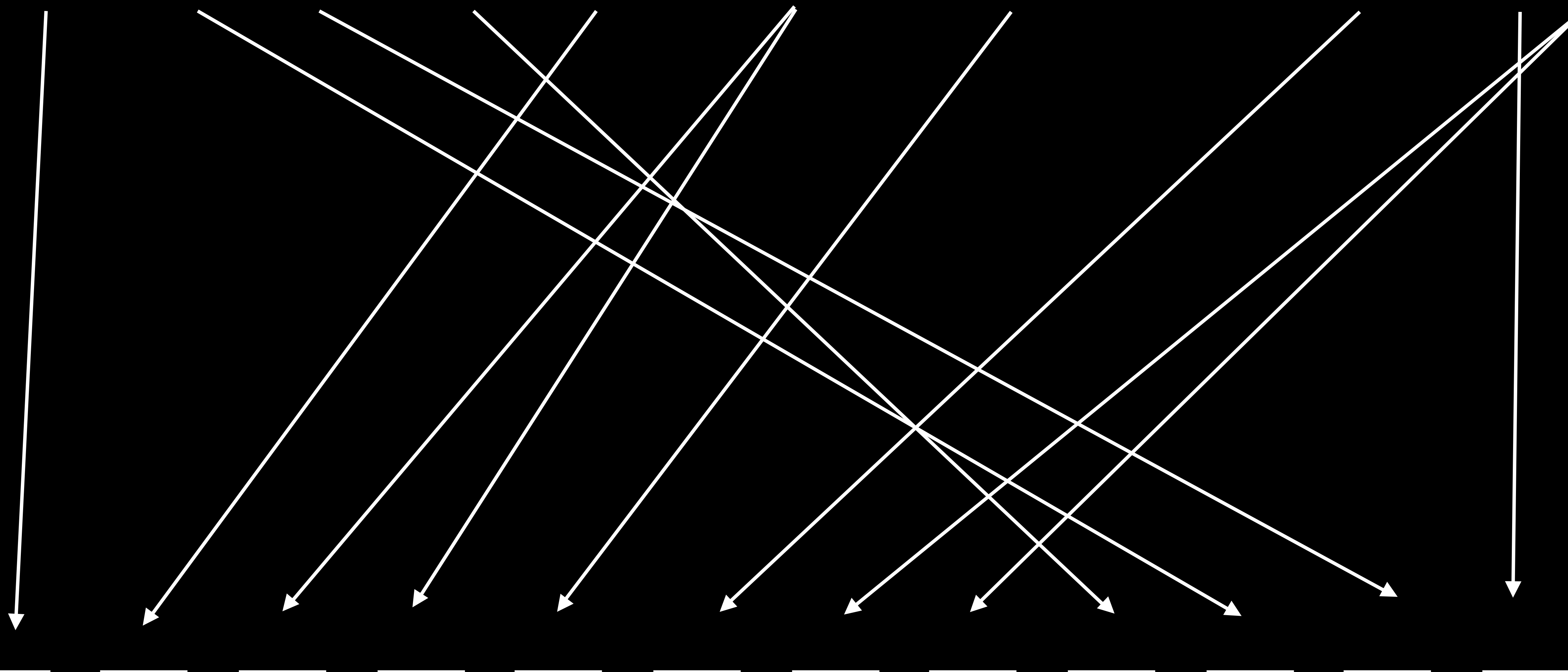
___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

However ,

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε



However , the sky remained clear under the strong north wind .

IBM Model 1

alignment of French
word at position i

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = p(I|J) \prod_{i=1}^{I} p(a_i|J) \cdot p(f_i|e_{a_i})$$

French, English
sentence lengths

French, English
word pair

IBM Model 1

$$\theta \left\{ \begin{array}{ll} p(despite|\text{虽然}) & ??? \\ p(however|\text{虽然}) & ??? \\ p(although|\text{虽然}) & ??? \\ \ldots & \\ p(northern|\text{北}) & ??? \\ p(north|\text{北}) & ??? \end{array} \right.$$

Where do these parameters come from?

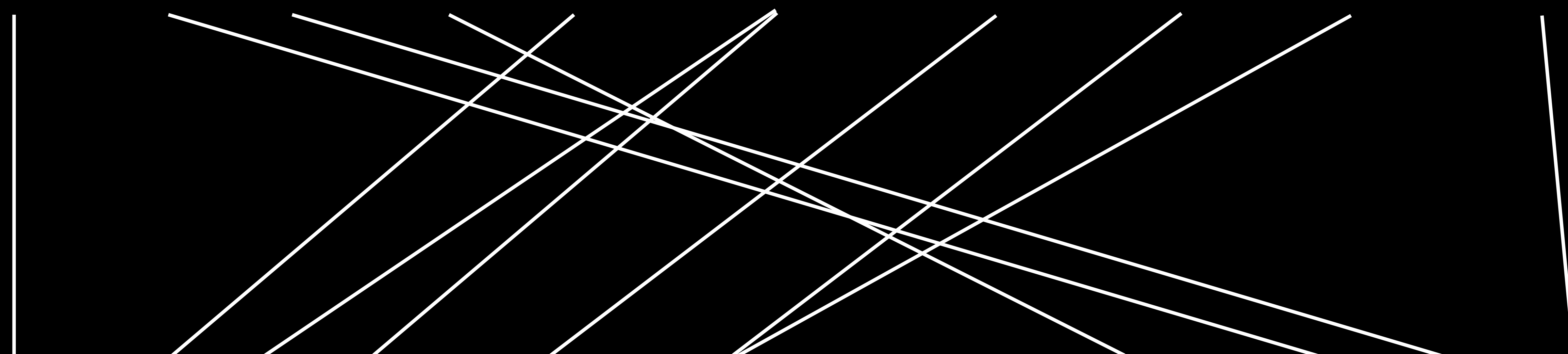
Conditional LMs

You may have noticed a small problem.

Our data does not contain word alignments.

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

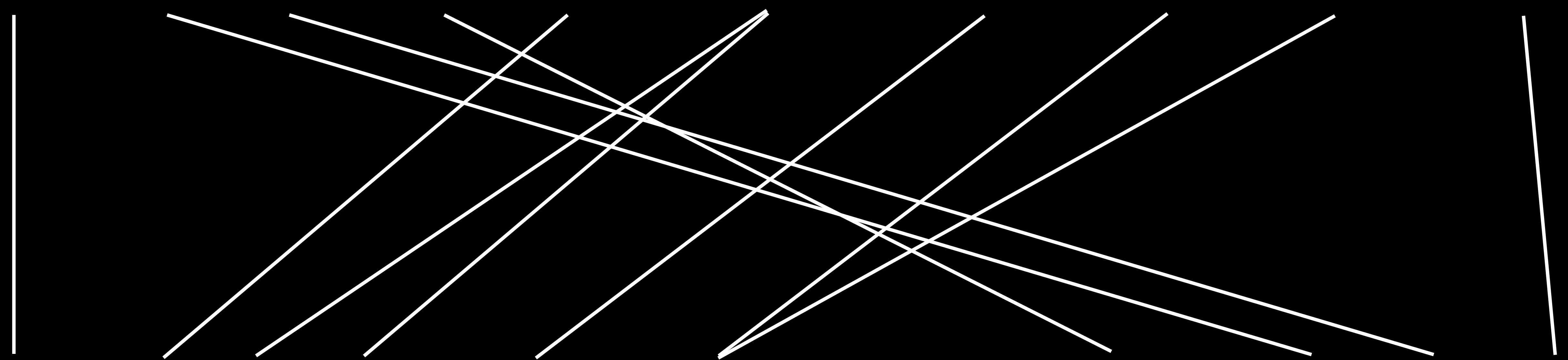


However , the sky remained clear under the strong north wind .

MLE for IBM Model 1 (observed)

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{f}, \mathbf{a} | \mathbf{e})$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)}) \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

number of sentences

alignment of French word at position *i*

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)}) \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

French, English sentence lengths

French, English word pair

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(\underbrace{p(I^{(n)}|J^{(n)})}_{\text{constant (w.r.t. words)!}} \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)}) \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)})$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

$$\log(a) < \log(b) \iff a < b$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{count(\langle f,e \rangle)} \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$

log of product = sum of logs

MLE for IBM Model 1 (observed)

$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e) \underbrace{- \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)}_{\text{Lagrange multiplier expresses normalization constraint}}$$

MLE for IBM Model 1 (observed)

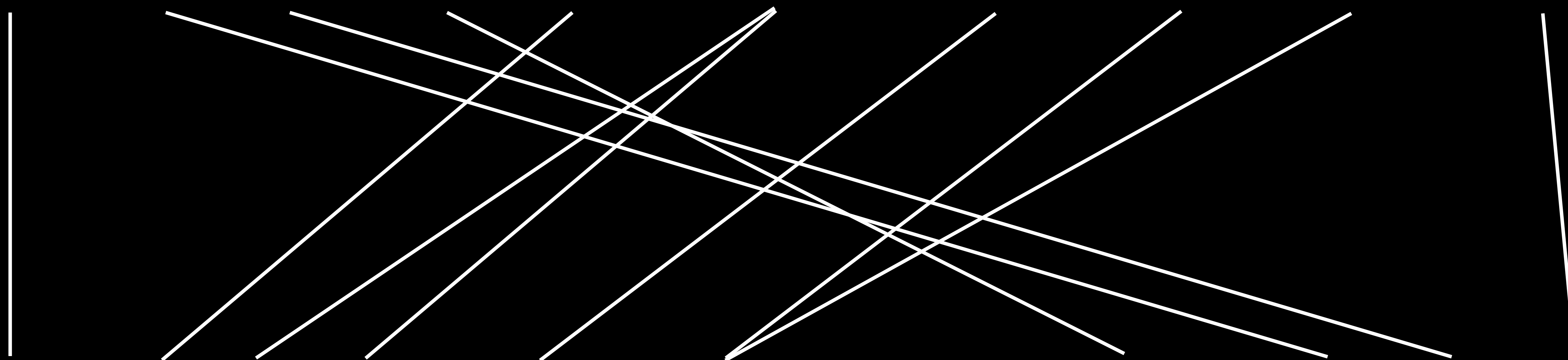
$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$
$$- \sum_e \lambda_e \left(\sum_f p(f|e) - 1 \right)$$

derivative $$\frac{\partial \Lambda(\theta, \lambda)}{\partial p(f|e)} = \frac{count(\langle f, e \rangle)}{p(f|e)} - \lambda_e$$

MLE for IBM Model 1 (observed)

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 aligns to any word}}$$

MLE for IBM Model 1 (unobserved)

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$$p(however|\text{虽然}) = \text{???}$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \sum_{a} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

Data likelihood = marginal probability of observed data

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

Not constant! Depends on parameters, no analytic solution.

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

Not constant! Depends on parameters, no analytic solution.

But it does strongly imply an iterative solution.

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

Parameters and alignments are both unknown.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。 ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。 ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

The Plan: Bootstrapping

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However , the sky remained clear under the strong north wind .

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

if we had observed the alignment, this line would either be here (count 1) or it wouldn’t (count 0).

However , the sky remained clear under the strong north wind .

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

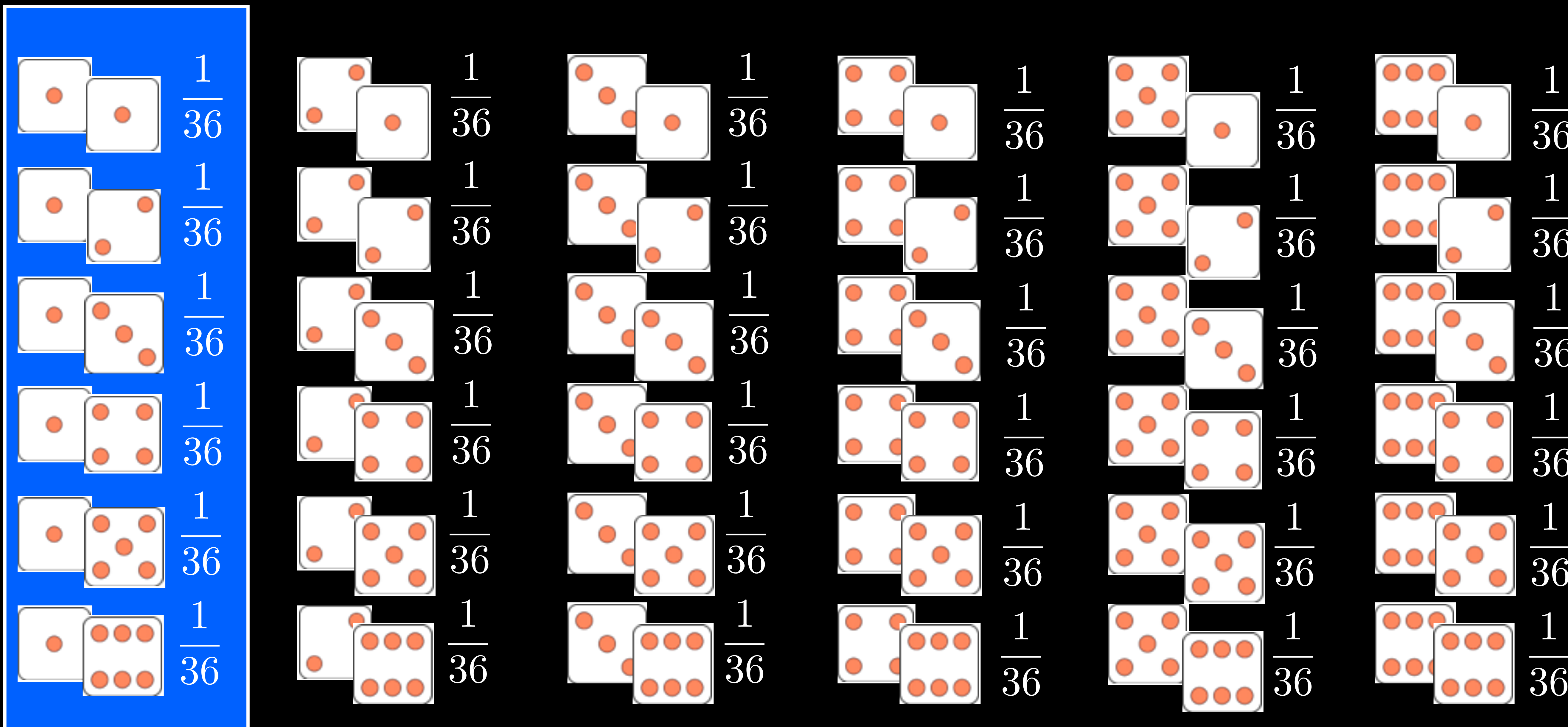
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

if we had observed the alignment, this line would either be here (count 1) or it wouldn't (count 0).

since we didn't observe the alignment, we calculate the probability that it's there.

However , the sky remained clear under the strong north wind .

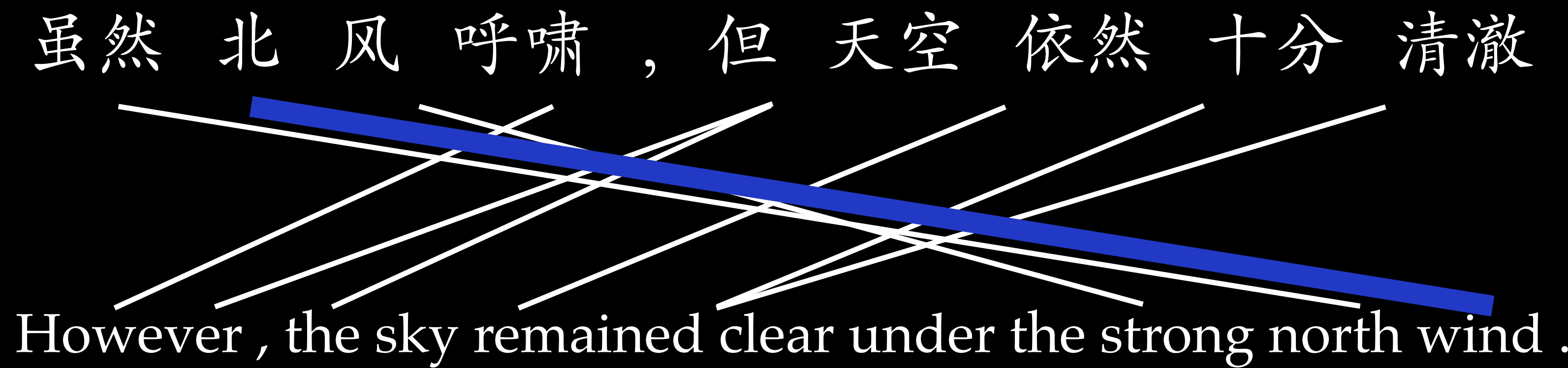
Reminder



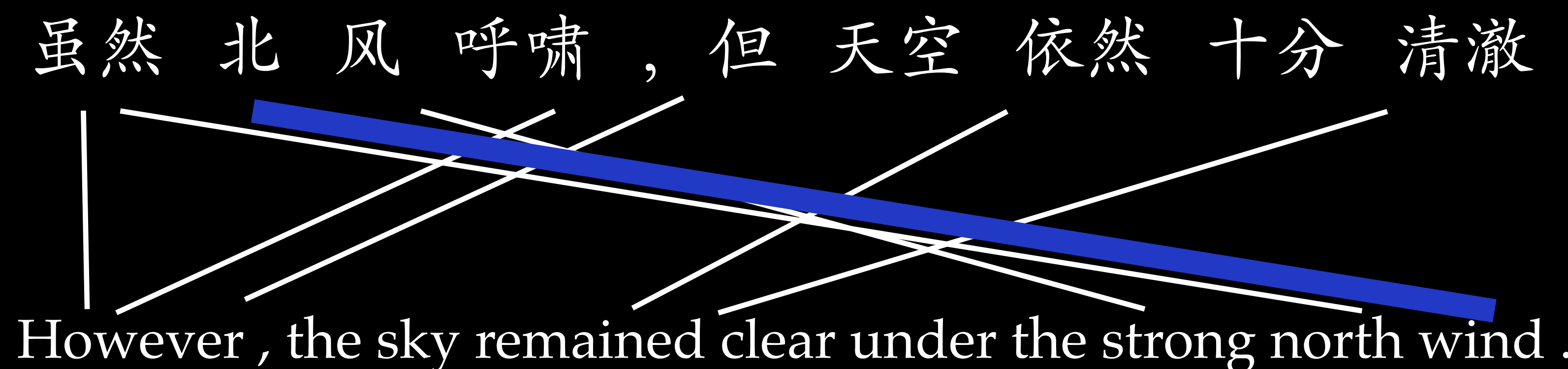
marginal probability

$$p(Y=1) = \sum_{x \in X} p(X=x, Y=1) = \frac{1}{6}$$

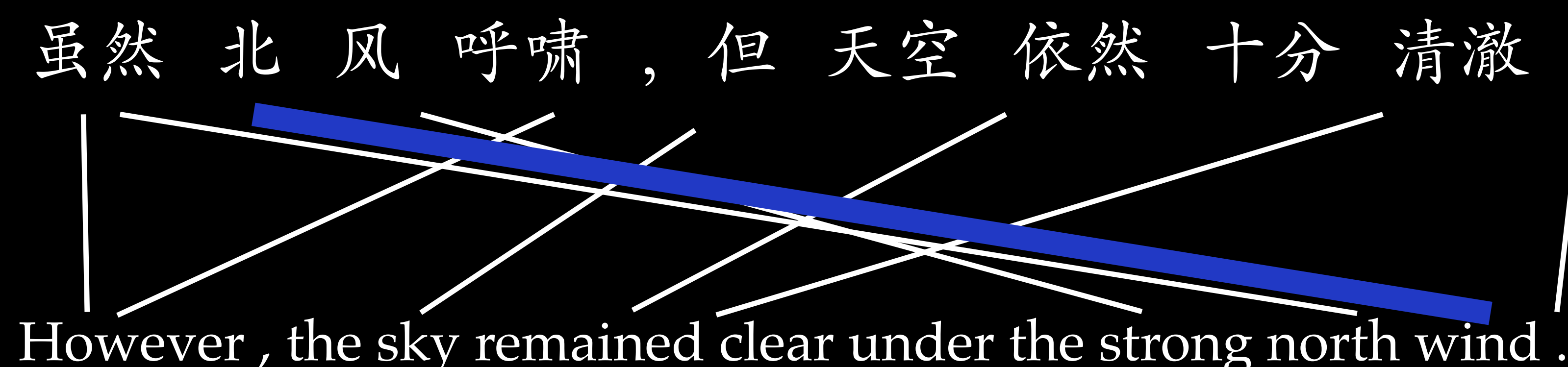
Marginalize: sum all alignments containing the link

p(

) +

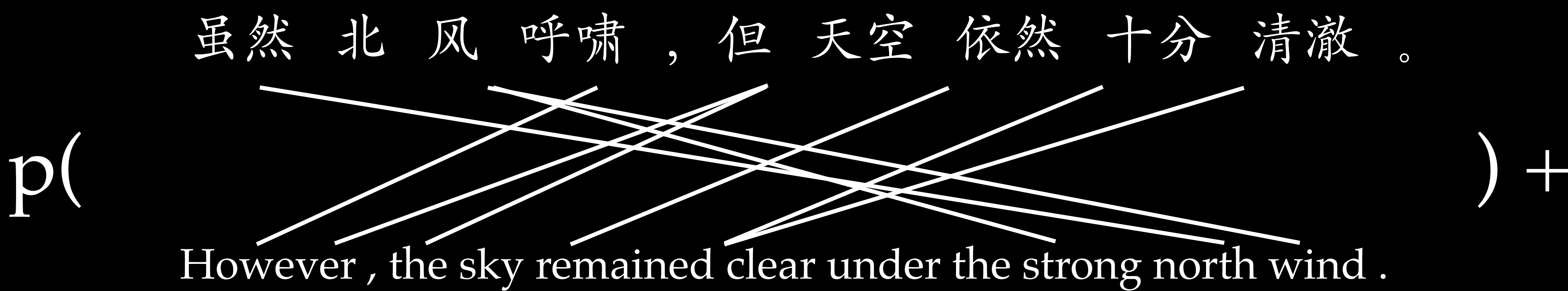
p(

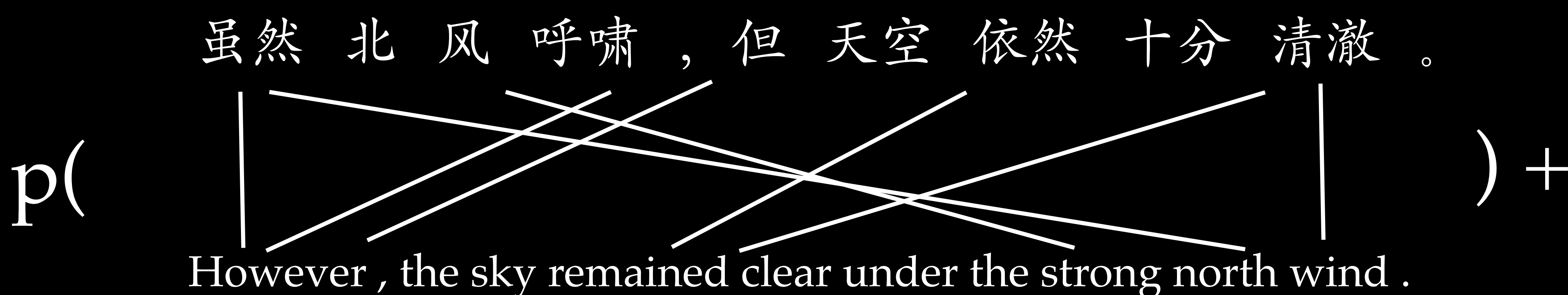
) +

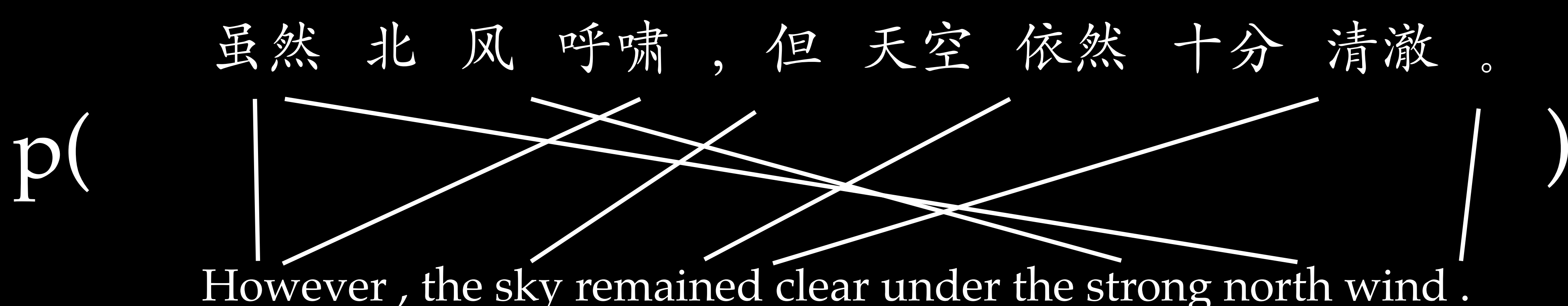
p(

)

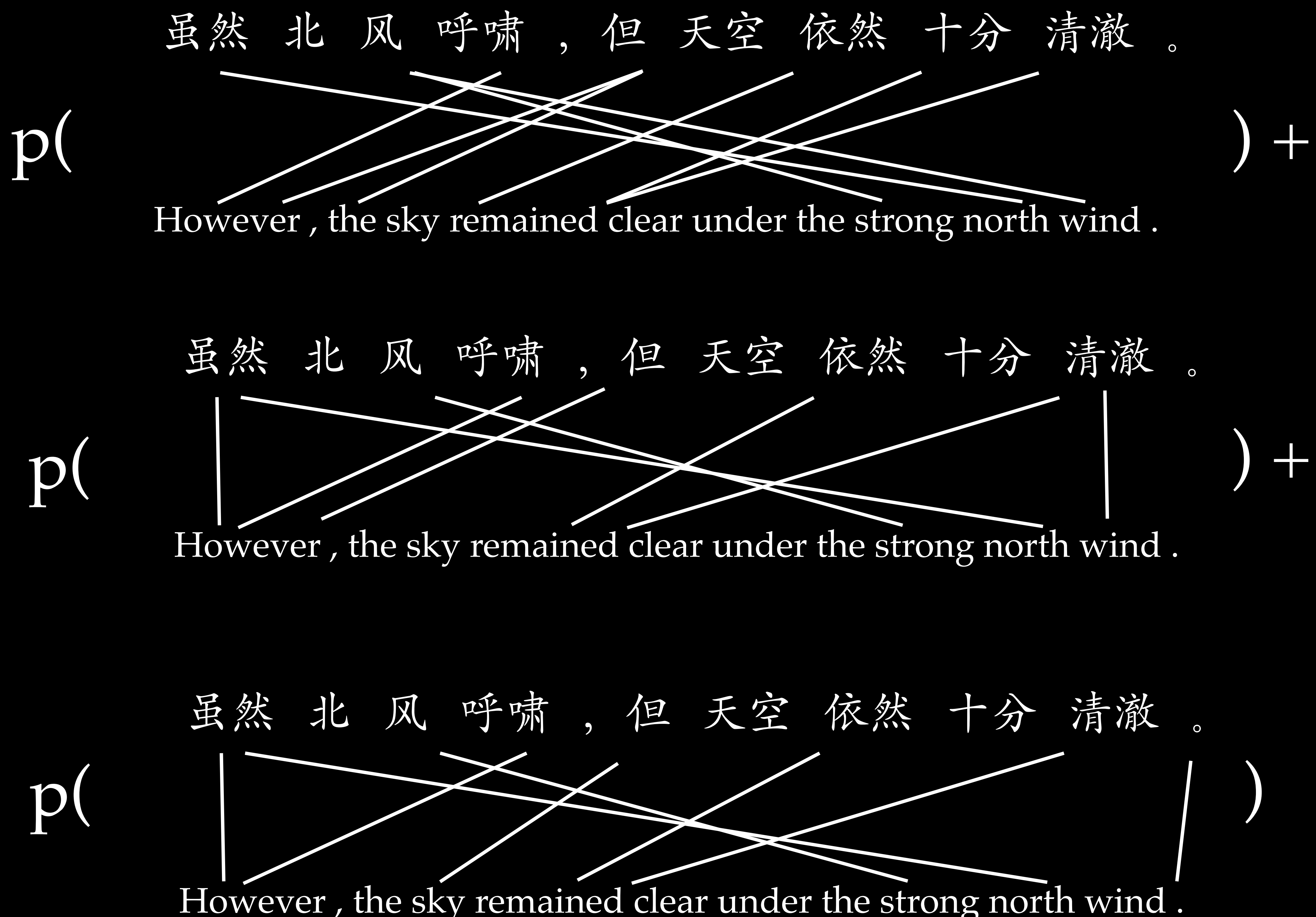
Divide by sum of all *possible* alignments







Divide by sum of all *possible* alignments



Is this hard? How many alignments are there?

Computing Expectations

probability of an alignment.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) p(f_i|e_j)$$

Computing Expectations

probability of an alignment.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) p(f_i|e_j)$$

observed

uniform

Computing Expectations

probability of an alignment.

factors across words.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) \boxed{p(f_i|e_j)}$$

observed

uniform

Computing Expectations

marginal probability of
alignments containing link

$$\sum_{a \in A: 北 \leftrightarrow north} p(north|北) \cdot p(rest\ of\ a)$$

Computing Expectations

marginal probability of
alignments containing link

$$p(north|北) \sum_{a \in A: 北 \leftrightarrow north} p(rest\ of\ a)$$

Computing Expectations

marginal probability of alignments containing link

$$\frac{p(north|北) \displaystyle\sum_{a \in A:北 \leftrightarrow north} p(rest\ of\ a)}{\displaystyle\sum_{c \in Chinese\ words} p(north|c) \sum_{a \in A:\,\mathrm{c} \leftrightarrow north} p(rest\ of\ a)}$$

marginal probability of all alignments

Computing Expectations

marginal probability of
alignments containing link

$$\frac{p(north|\text{北}) \sum_{a \in A: \text{北} \leftrightarrow north} p(rest\ of\ a)}{\sum_{c \in Chinese\ words} p(north|c) \sum_{a \in A:\, c \leftrightarrow north} p(rest\ of\ a)}$$

marginal probability of all
alignments

Computing Expectations

marginal probability of
alignments containing link

$$\frac{p(north|\text{北}) \sum_{a \in A: \text{北} \leftrightarrow north} p(rest\ of\ a)}{\sum_{c \in Chinese\ words} p(north|c) \sum_{a \in A: c \leftrightarrow north} p(rest\ of\ a)}$$

identical!

marginal probability of all
alignments

Computing Expectations

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Computing Expectations

marginal probability (expected count) of an alignment containing the link

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Computing Expectations

marginal probability (expected count) of an alignment containing the link

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

For each sentence, use this quantity instead of 0 or 1

Maximum Likelihood

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 aligns to any word}}$$

Expectation Maximization

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$$p(however | \text{虽然}) = \frac{\textit{Expected}\ \#\ \text{of times 虽然 aligns to However}}{\textit{Expected}\ \#\ \text{of times 虽然 aligns to any word}}$$

Expectation Maximization

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate. (Until when?)

Expectation Maximization

Why does this even work?

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

MLE: choose parameters that maximize this expression.

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

MLE: choose parameters that maximize this expression.

Minor problem: there is no analytic solution.

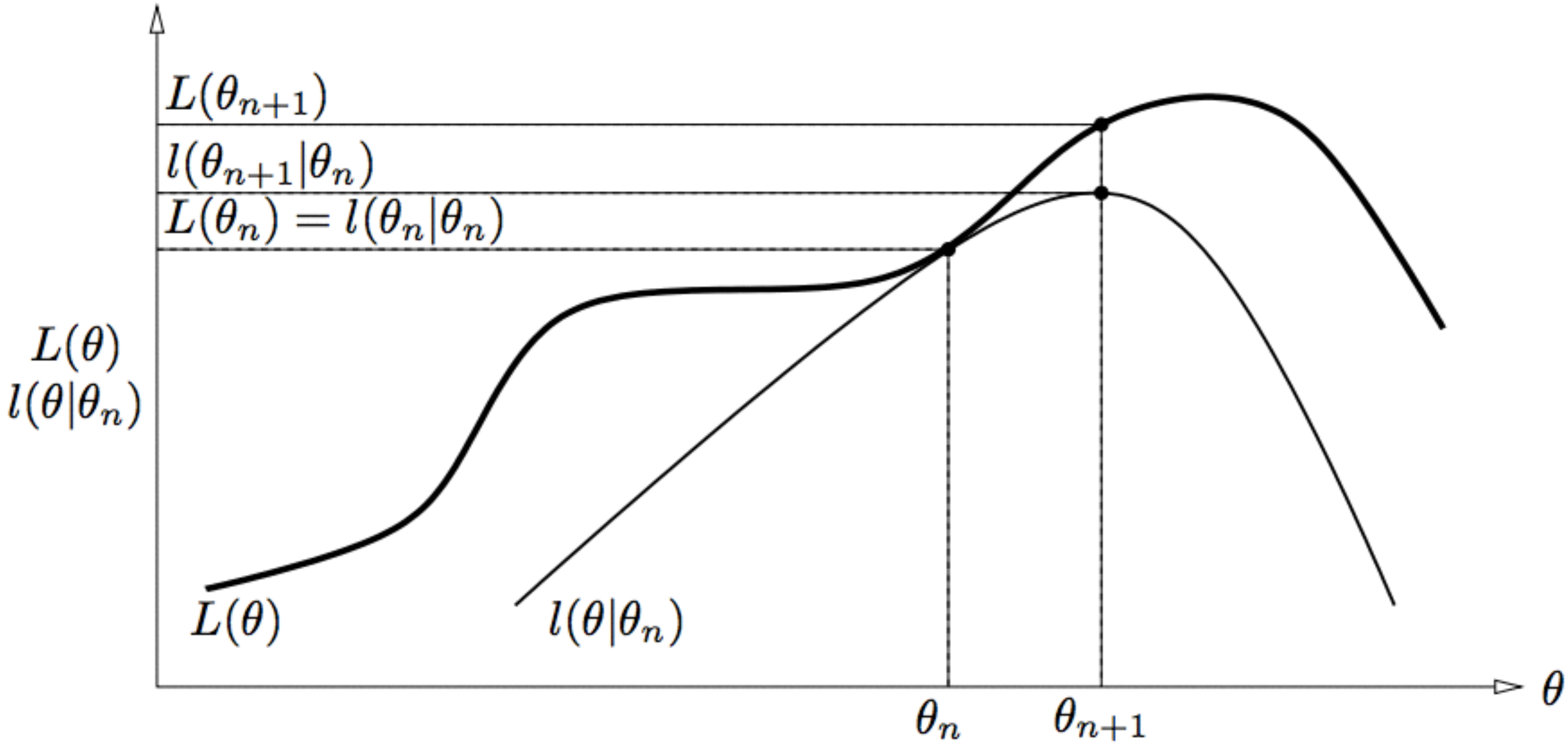


Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

(from Boorman ’04)

Summary

- *Learning is optimization*: choose parameters that optimize some function, such as likelihood.
- Supervised: *maximum likelihood*.
  - Beware of overfitting.
- Unsupervised: *expectation maximization*.
- Many other objective functions and algorithms.

Next week

- No in-person lectures. I will post video lectures.
- Lab 1 (Tuesday 10am, FH 1.B30): implement model 1 and extensions to it. Several demonstrators will be there to help.
- Coursework 1 release: use a learned translation model to translate new sentences.