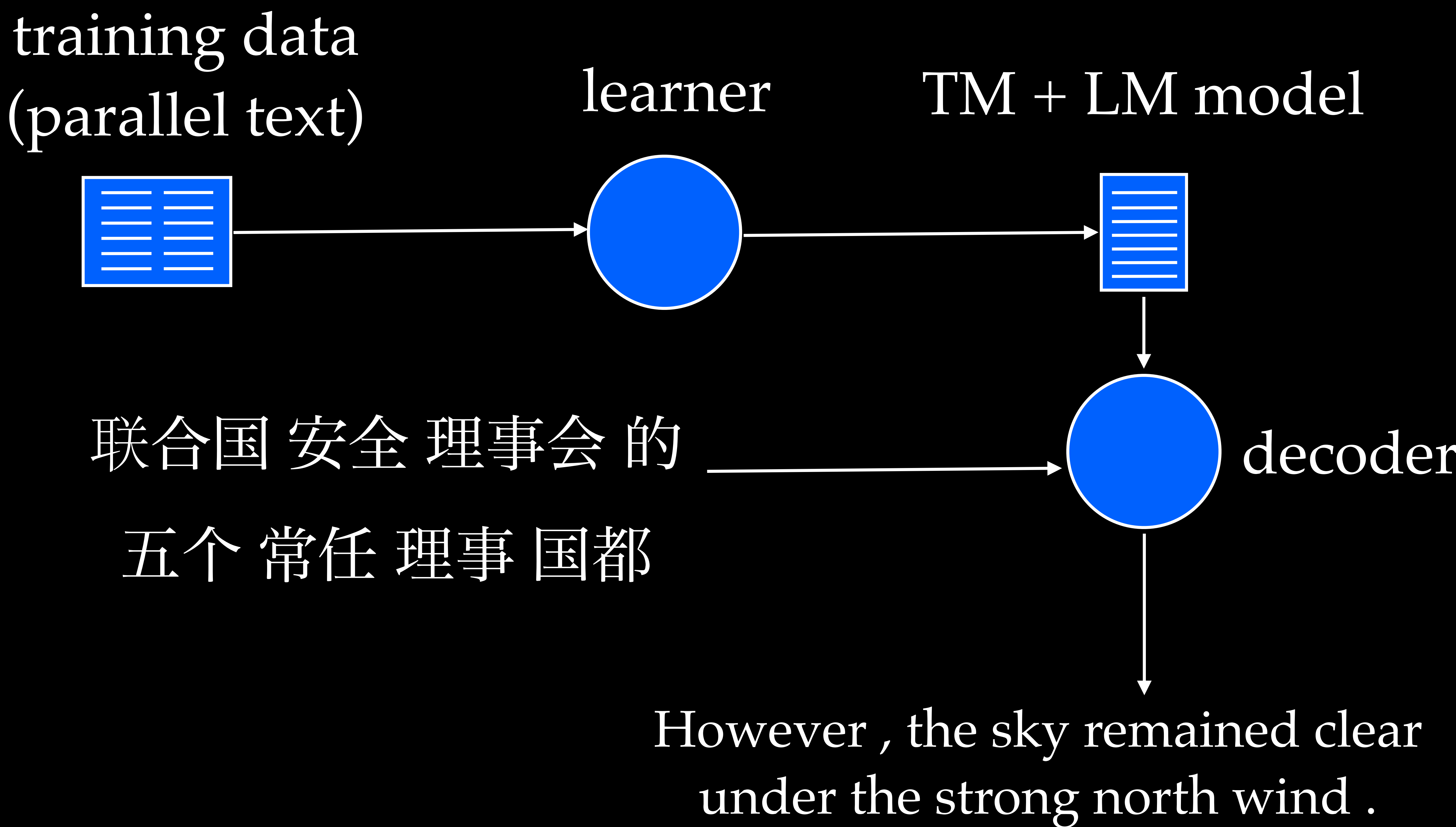
Word Alignment

Adam Lopez
Johns Hopkins

Quick Recap


training data
(parallel text)
learner
TM + LM model
decoder
联合国 安全 理事会 的
五个 常任 理事 国都
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 。

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

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

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

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

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

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

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

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

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

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

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

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 ,

IBM Model 1

Although north wind howls , but sky still very clear .

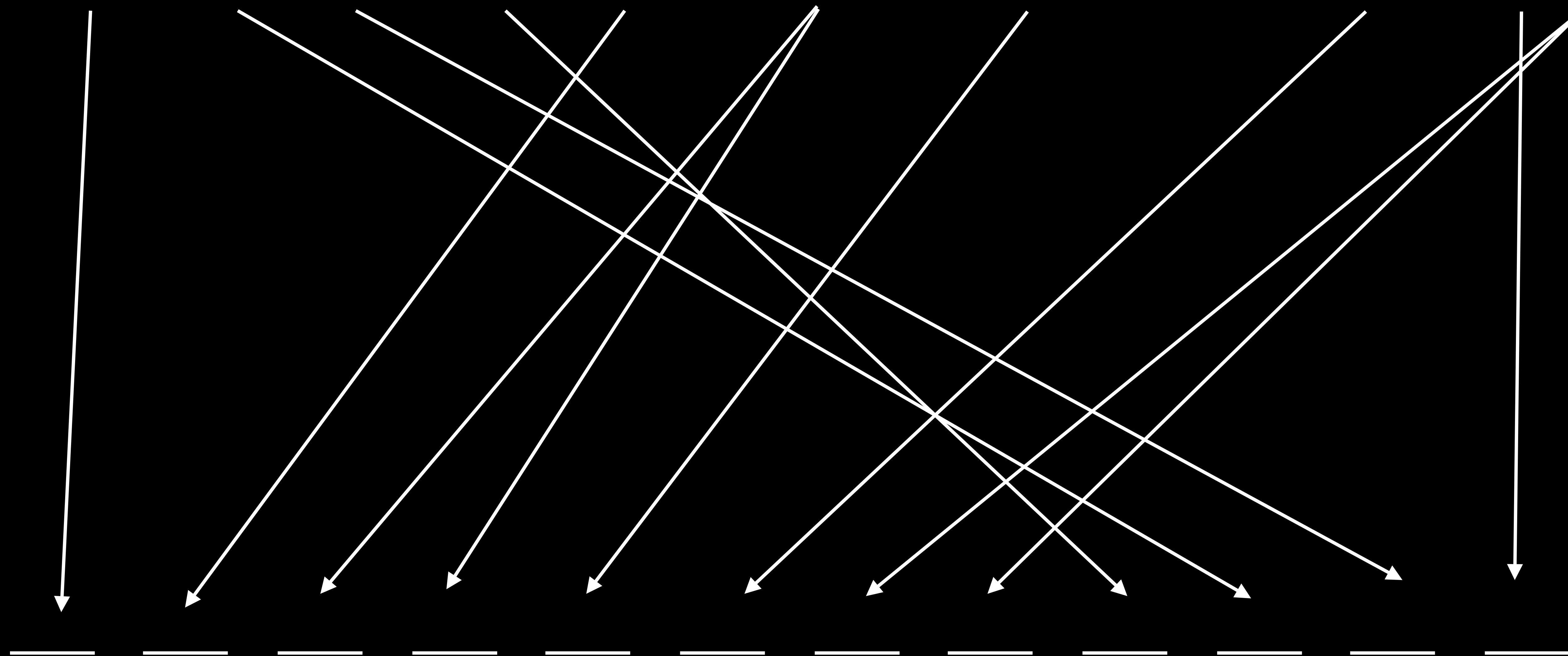
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However , the

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

$$p(despite|\text{虽然})$$

$$p(however|\text{虽然})$$

$$p(although|\text{虽然})$$

$$p(northern|\text{北})$$

$$p(north|\text{北})$$

IBM Model 1

$p(despite|\text{虽然})$???

$p(however|\text{虽然})$???

$p(although|\text{虽然})$???

$p(northern|\text{北})$???

$p(north|\text{北})$???

IBM Model 1

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


ČESKÁ REPUBLIKA
50 KČ

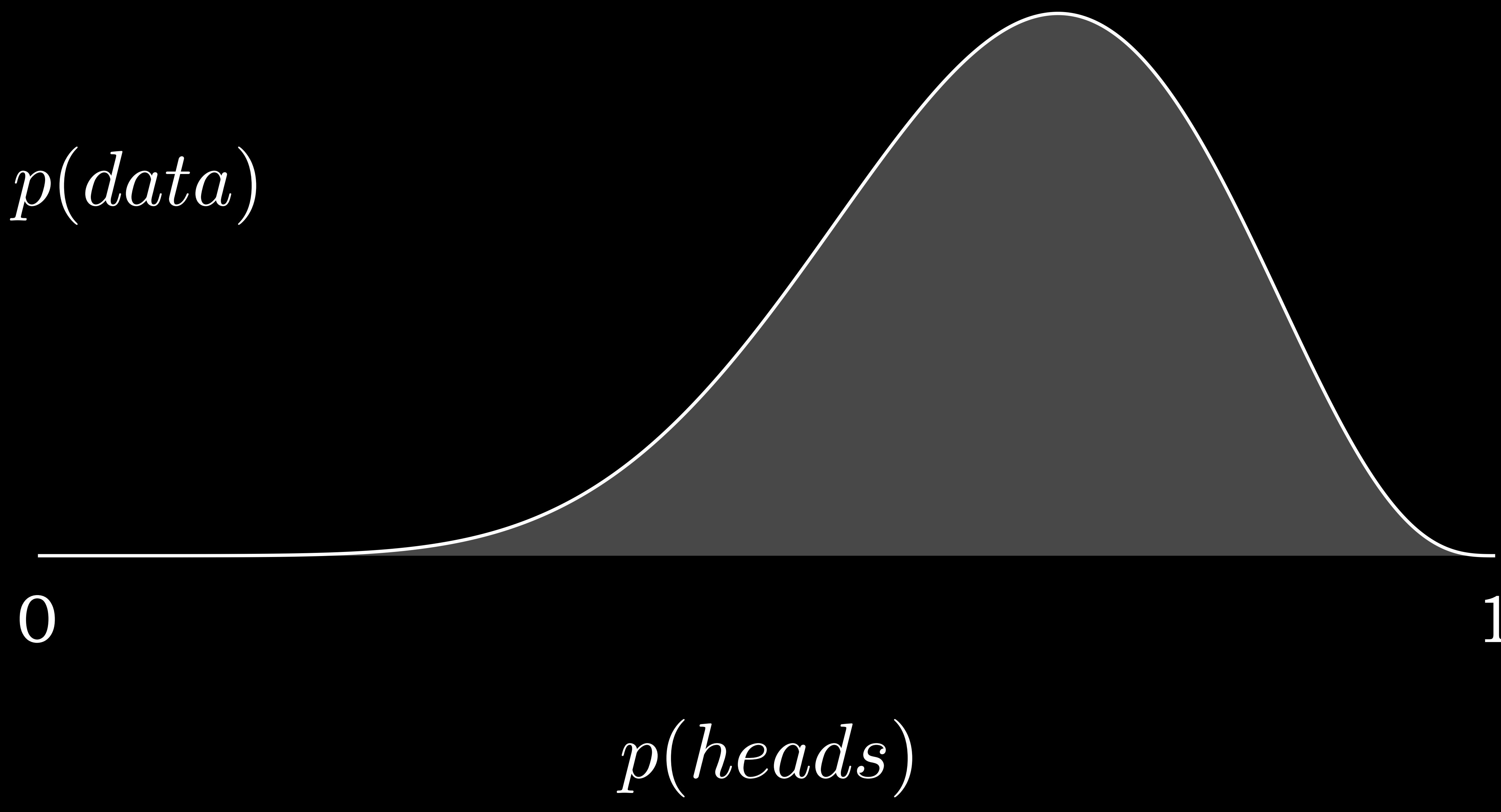

PRAGA · MATER · URBIUM

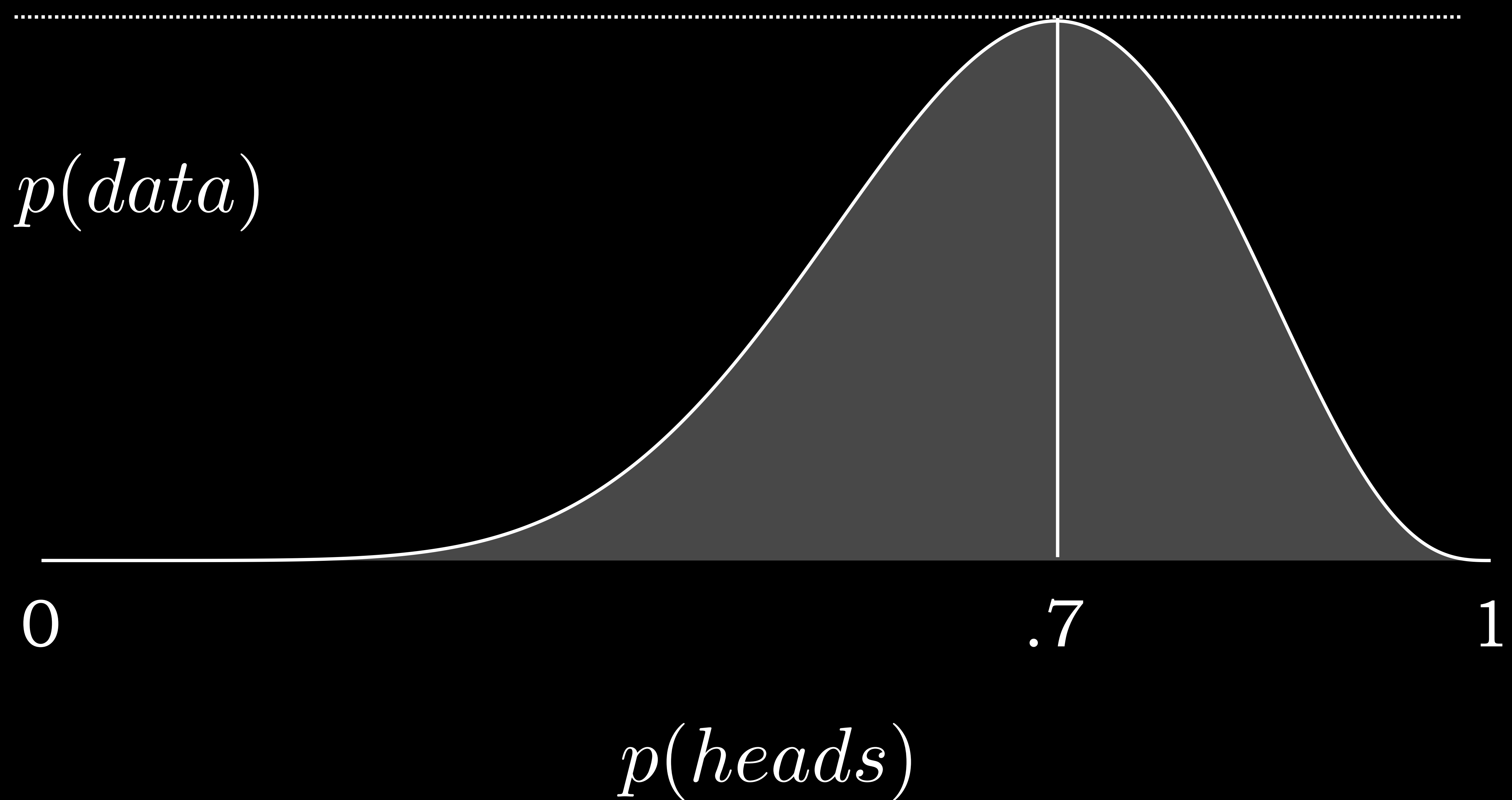
$p(heads)$?


ČESKÁ REPUBLIKA 50 KČ
PRAGA MATER URBIUM
ČESKÁ REPUBLIKA 50 KČ
ČESKÁ REPUBLIKA 50 KČ
PRAGA MATER URBIUM
ČESKÁ REPUBLIKA 50 KČ
ČESKÁ REPUBLIKA 50 KČ
ČESKÁ REPUBLIKA 50 KČ
PRAGA MATER URBIUM
ČESKÁ REPUBLIKA 50 KČ

$$p(data) = p(heads)^7 \times p(tails)^3$$

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


$p(data)$
0
1
$p(heads)$

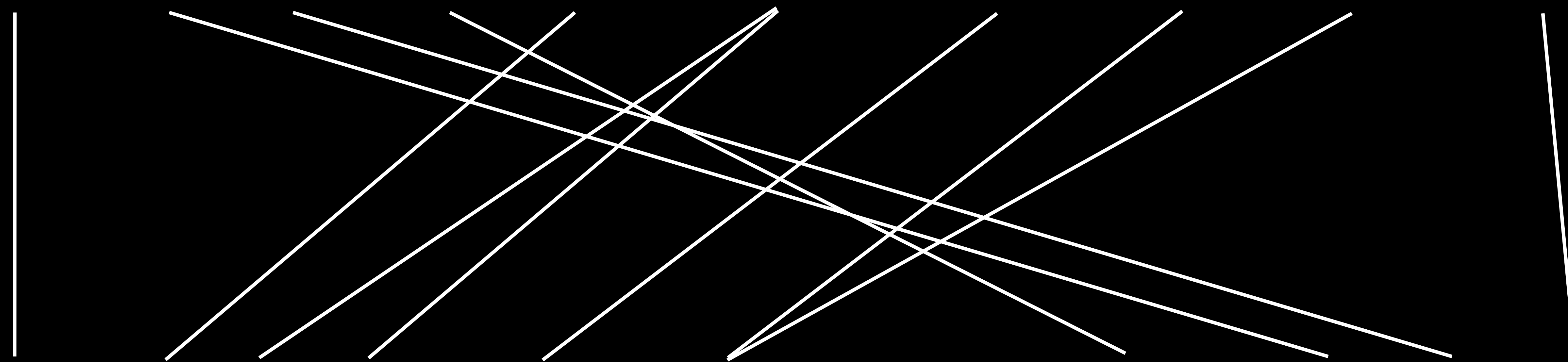


$p(data)$
0
.7
1
$p(heads)$

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 .

$$p(however|\text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

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 。

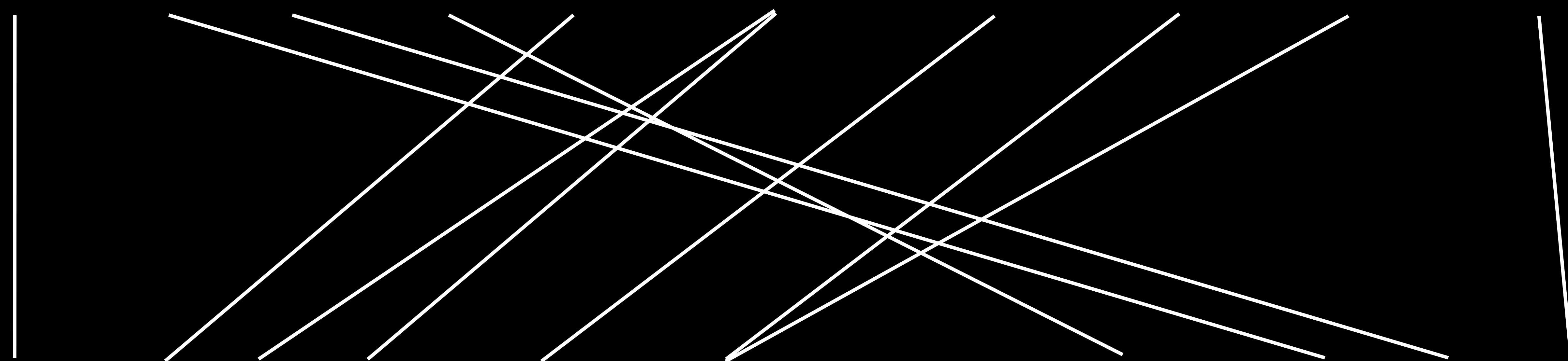
However , the sky remained clear under the strong north wind .

$$p(however|\text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

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 .

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

- **Optimization**

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)

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

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)

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

MLE for IBM Model 1 (observed)

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

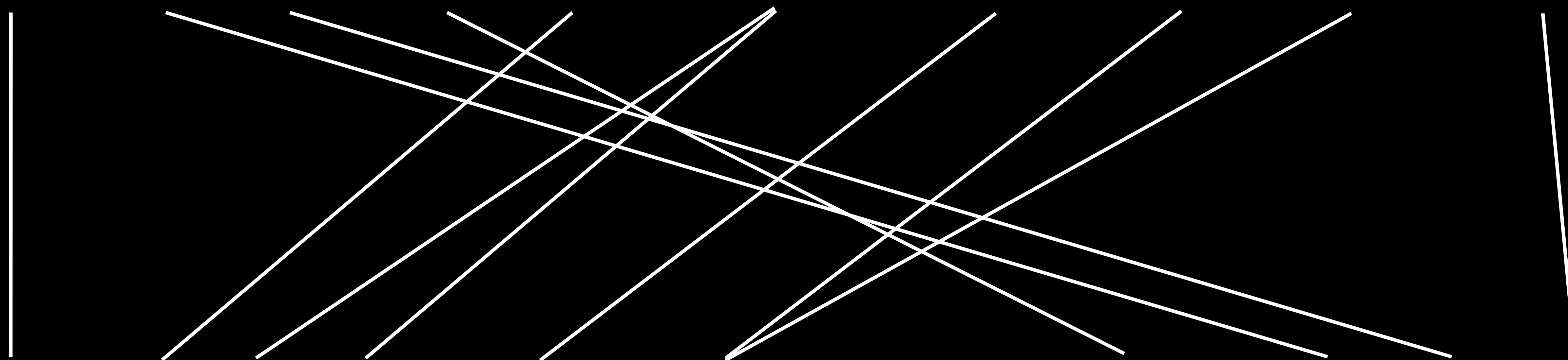
$$- \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)$$

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 .

$$p(however|\text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

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 .

$$p(however|\text{虽然}) = ???$$

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)

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

marginalize over alignments:

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

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.
- Guarantee: likelihood will be monotonically nondecreasing.

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 .

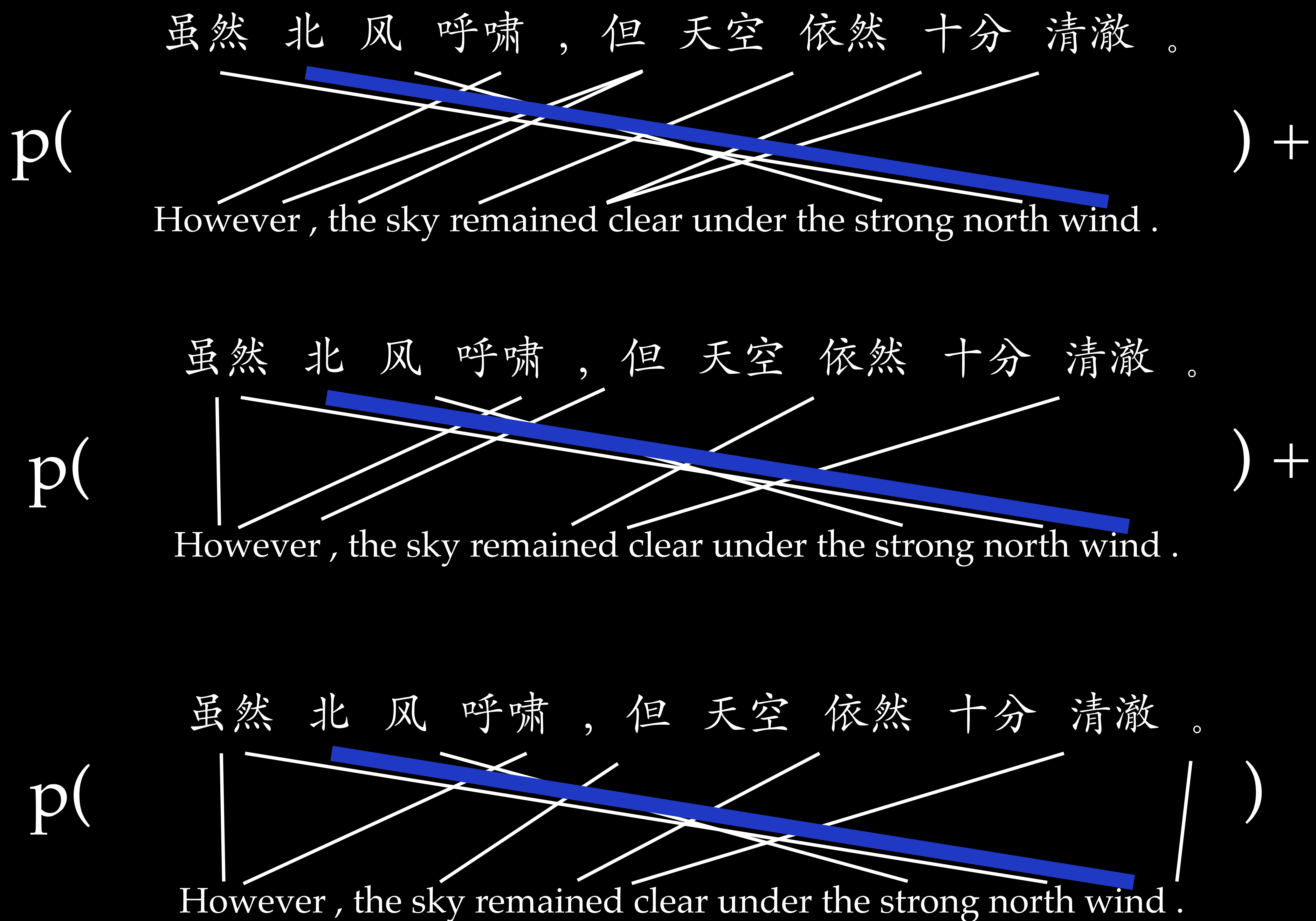
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

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

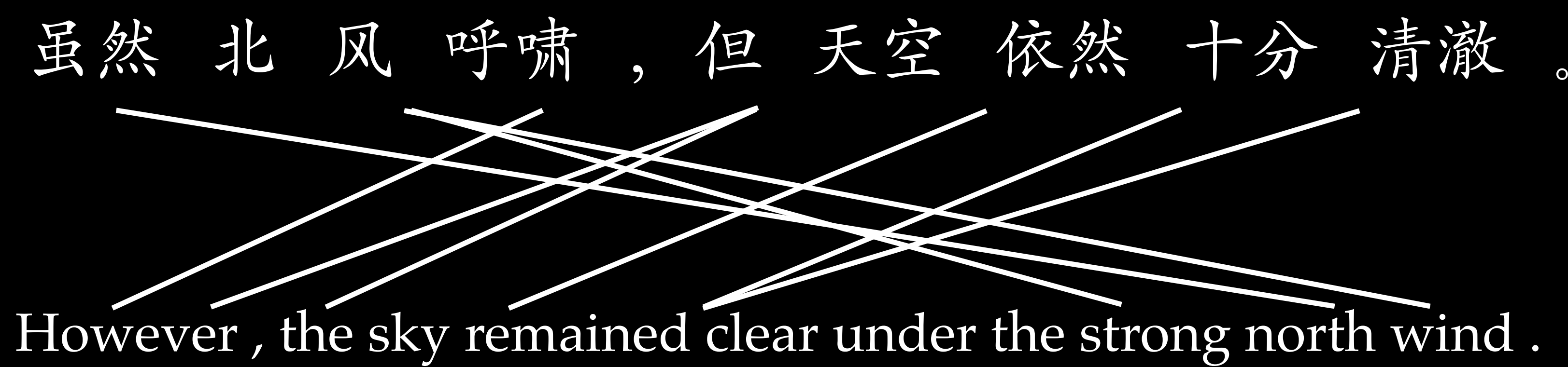
since we didn't observe the alignment, we calculate the probability that it's there.

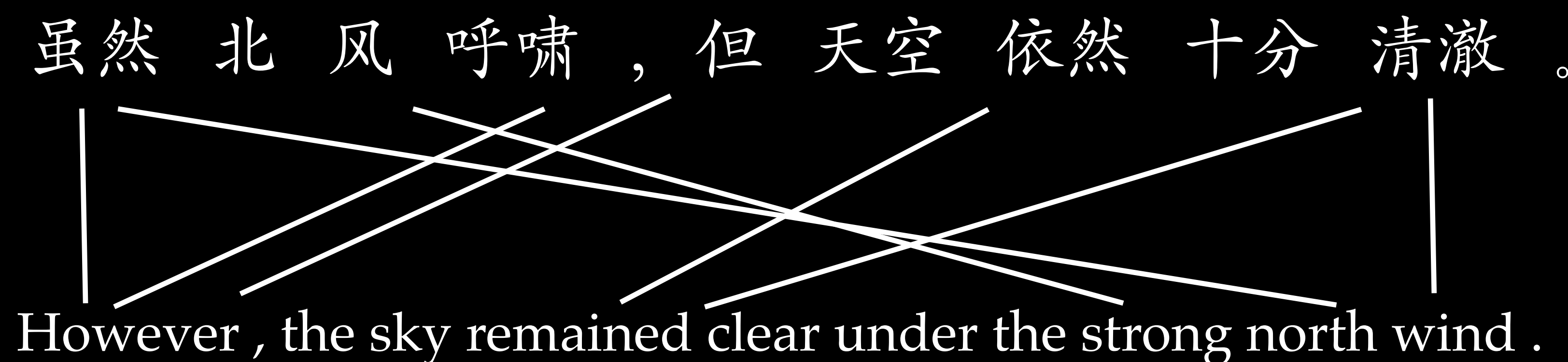
However , the sky remained clear under the strong north wind .

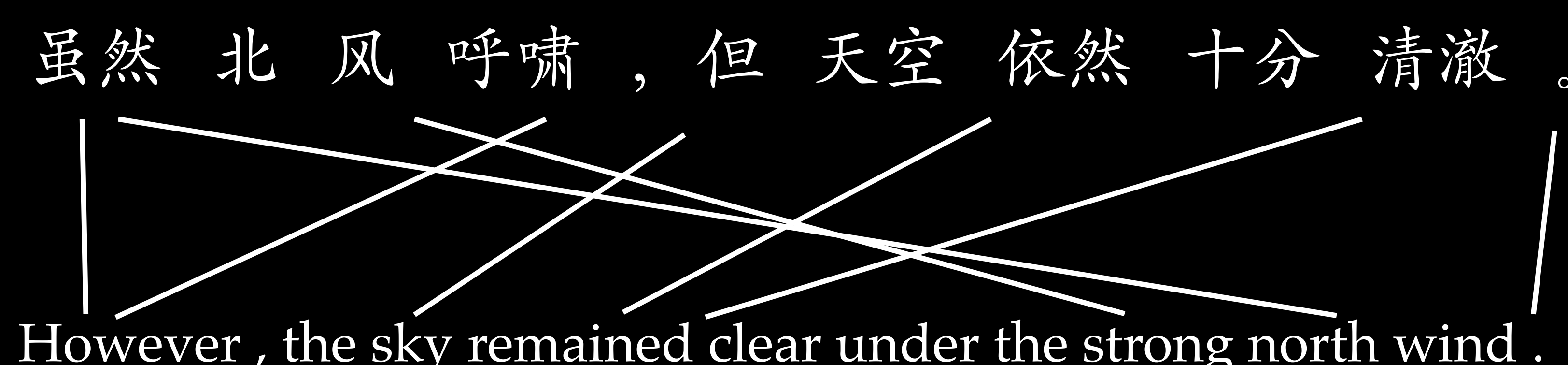
Marginalize: sum all alignments containing the link



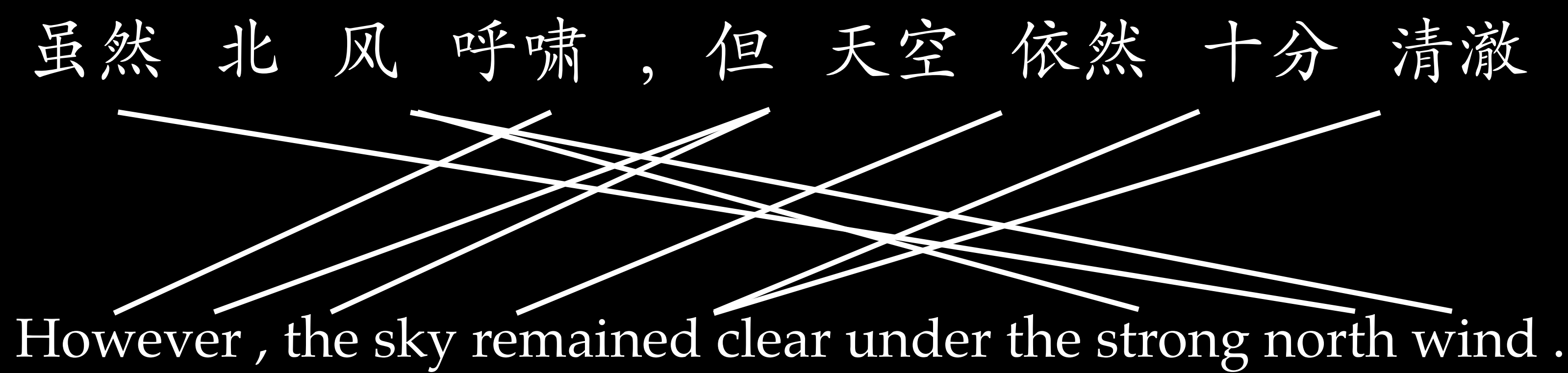
Divide by sum of all *possible* alignments

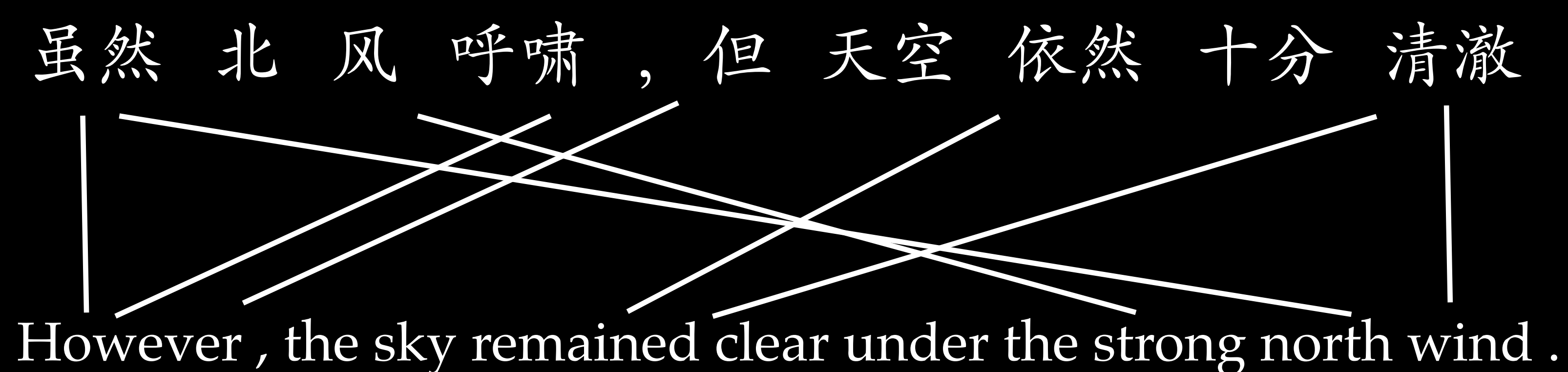
p() +

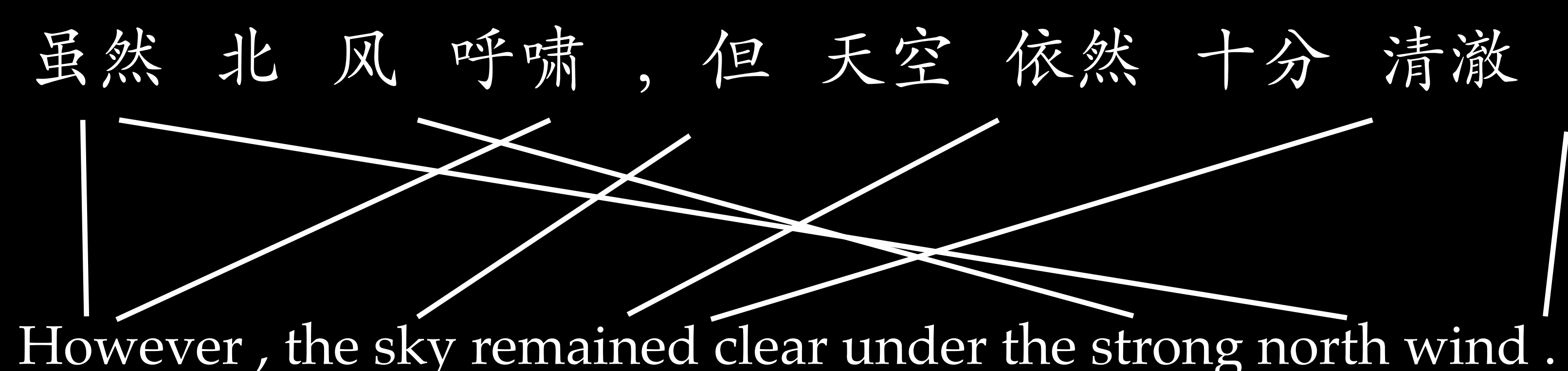
p() +

p()

Divide by sum of all *possible* alignments

p() +

p() +

p()

Is this hard? How many alignments are there?

Expectation Maximization

probability of an alignment.

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

Expectation Maximization

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

Expectation Maximization

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

Expectation Maximization

marginal probability of
alignments containing link

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

Expectation Maximization

marginal probability of
alignments containing link

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

Expectation Maximization

marginal probability of
alignments containing link

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

marginal probability of all
alignments

Expectation Maximization

marginal probability of
alignments containing link

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

identical!

marginal probability of all
alignments

Expectation Maximization

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

Expectation Maximization

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

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

Expectation Maximization

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

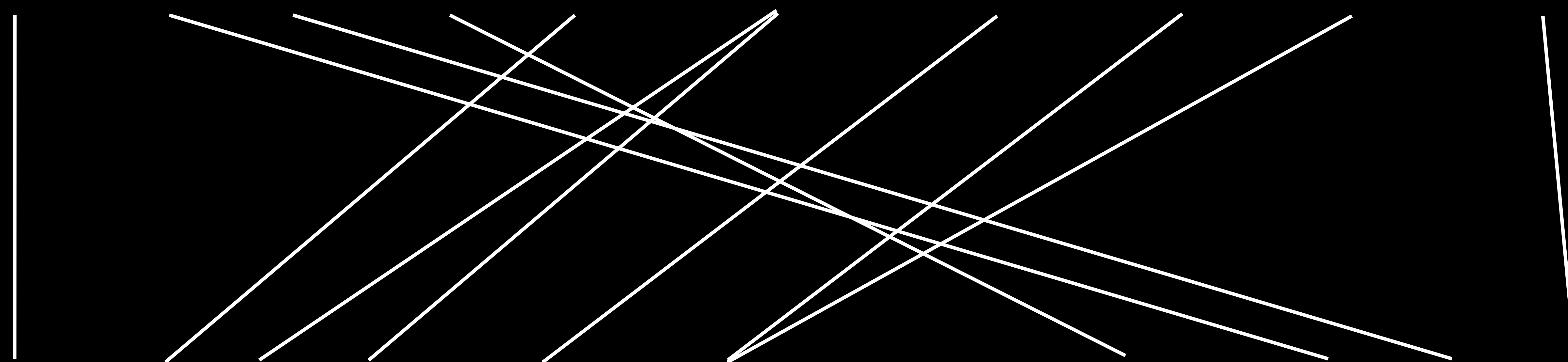
$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

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

Translation Models

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 .

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

Translation Models

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 .

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

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.

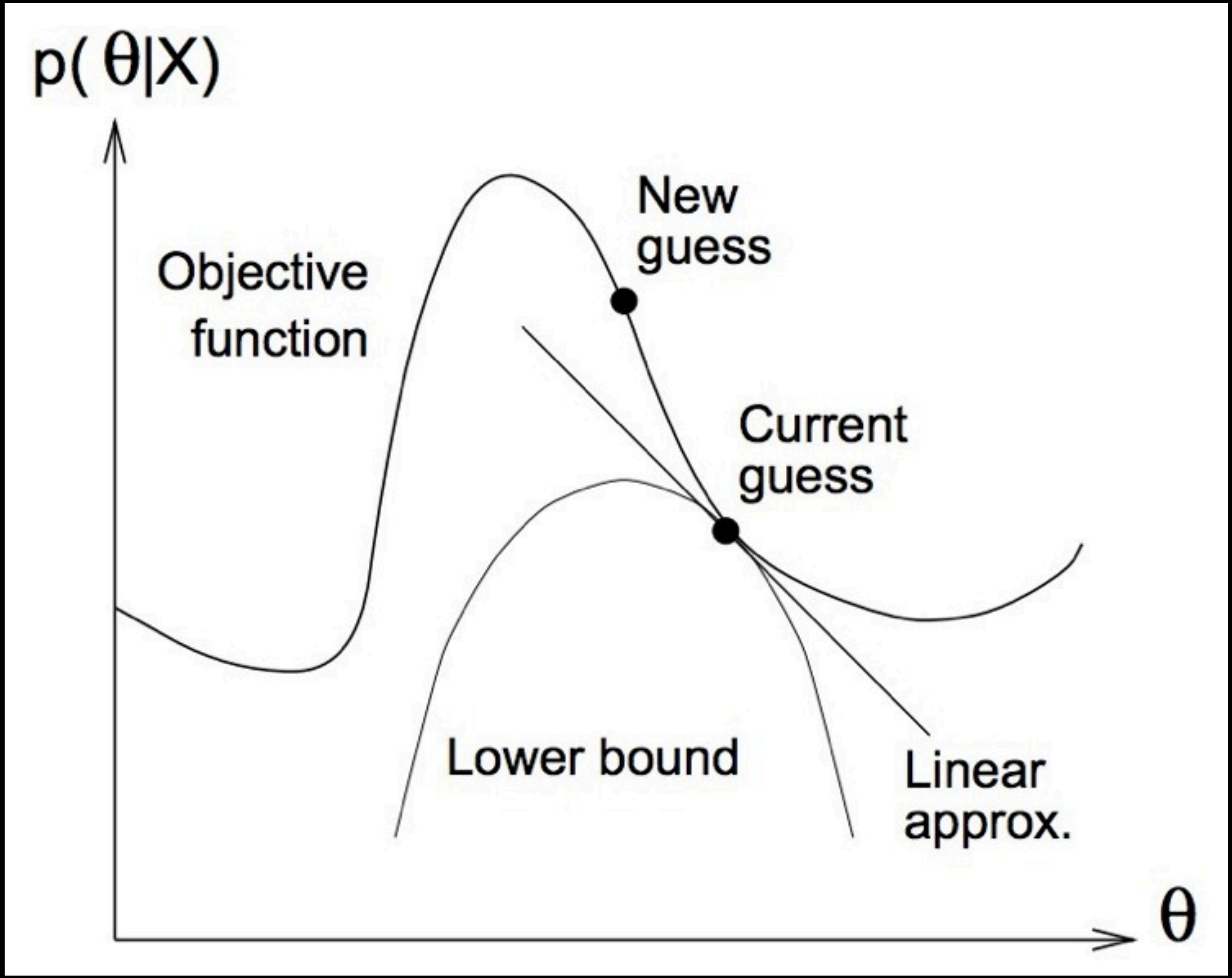
Expectation Maximization

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

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

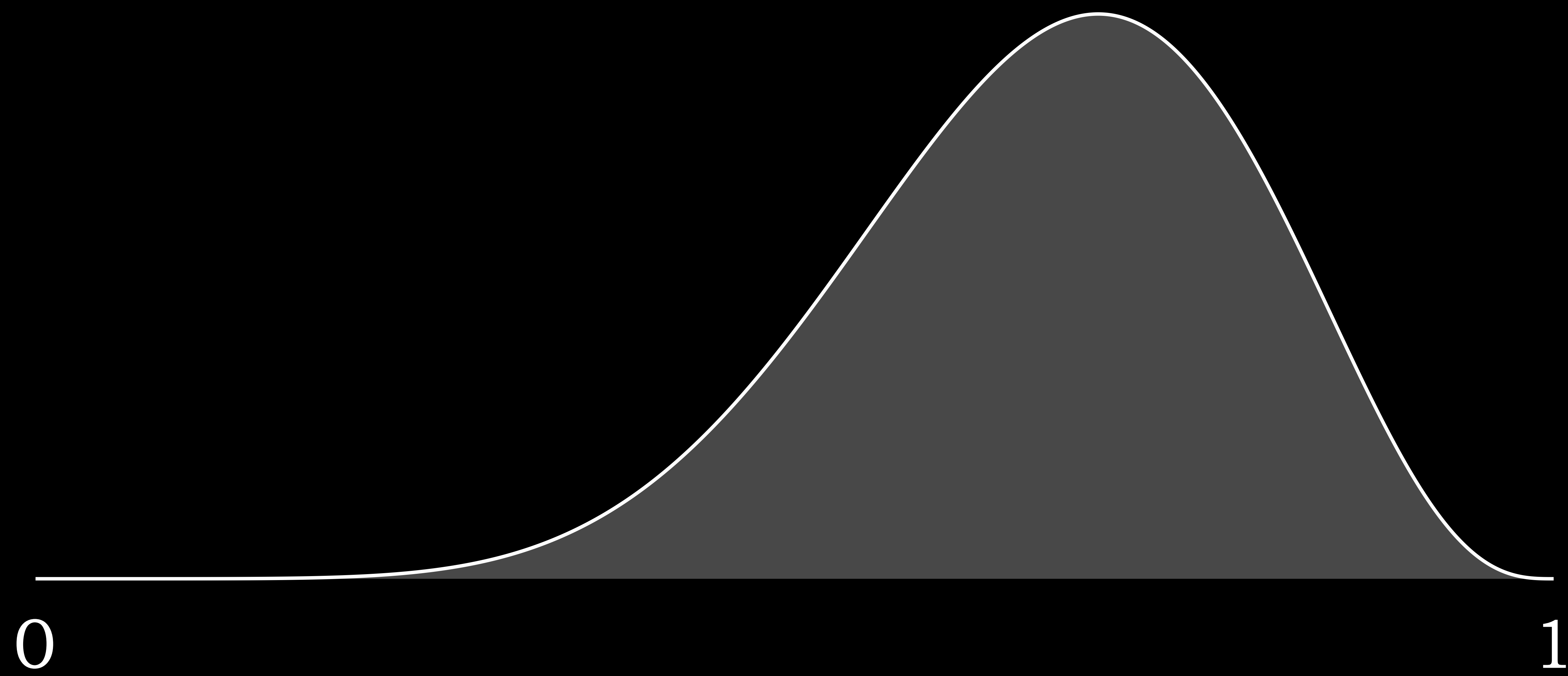
MLE: choose parameters that maximize this expression.

Minor problem: there is no analytic solution.



(from Minka ’98)

... and, likelihood is ***convex*** for this model:



Summary

- Many possible models.
- Many possible objective functions.
- ***Learning is optimization***: choose parameters that optimize some function, such as likelihood.
- Try some out this afternoon in the lab!