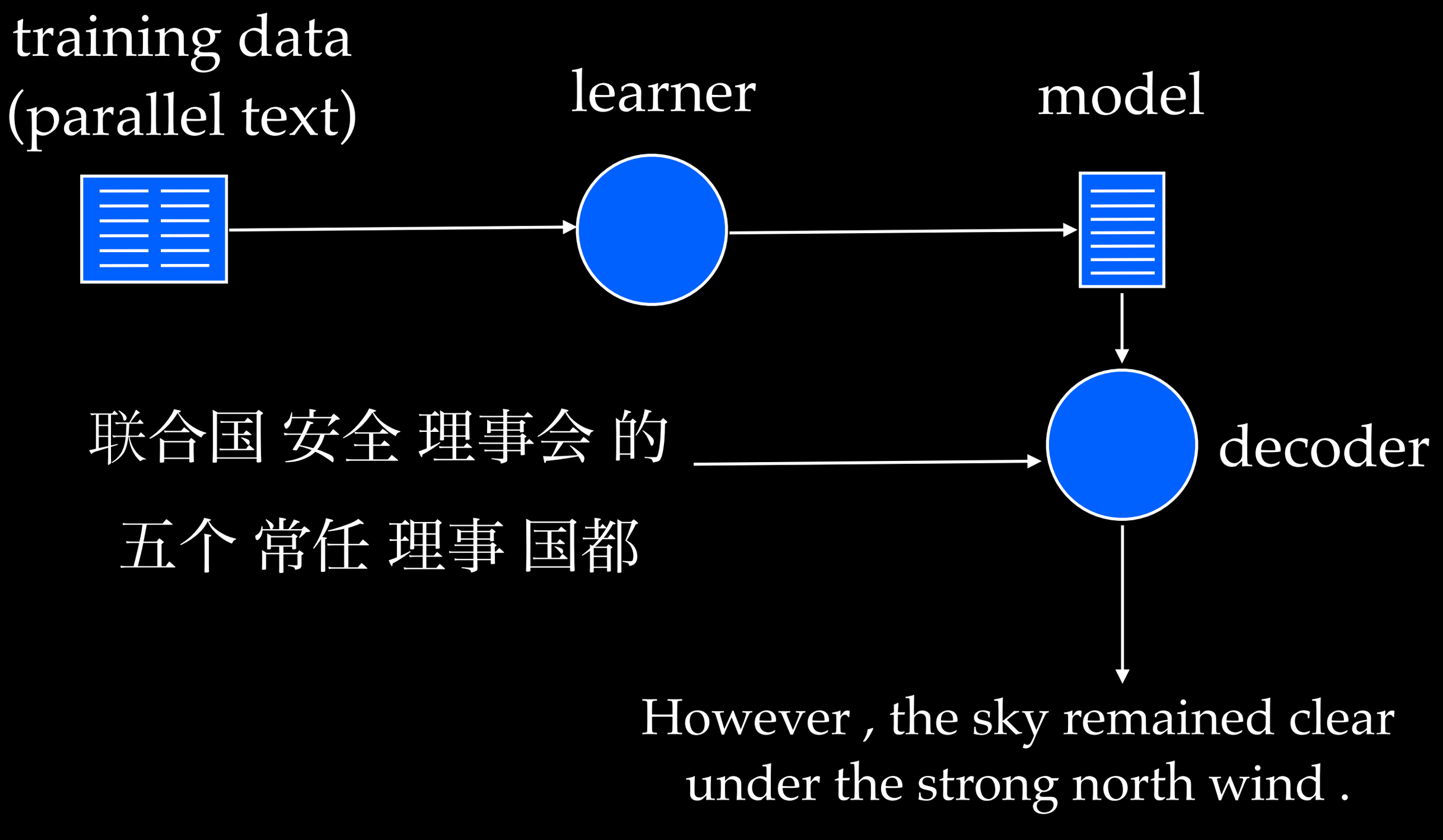
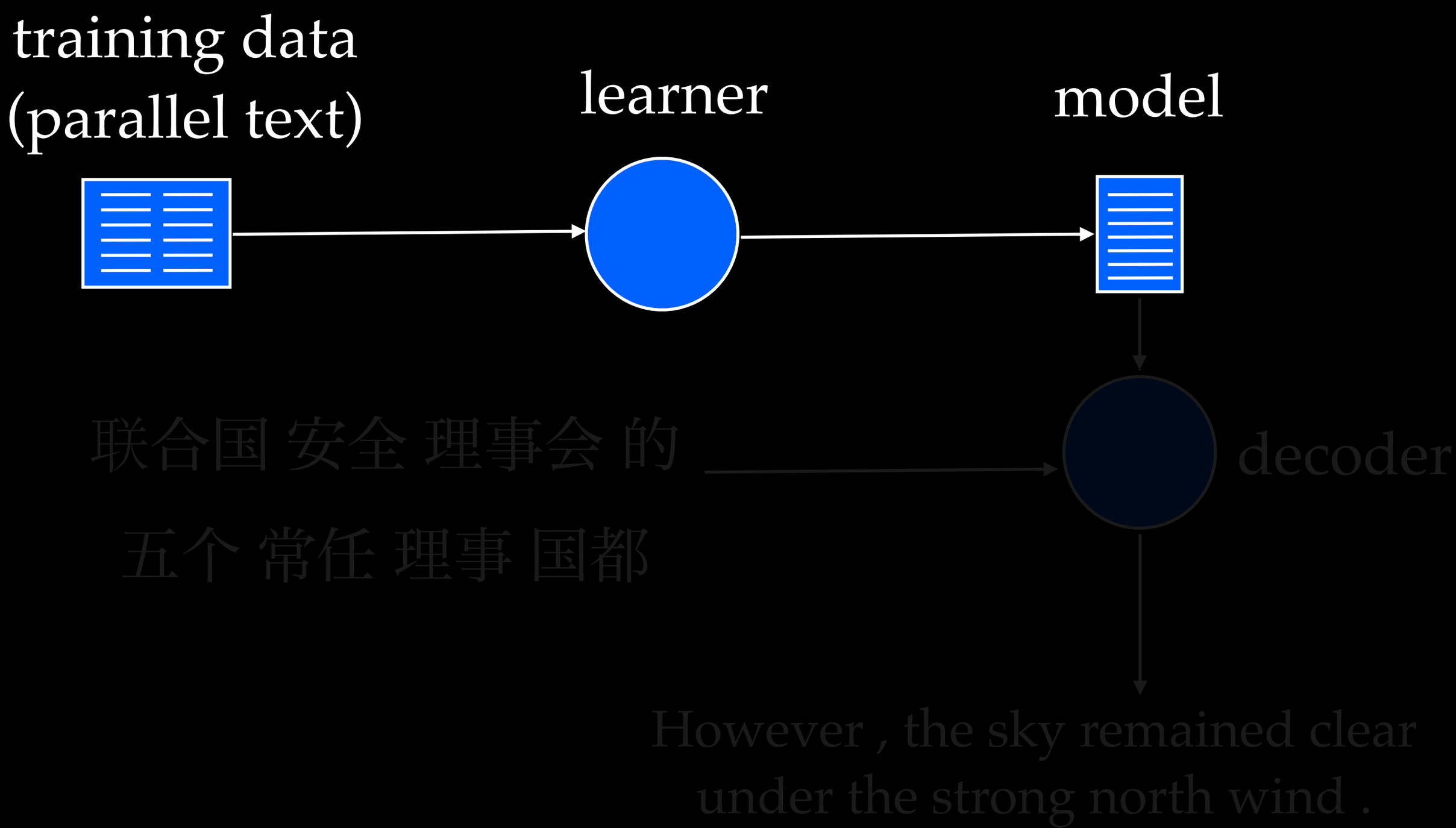
# Learning Translation Models: Word Alignment

# Overview


training data
(parallel text)
learner
model
联合国 安全 理事会 的
五个 常任 理事 国都
decoder
However , the sky remained clear
under the strong north wind .

# Overview

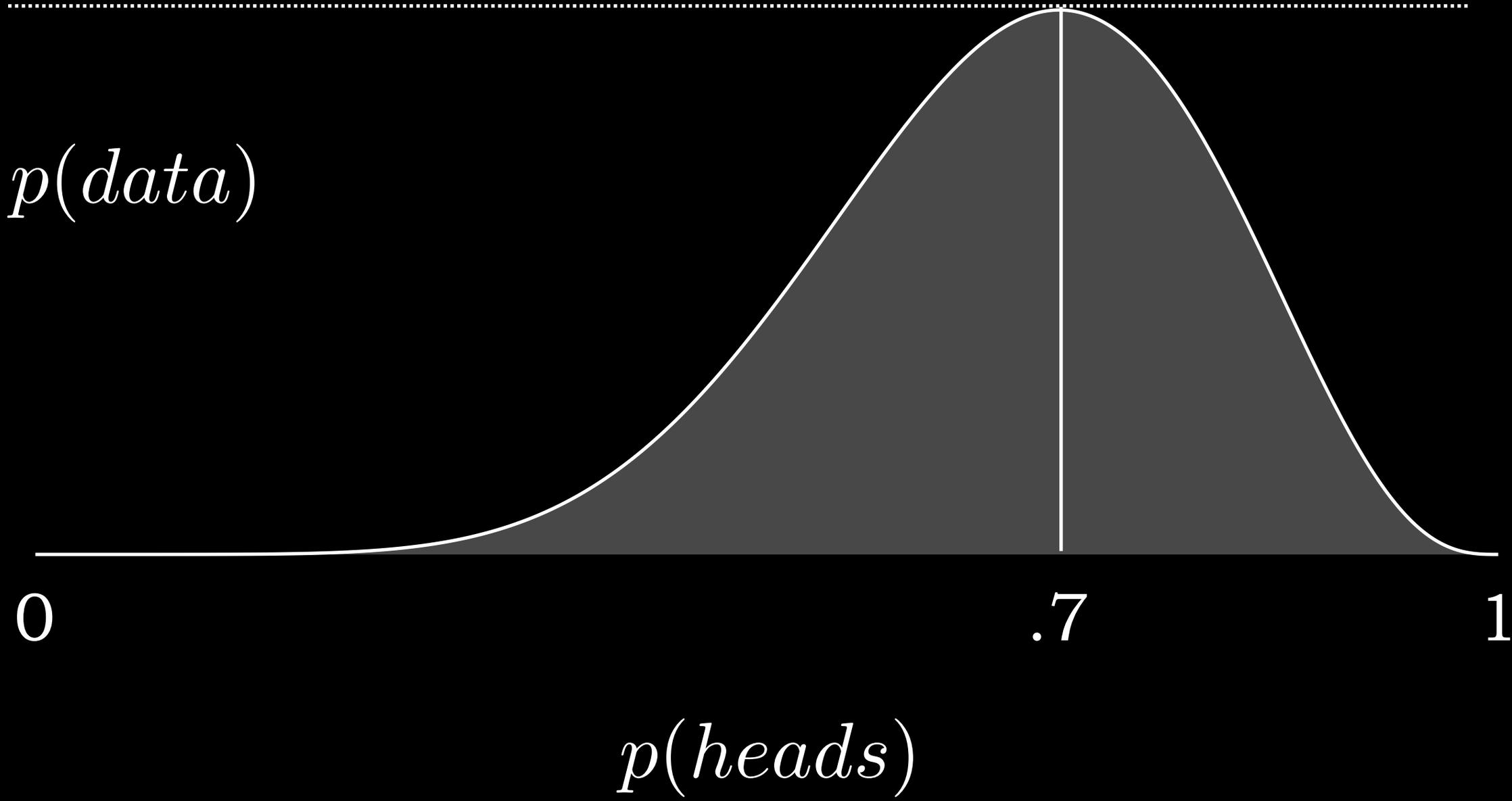

training data
(parallel text)
learner
model
联合国 安全 理事会 的
五个 常任 理事 国都
decoder
However , the sky remained clear
under the strong north wind .


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

$p(heads)$ ?


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

can be derived analytically using Lagrange multipliers




- Øptimization

# Maximum Likelihood Estimation

The ***maximum likelihood estimate*** of a set of parameters is the one that solves:

$$\text{argmax}_{parameters}\, p(data|parameters)$$

# Maximum Likelihood Estimation

The ***maximum likelihood estimate*** of a set of parameters is the one that solves:

$$\text{argmax}_{parameters}\, p(data|parameters)$$

maximum

likelihood

# Language Models Again

Maximum likelihood estimate of a probability in an n-gram language model:

$$p(word_n|word_1...word_{n-1}) = \frac{count(word_1...word_n)}{count(word_1...word_{n-1})}$$

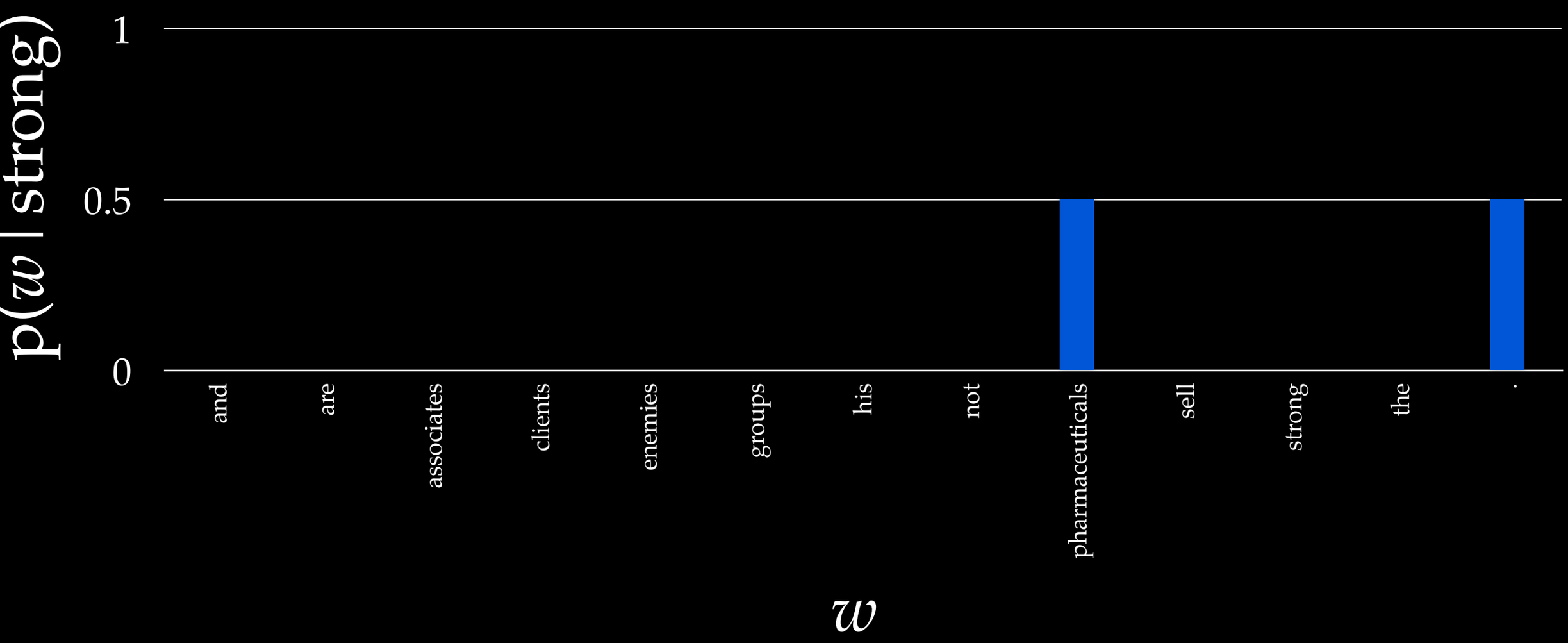
What happens if we have not seen an *n*-gram?

# Maximum Likelihood Estimation

- If an event is not observed, its maximum likelihood estimate will be zero. This is an example of *overfitting*, and is often not what we want.
- In the case of language models, we do not want any non-zero probabilities!
- Can be solved by various techniques to move probability mass onto the unobserved events.

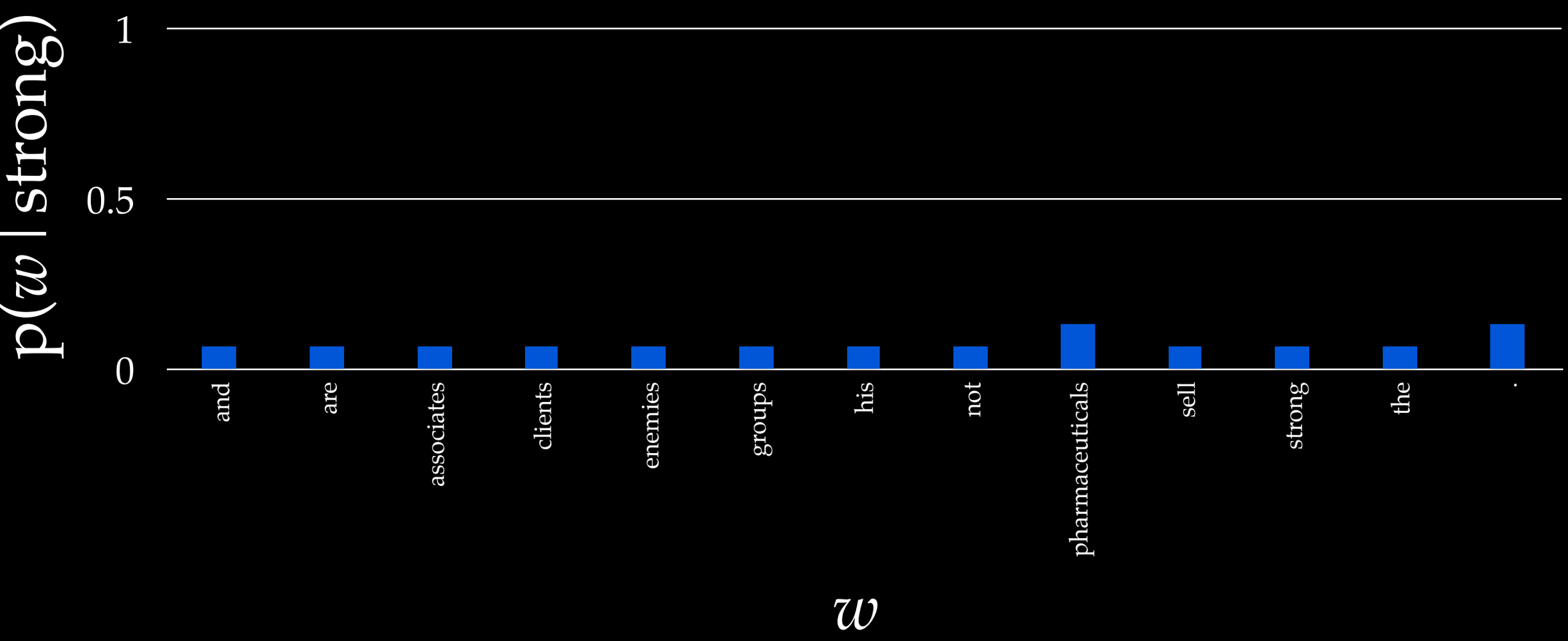
# Maximum Likelihood

his associates are not strong .
the clients and the associates are enemies .
the modern groups sell strong pharmaceuticals .



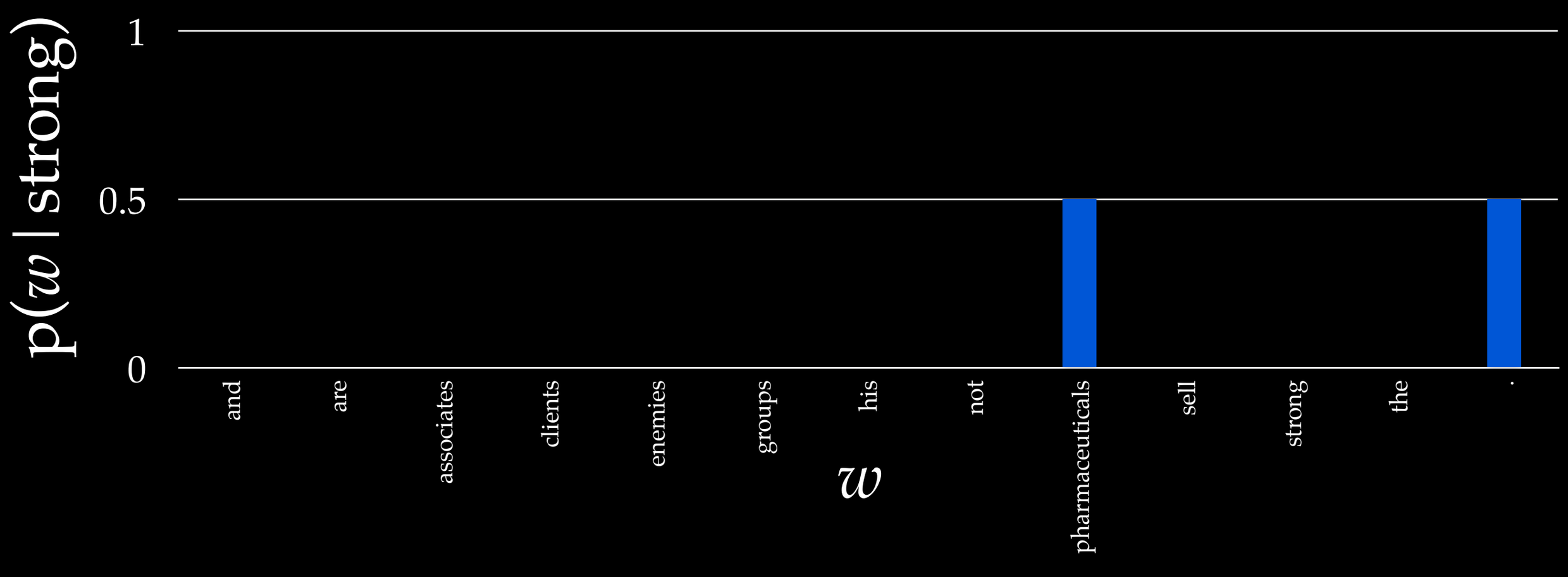
# Smoothed Maximum Likelihood

his associates are not strong .
the clients and the associates are enemies .
the modern groups sell strong pharmaceuticals .

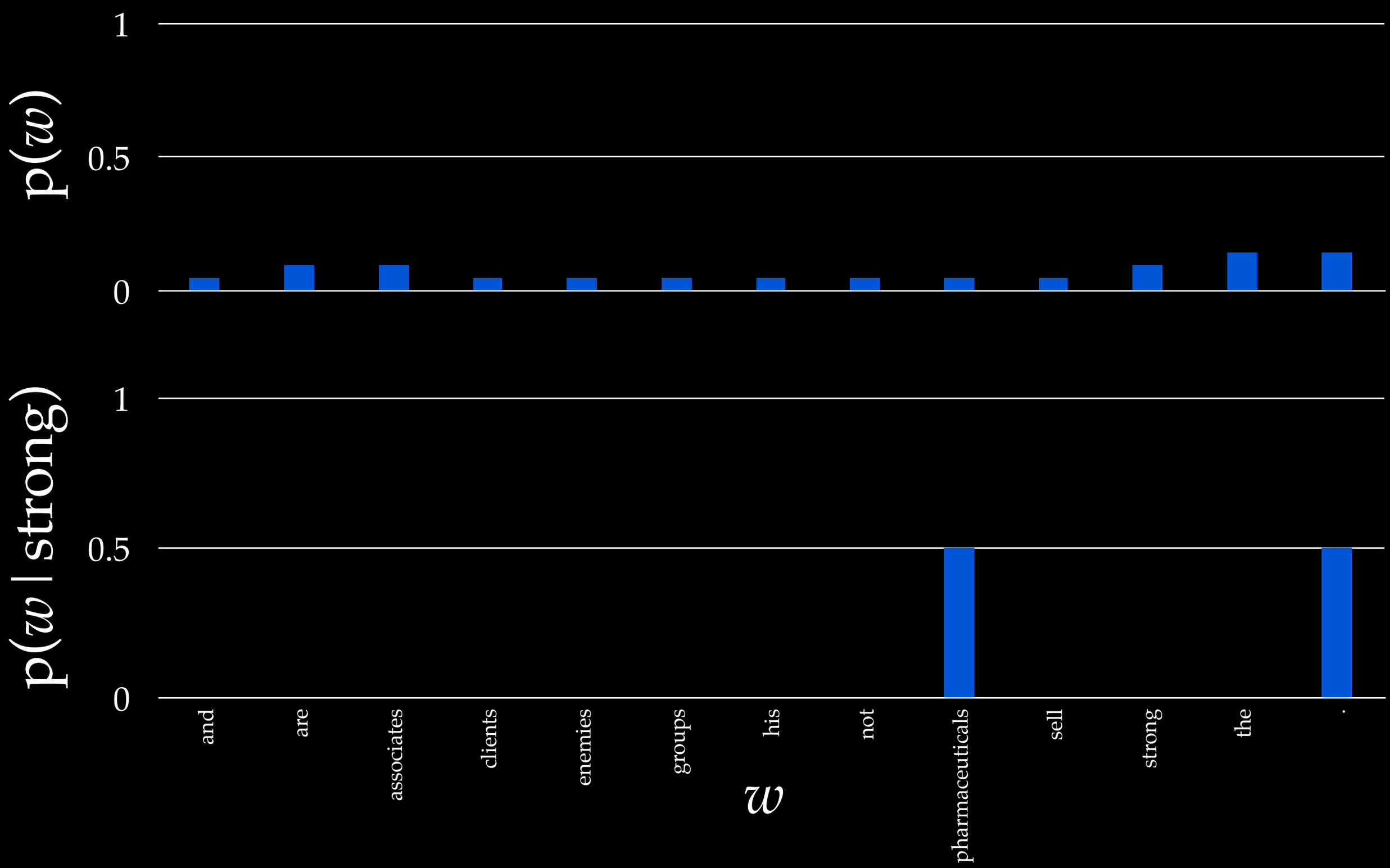


# Maximum Likelihood

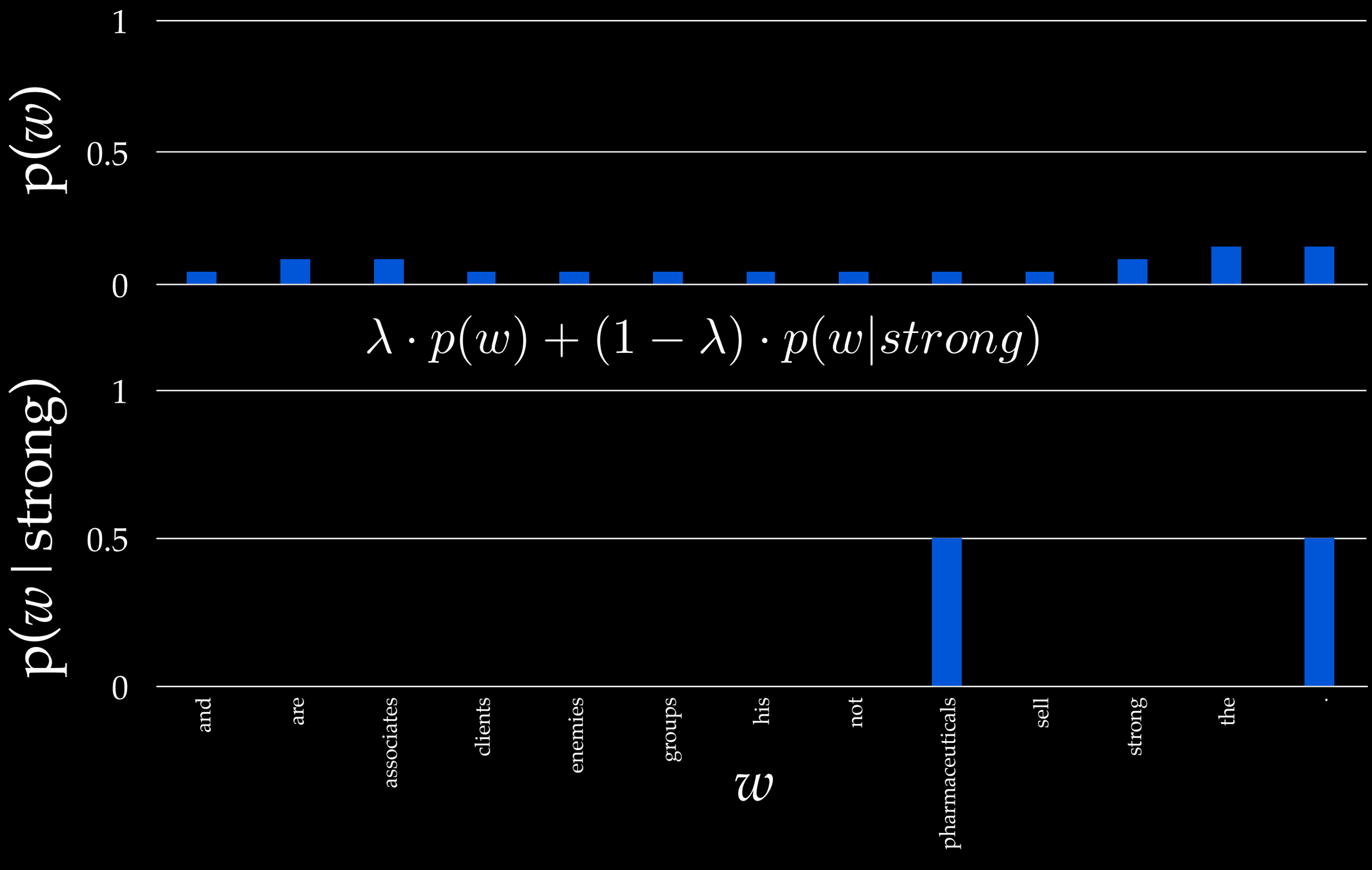
his associates are not strong .
the clients and the associates are enemies .
the modern groups sell strong pharmaceuticals .



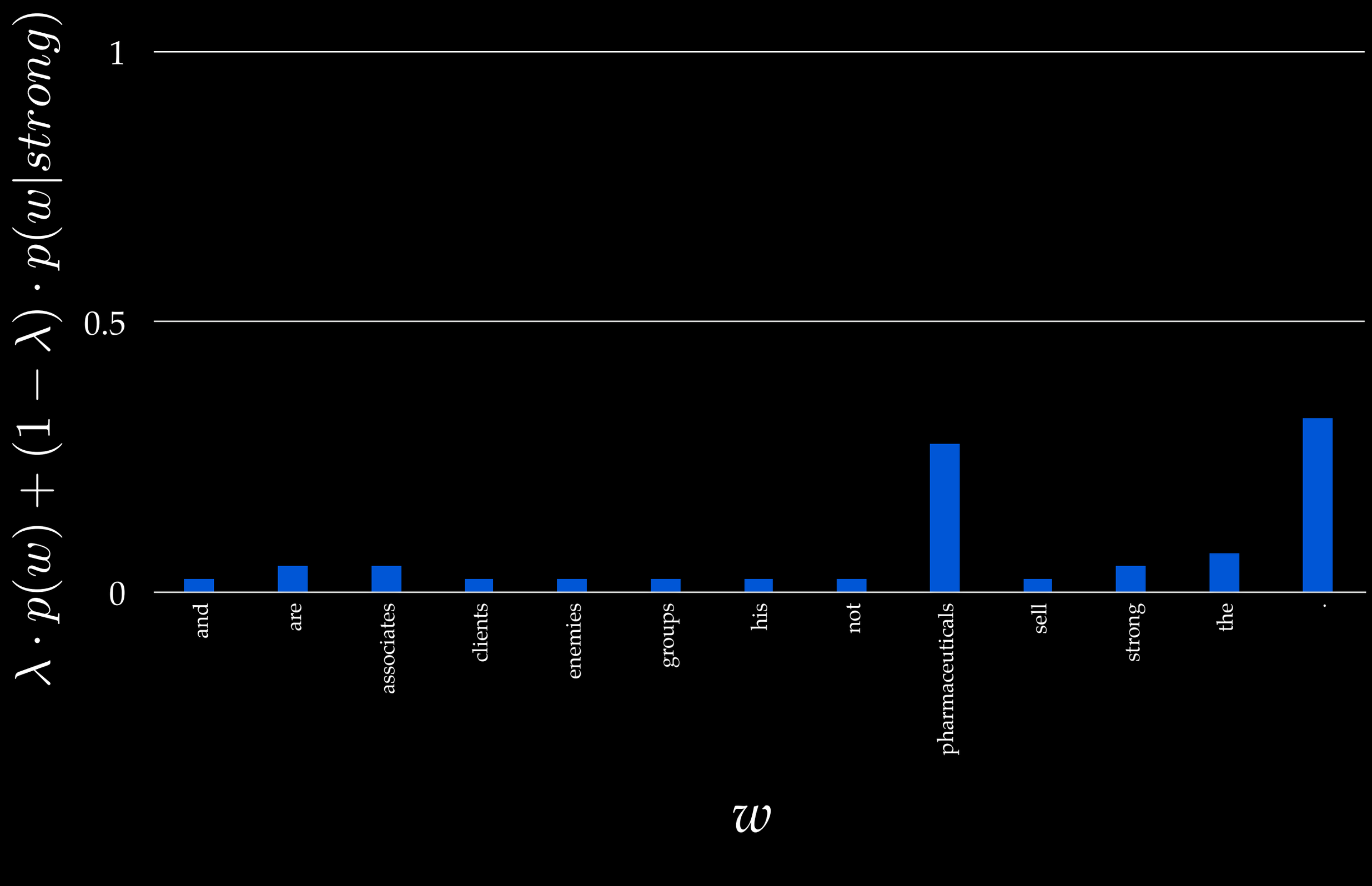
# Maximum Likelihood



# Maximum Likelihood



# Backed-off Maximum Likelihood



# Smoothing & Backoff

Intuitively, we can think of these techniques as optimizing likelihood under a definition that accounts for unseen data.

Further reading:

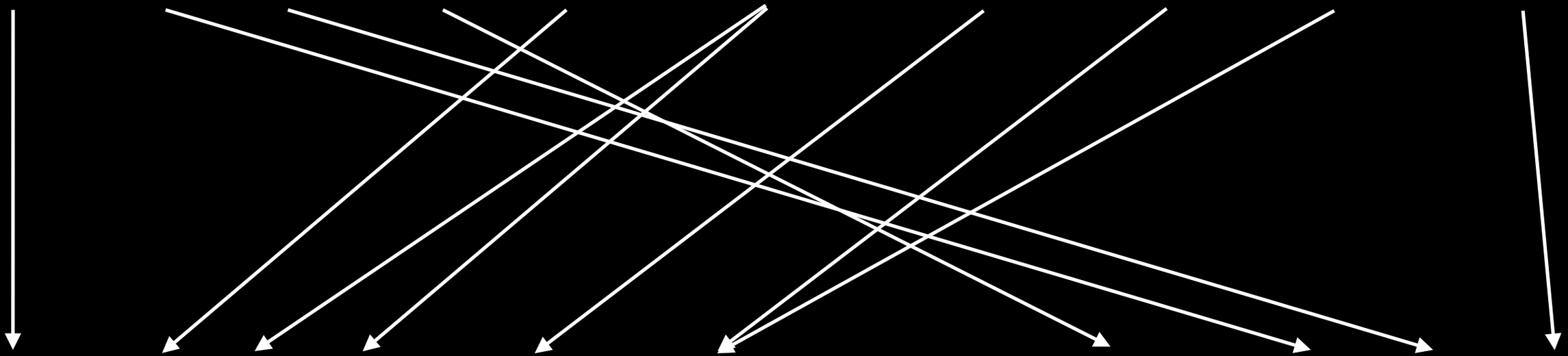
An empirical study of smoothing techniques for language modeling.

Stanley F. Chen & Josh Goodman, 1998.

# 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 。

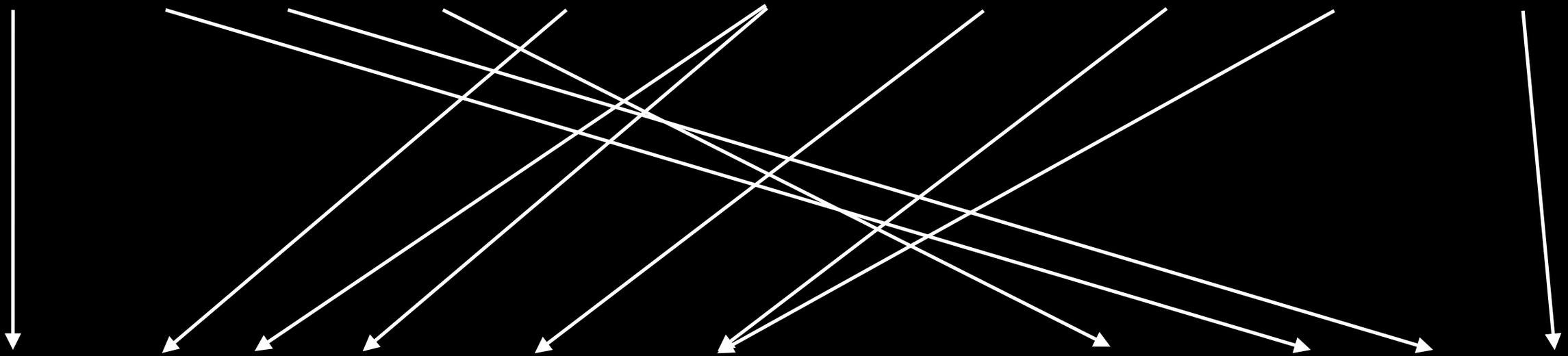


However , the sky remained clear under the strong north wind .

# 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 .

$$p(English|Chinese)?$$

# 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

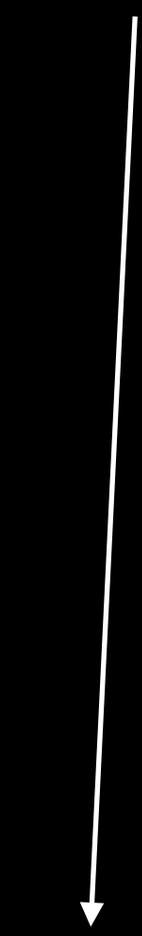
*Although north wind howls , but sky still very clear .*

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

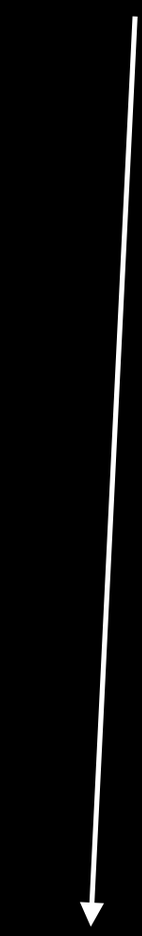


However

# 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



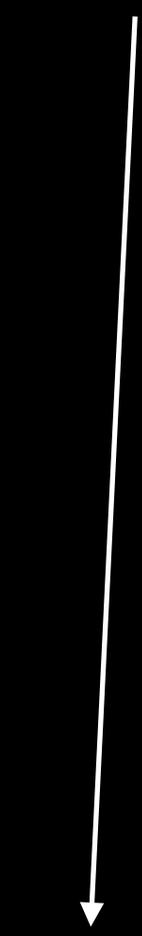
However

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

# 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

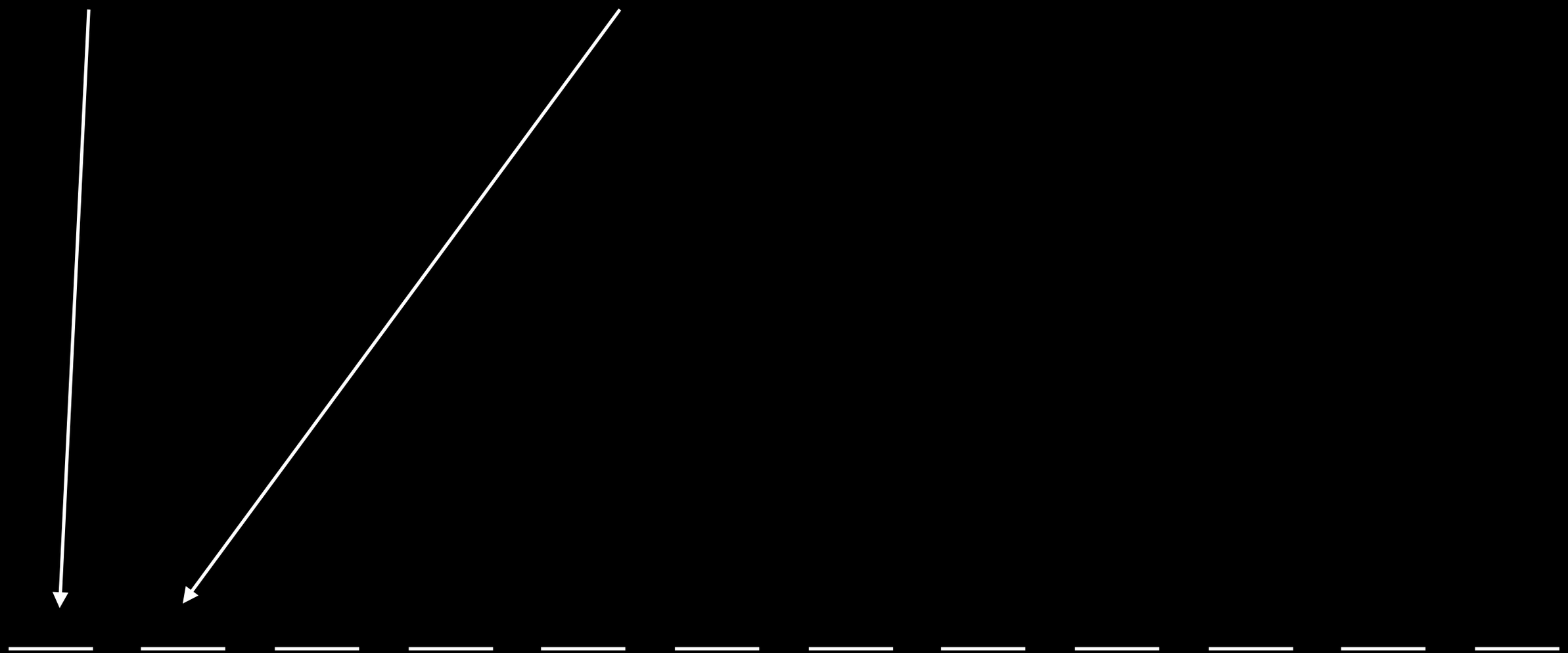


However

# 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

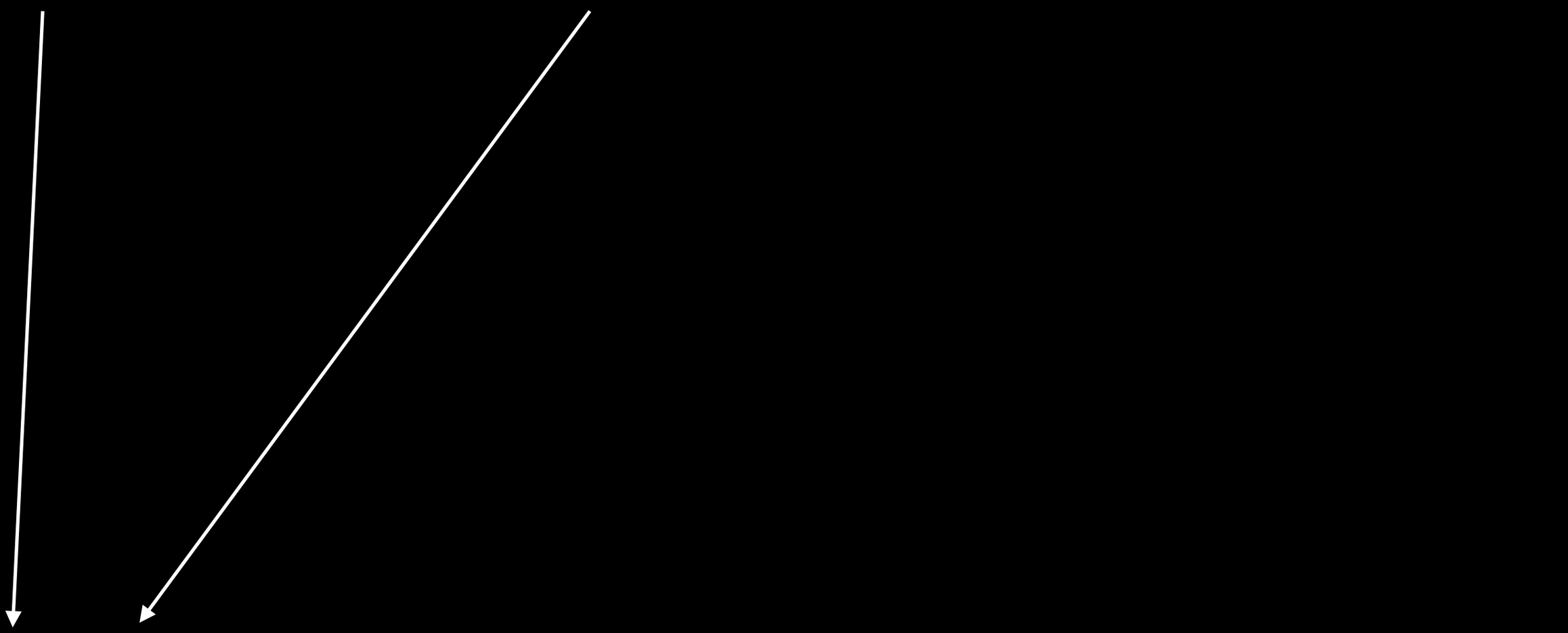


However

# 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

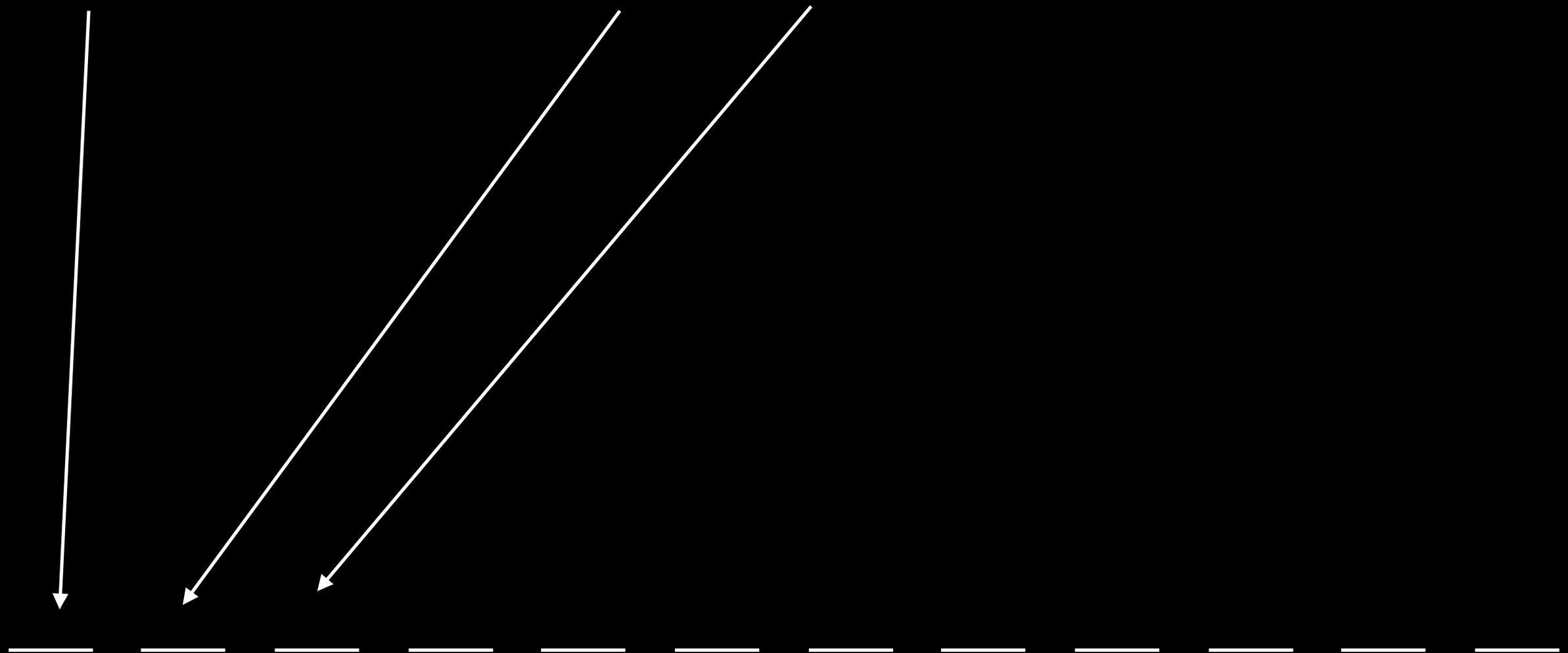


However ,

# 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

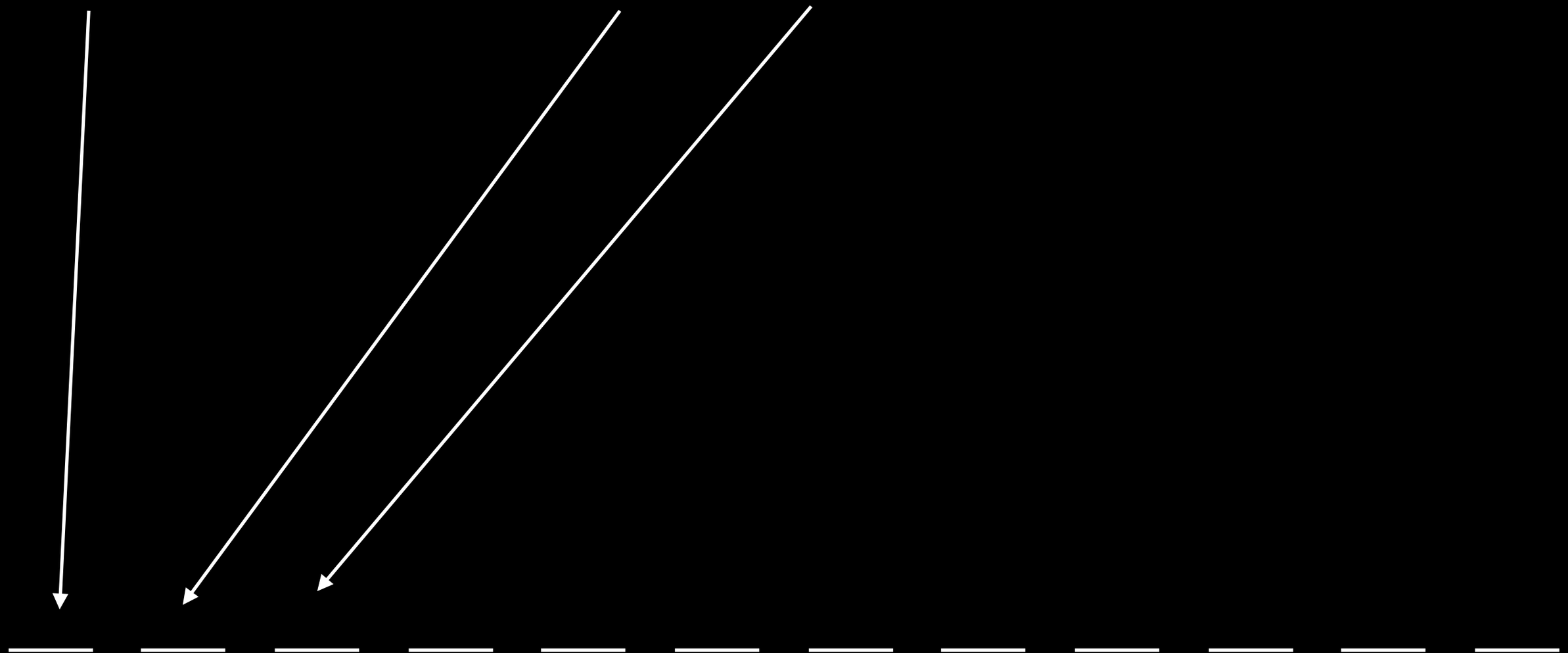


However ,

# 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

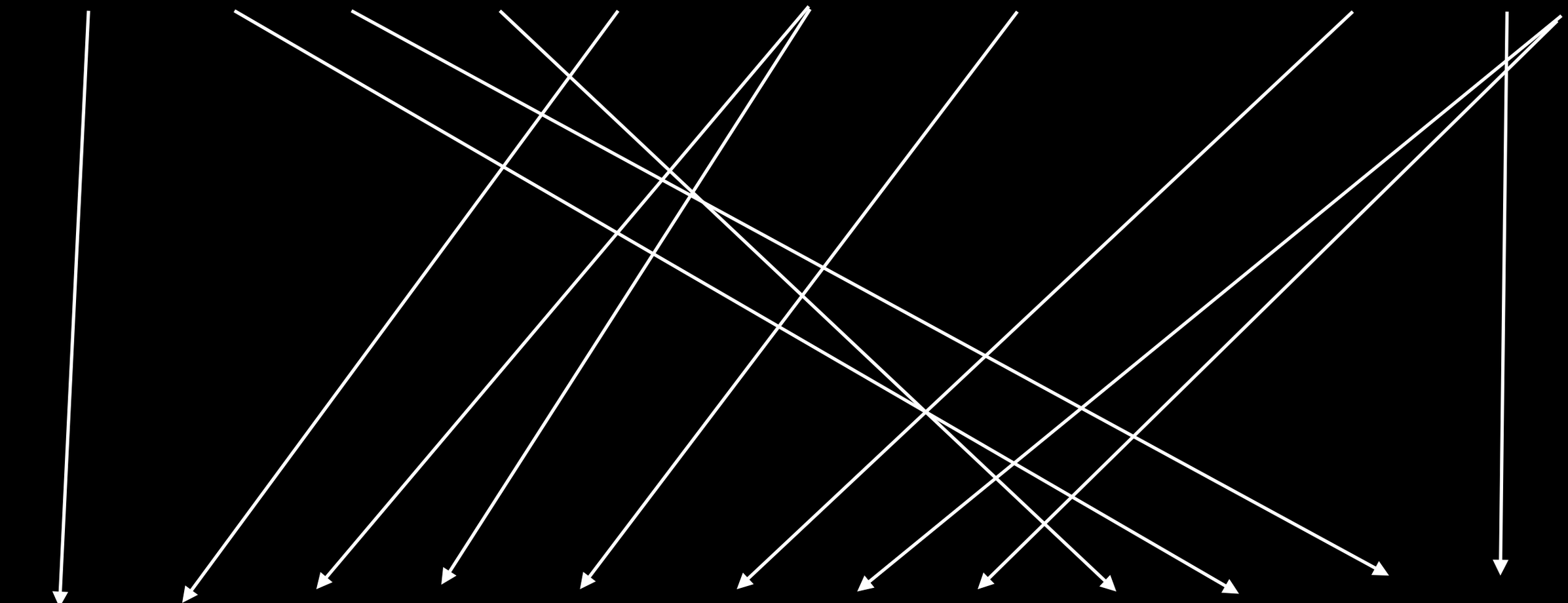


However , the

# 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

# IBM Model 1

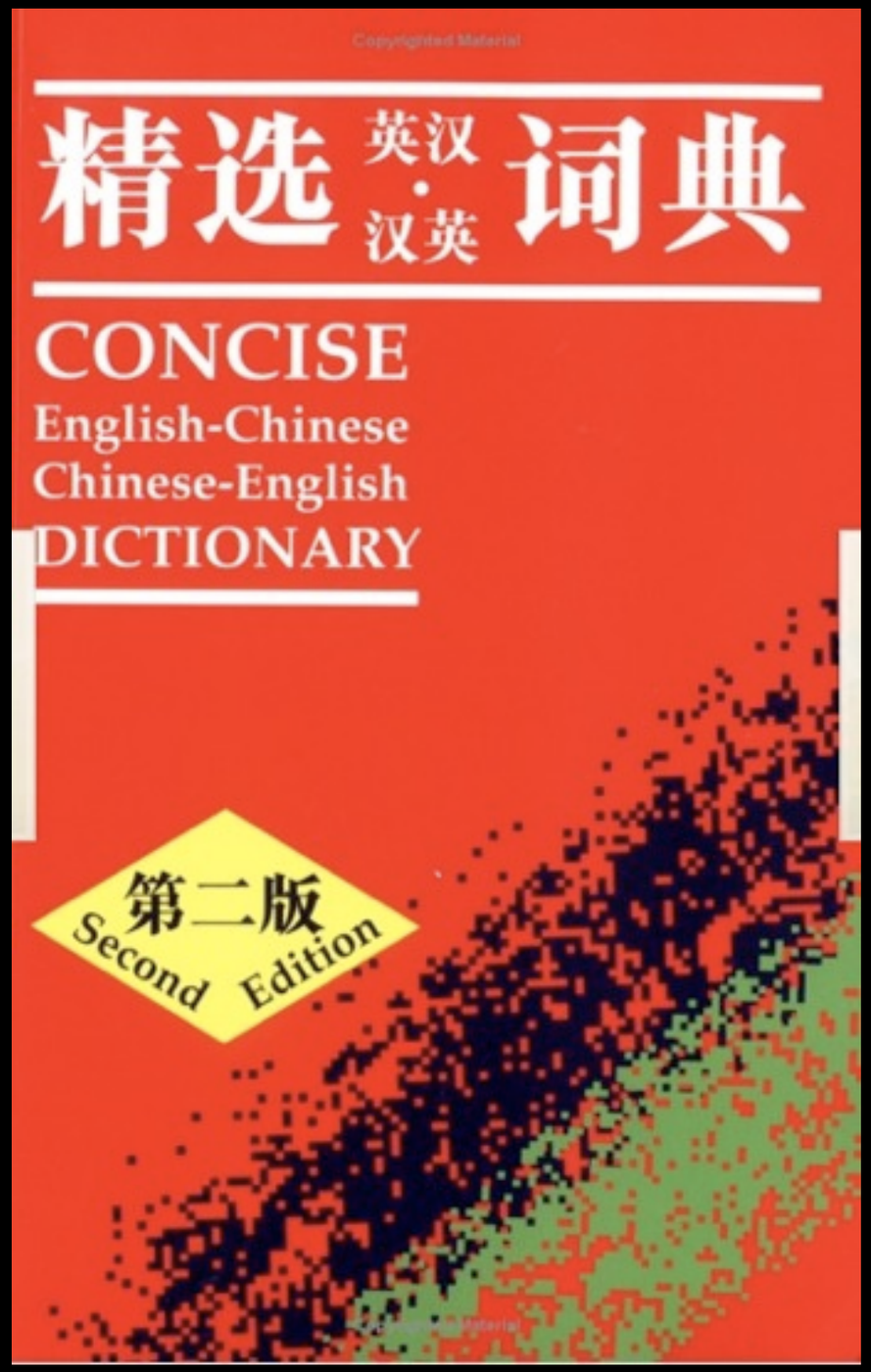
- Word translation probabilities.

# IBM Model 1

- Word translation probabilities.
- No real ordering model.
  - This is left to the LM.

# IBM Model 1

- Word translation probabilities.
- No real ordering model.
  - This is left to the LM.



# IBM Model 1

- Word translation probabilities.
- No real ordering model.
  - This is left to the LM.

# 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

$p(despite|\text{虽然})$ ???

$p(however|\text{虽然})$ ???

$p(although|\text{虽然})$ ???

$p(northern|\text{北})$ ???

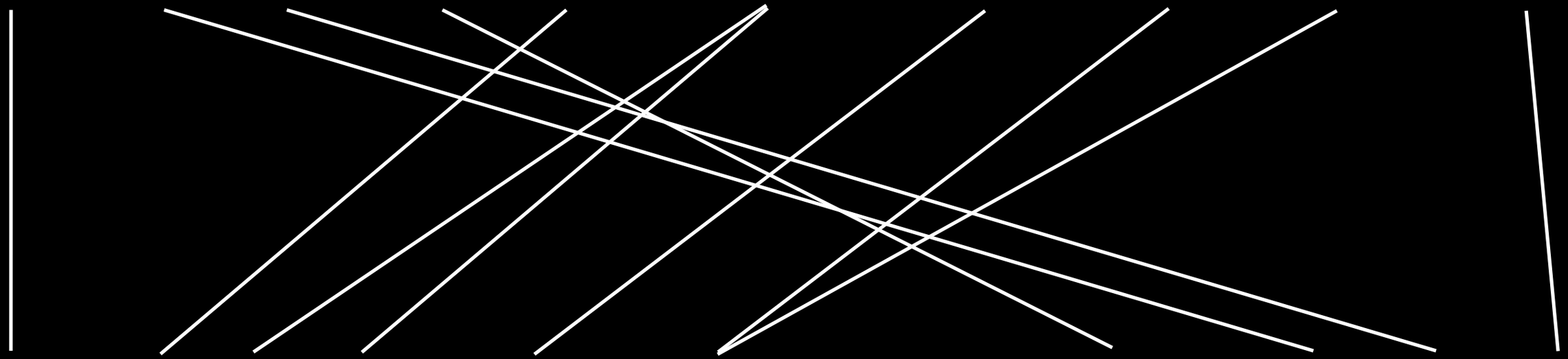
$p(north|\text{北})$ ???

Side note: p(X | Y) is given here (but not always)

# 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 .

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

# 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 .

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

# Likelihood Estimation for Model 1

# 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 。

# 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

# 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

# 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

— — — — — — — — — — — — —

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

# 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

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

# 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

# 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

$p(Chinese\ word\ position)$

# Likelihood Estimation for Model 1

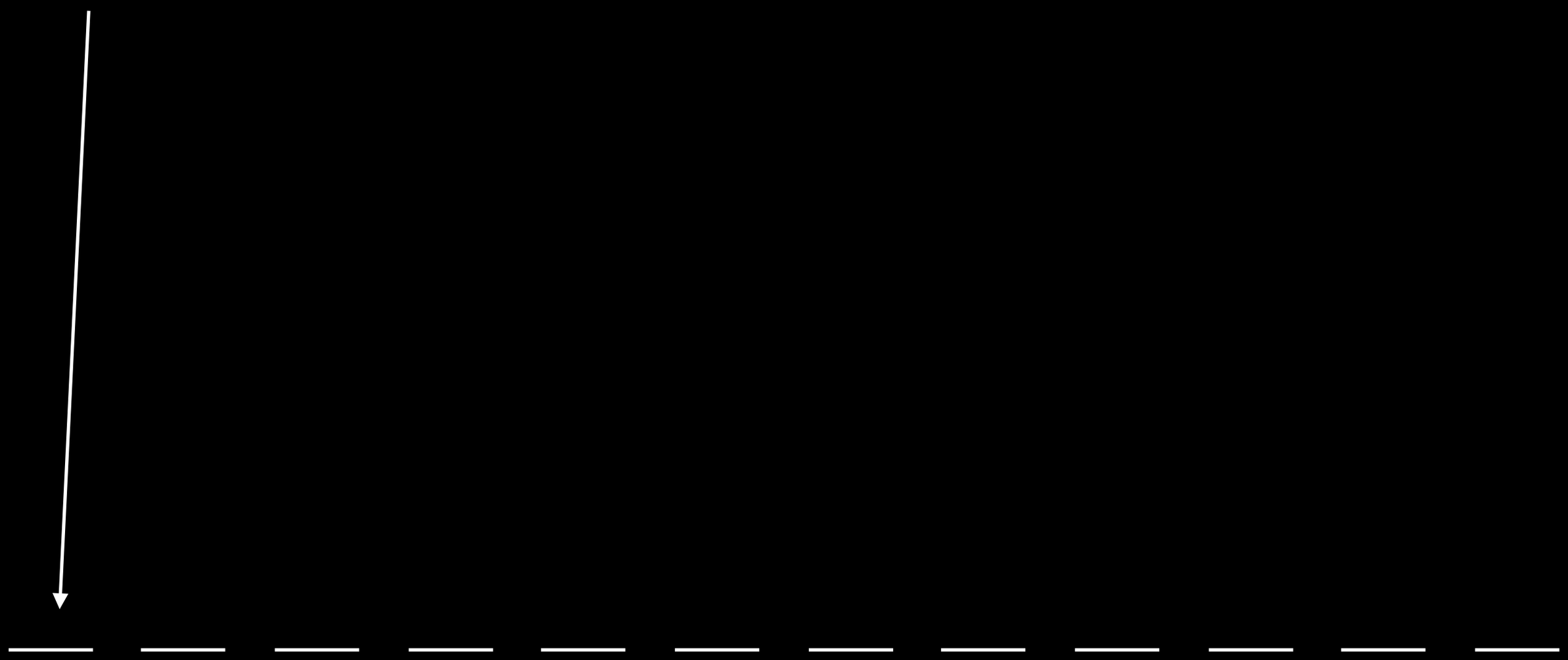
*Although north wind howls , but sky still very clear .*

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

# 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



However

# 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

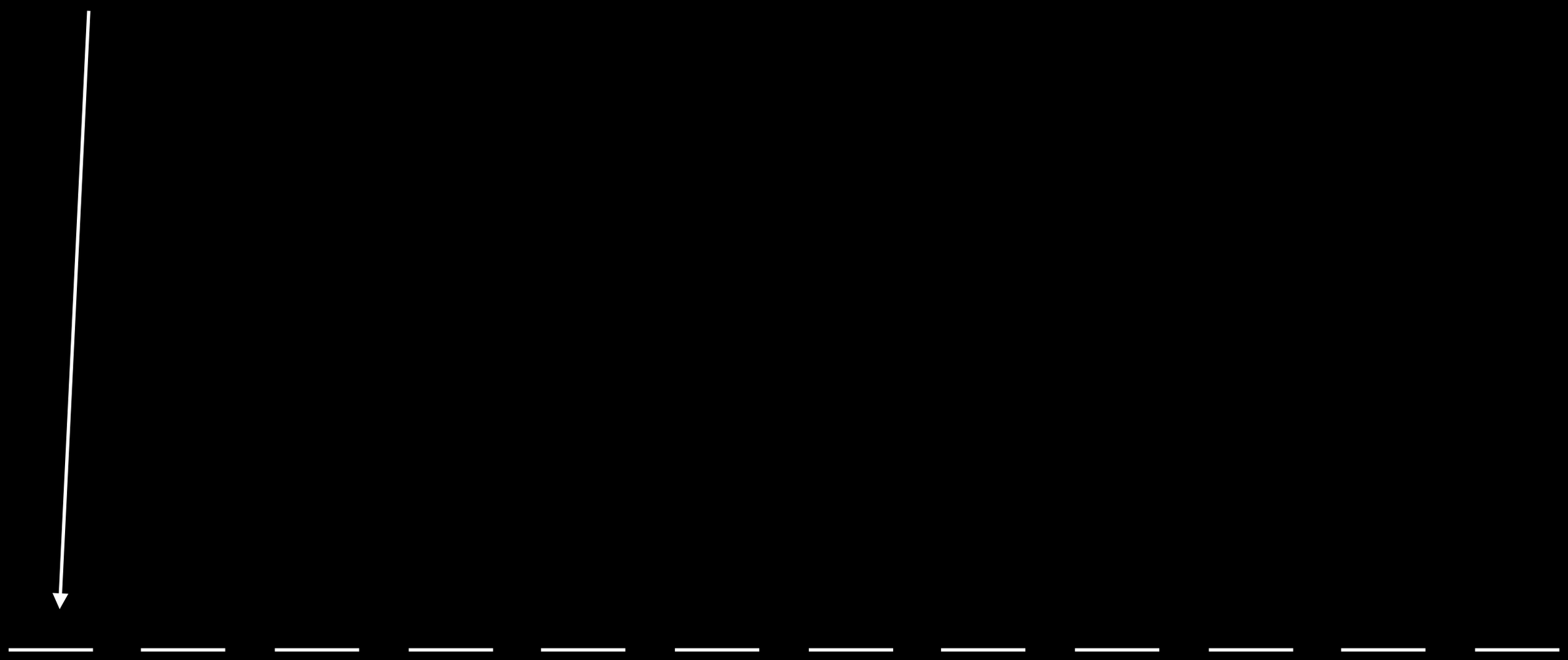
However

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

# 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



However

# Likelihood Estimation for Model 1

*Although north wind howls , but sky still very clear .*

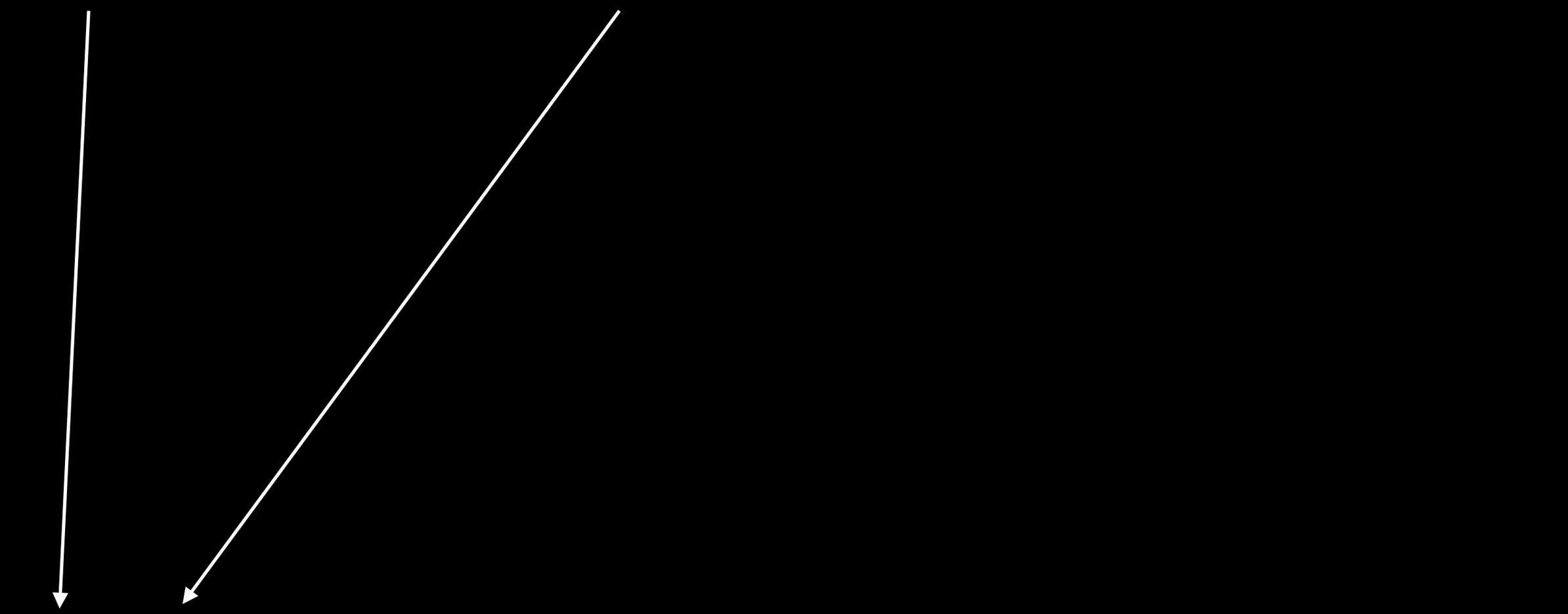
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However

# 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

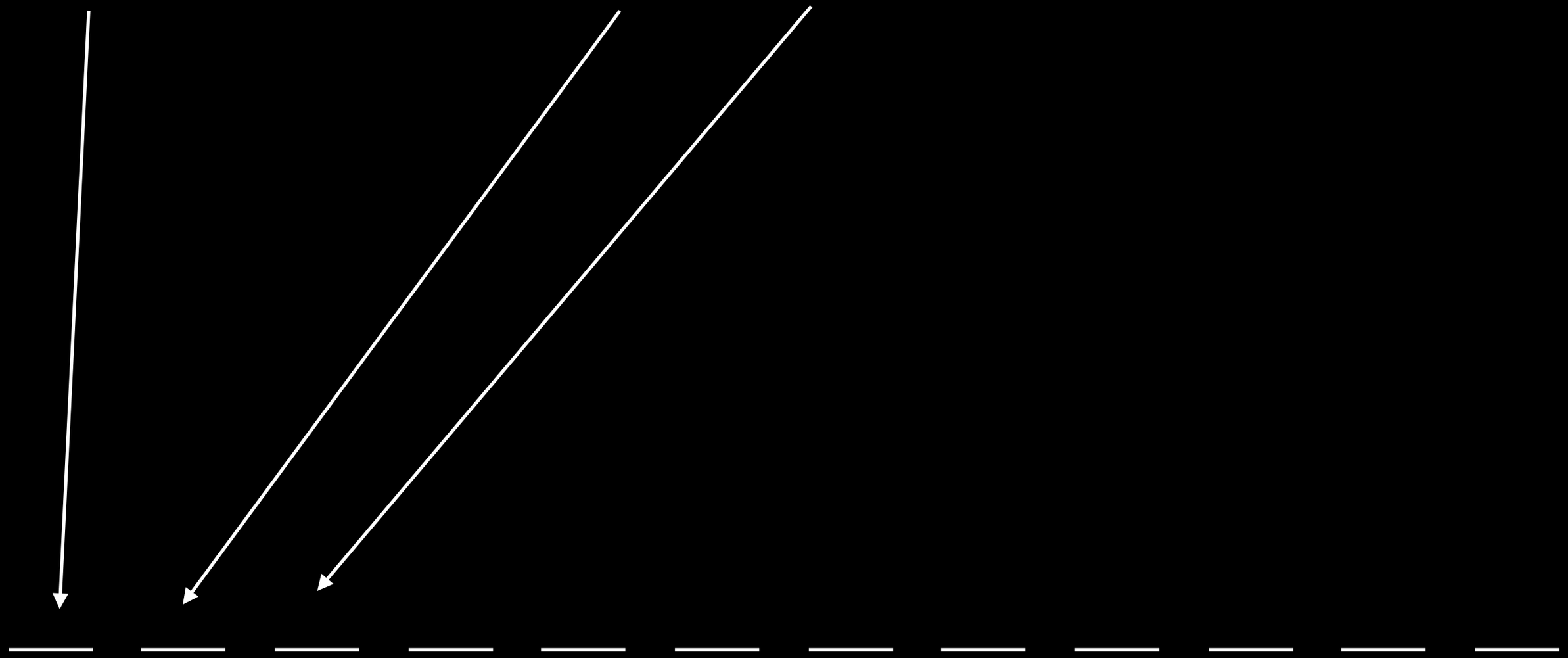


However ,

# 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

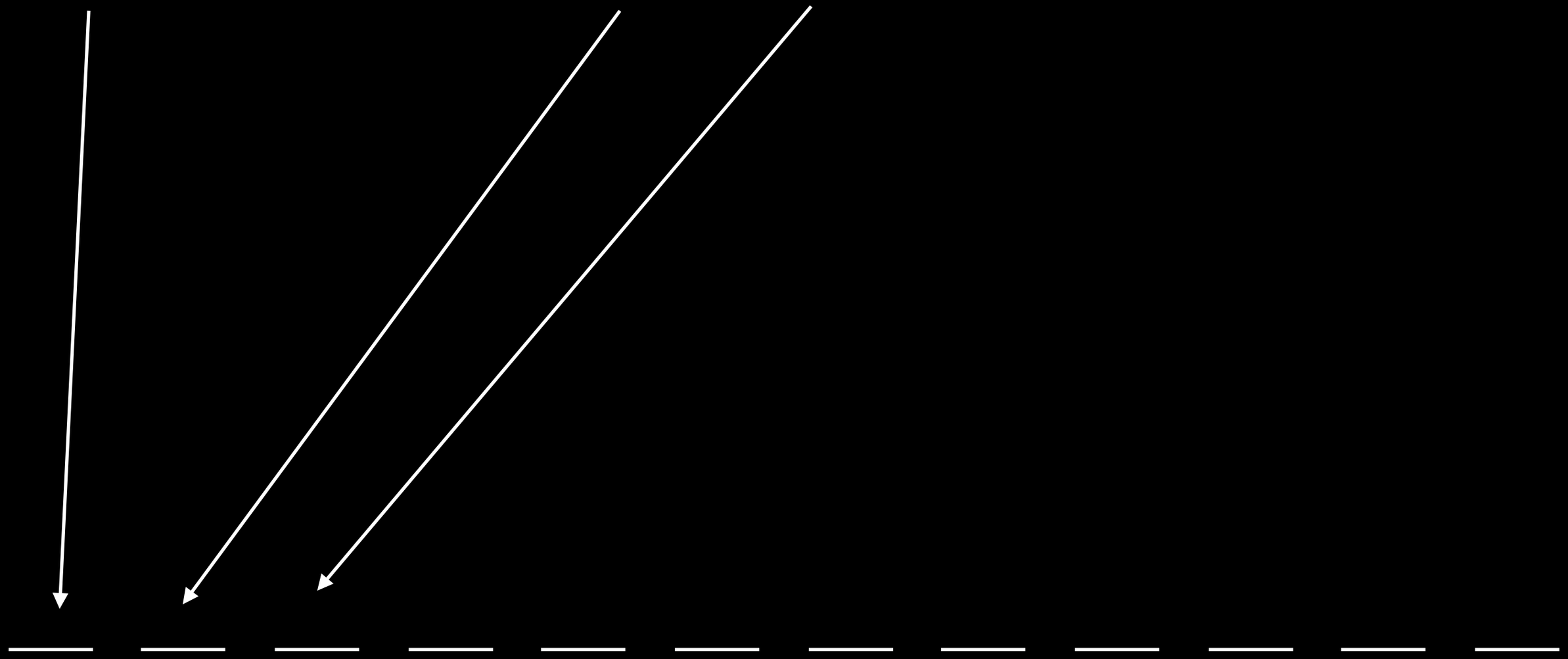


However ,

# 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

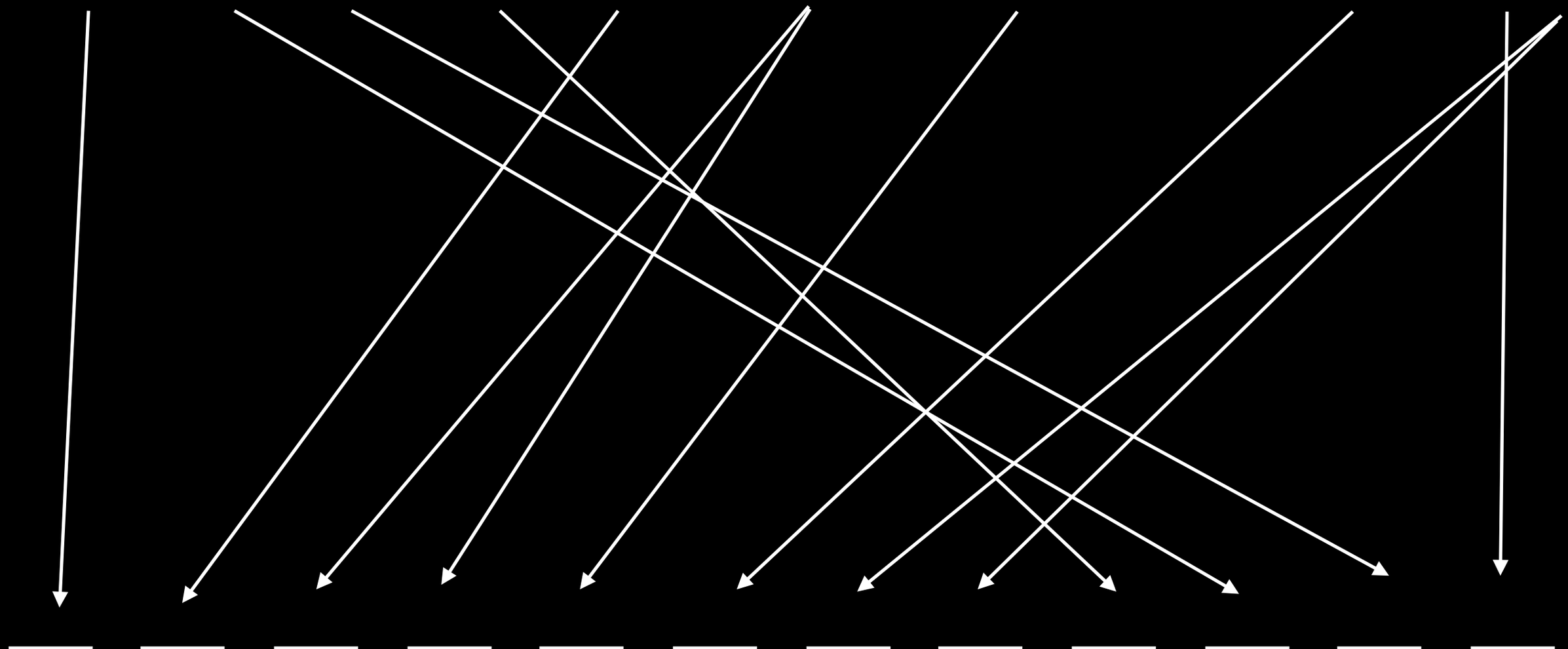


However , the

# 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

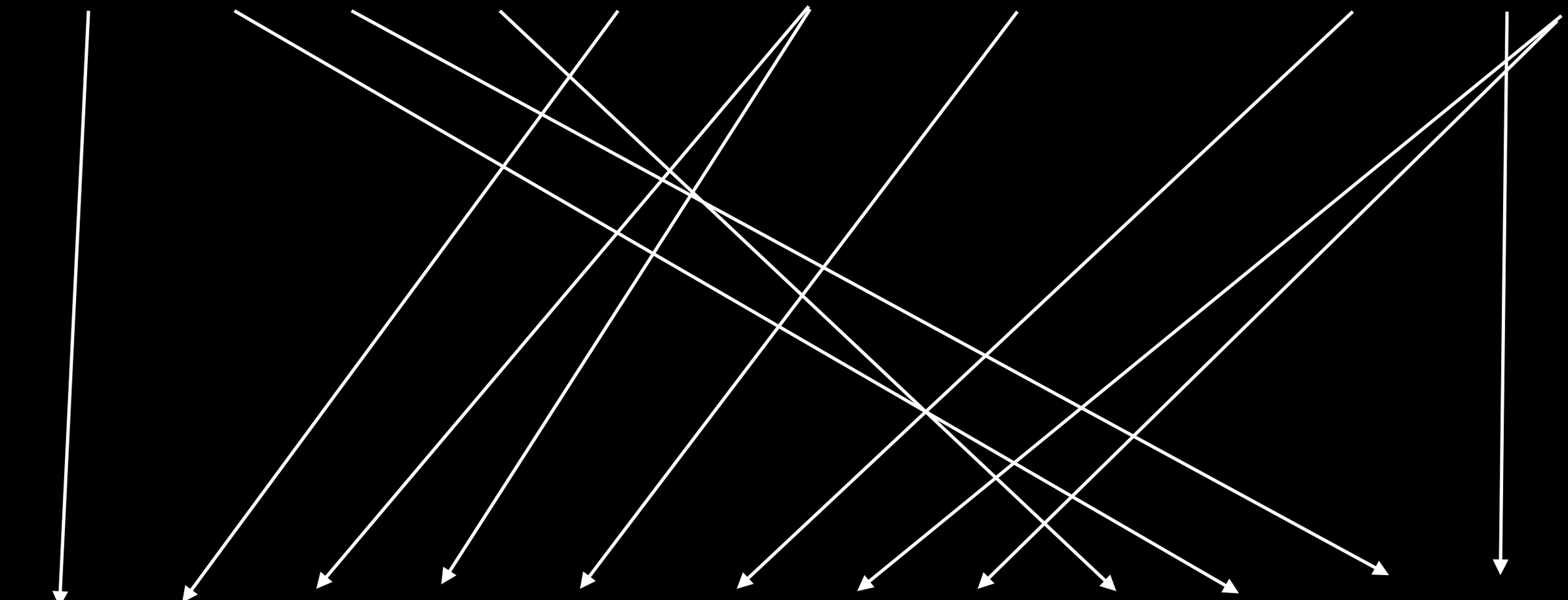


However , the sky remained clear under the strong north wind .

# 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



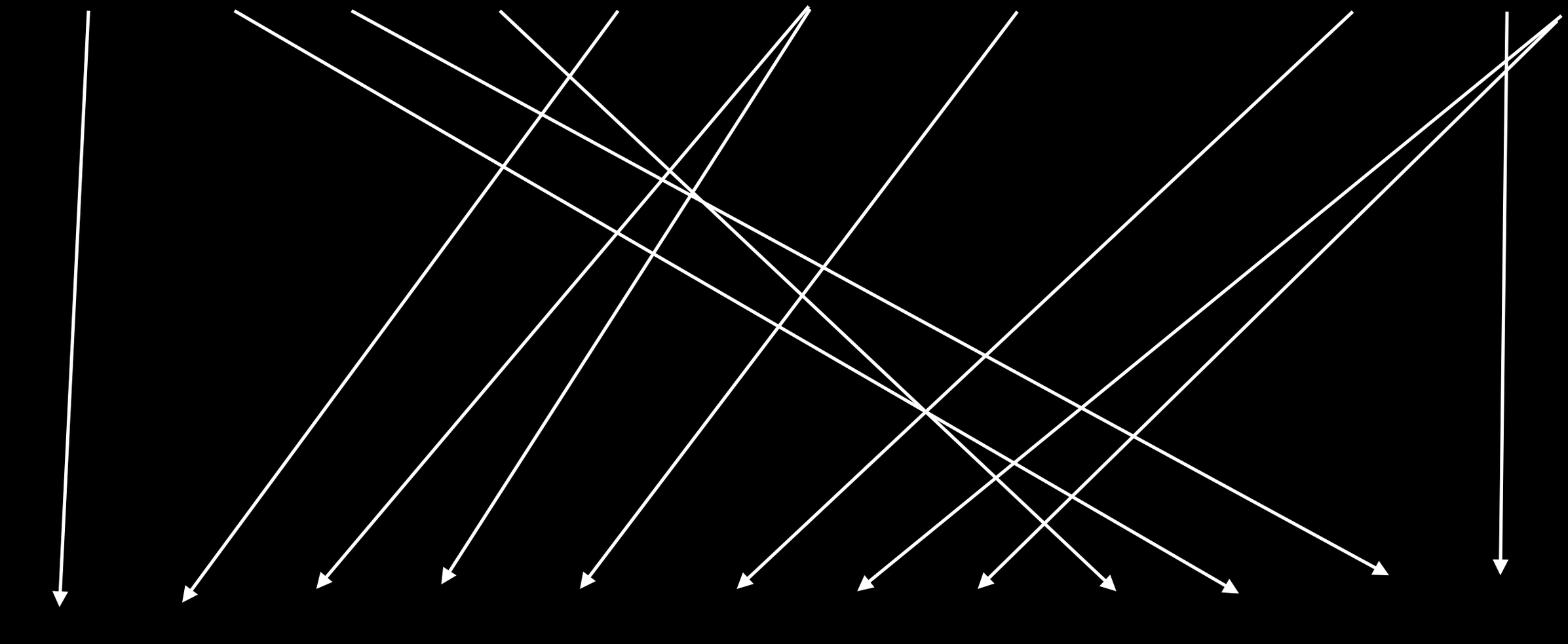
However , the sky remained clear under the strong north wind .

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

# 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



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

$p(Chinese\ word\ position)$ uniform, no need to estimate

# 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

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

the missing alignment is a ***latent variable***

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.

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 .*

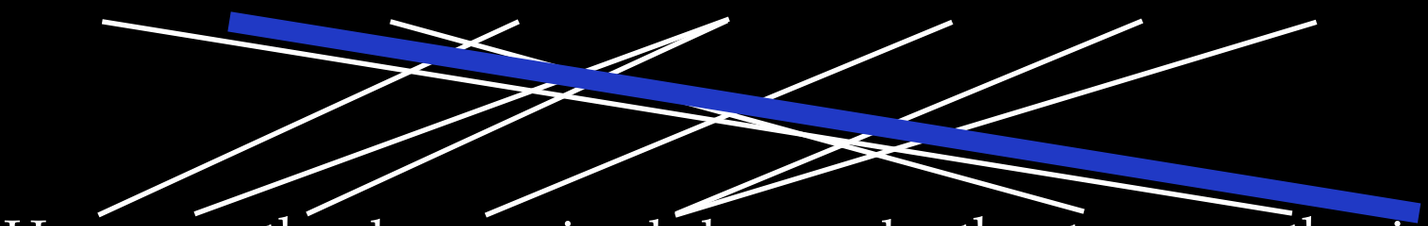
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

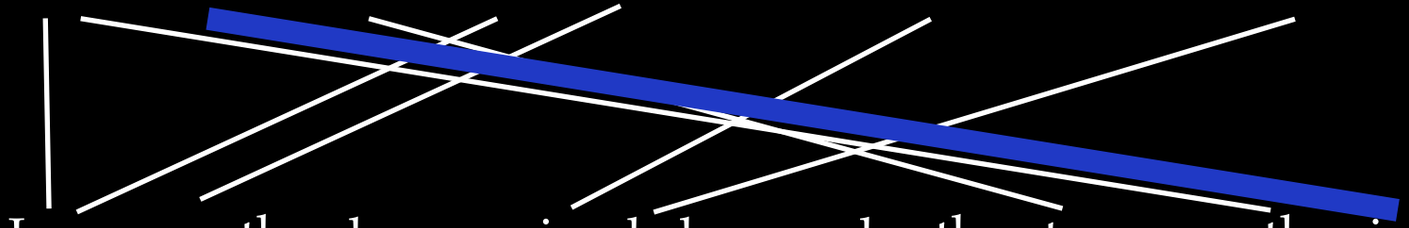
if we had observed the alignment, this line would either be here (count 1) or it wouldn't (count 0).

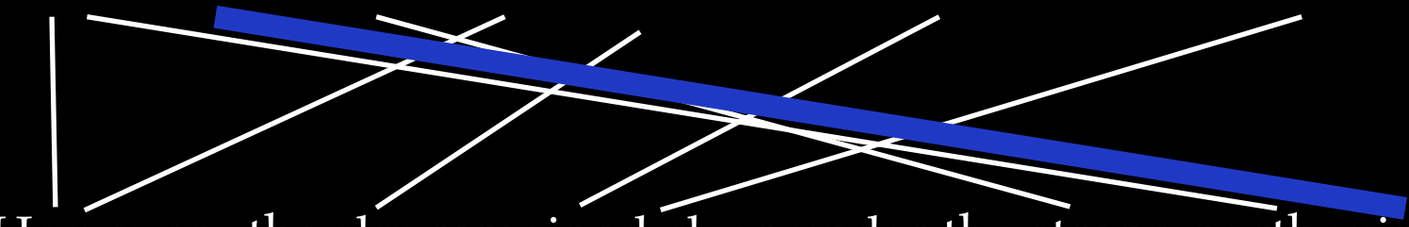
since we didn't observe the alignment, we calculate the probability that it's there.

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

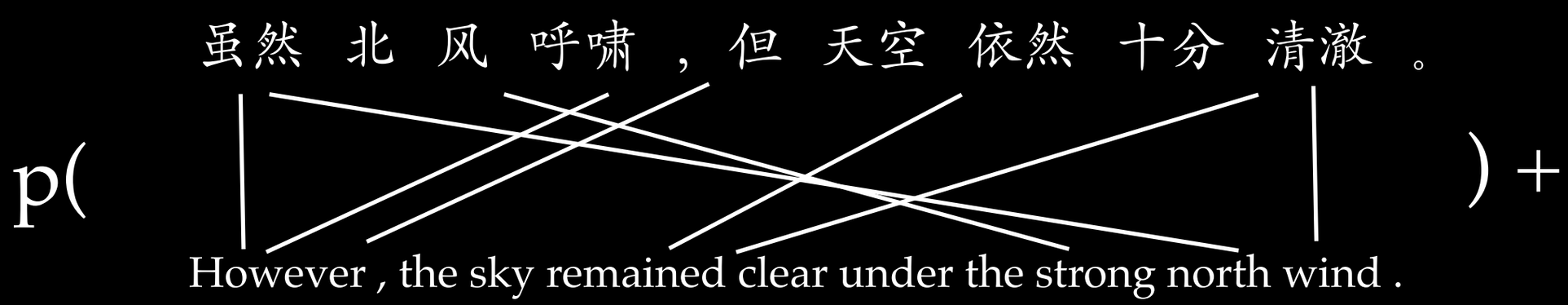
# Marginalize: sum all alignments containing the link

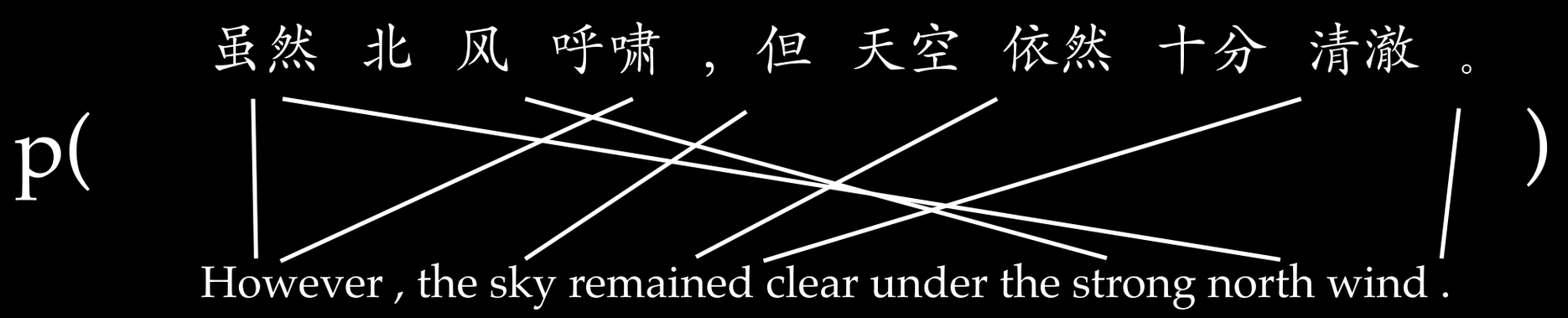
p( 
虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。
However , the sky remained clear under the strong north wind . ) +

p( 
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 . ) +

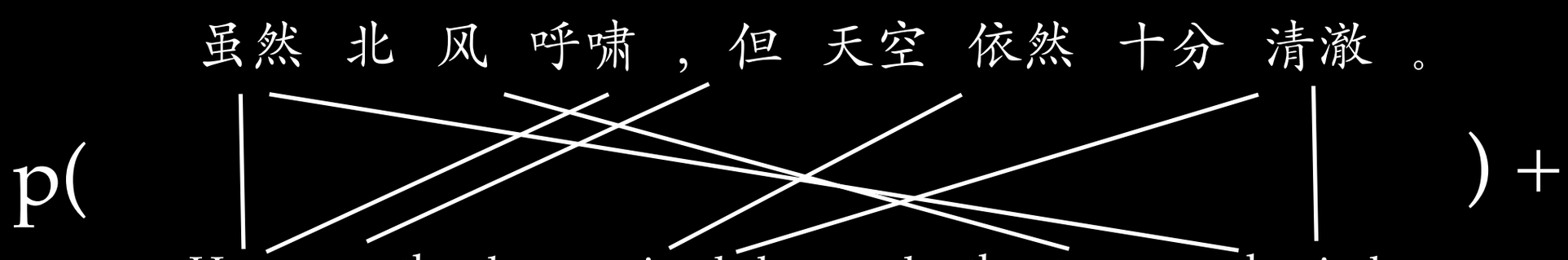
p( 
虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。
However , the sky remained clear under the strong north wind . )

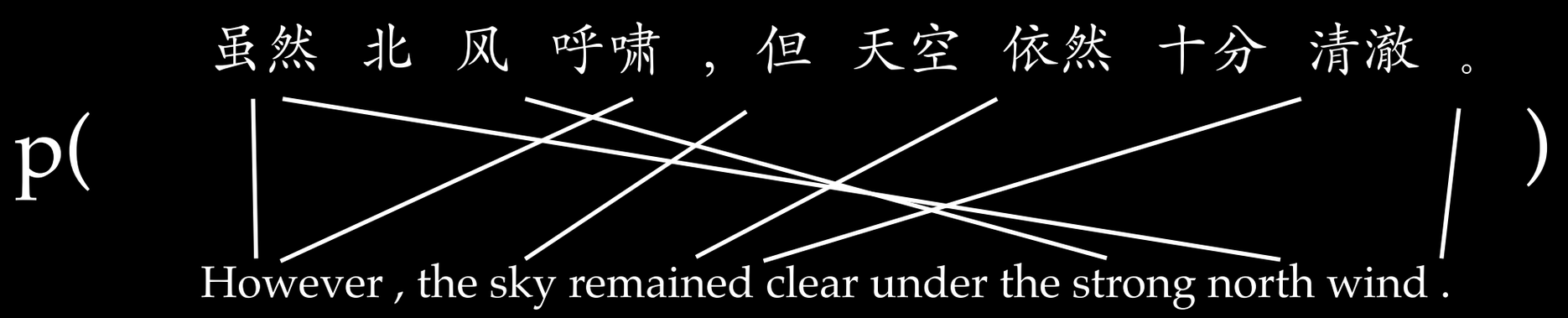
# Divide by sum of all *possible* alignments

# Divide by sum of all *possible* alignments


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.

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

$a_i$

observed

uniform

# Expectation Maximization

marginal probability of
alignments containing link

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

# Expectation Maximization

marginal probability of
alignments containing link

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

# Expectation Maximization

marginal probability of
alignments containing link

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

marginal probability of all
alignments

# Expectation Maximization

marginal probability of
alignments containing link

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

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

marginal probability of all
alignments

# Expectation Maximization

marginal probability of
alignments containing link

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

identical!

marginal probability of all
alignments

# 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

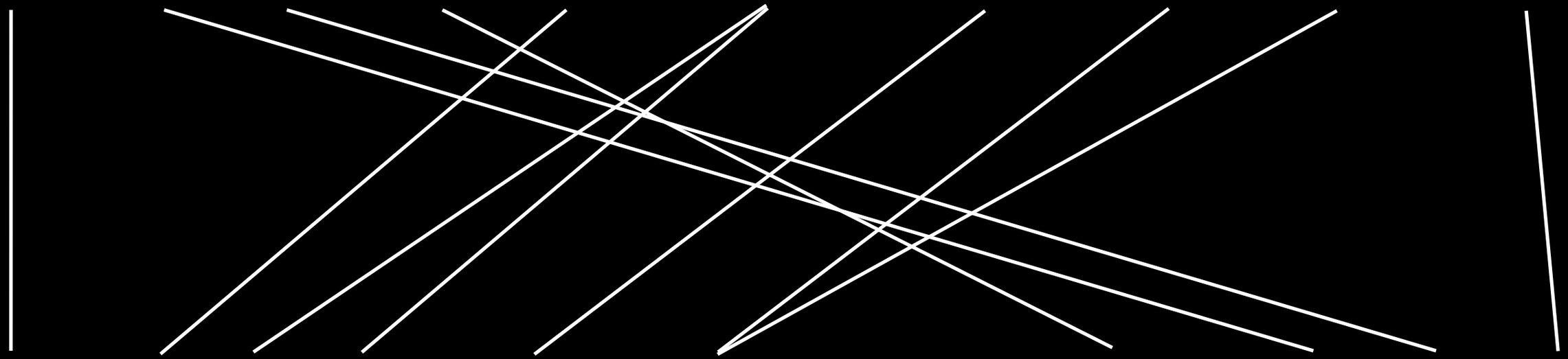
$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

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 .

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

# 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.

# Expectation Maximization

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

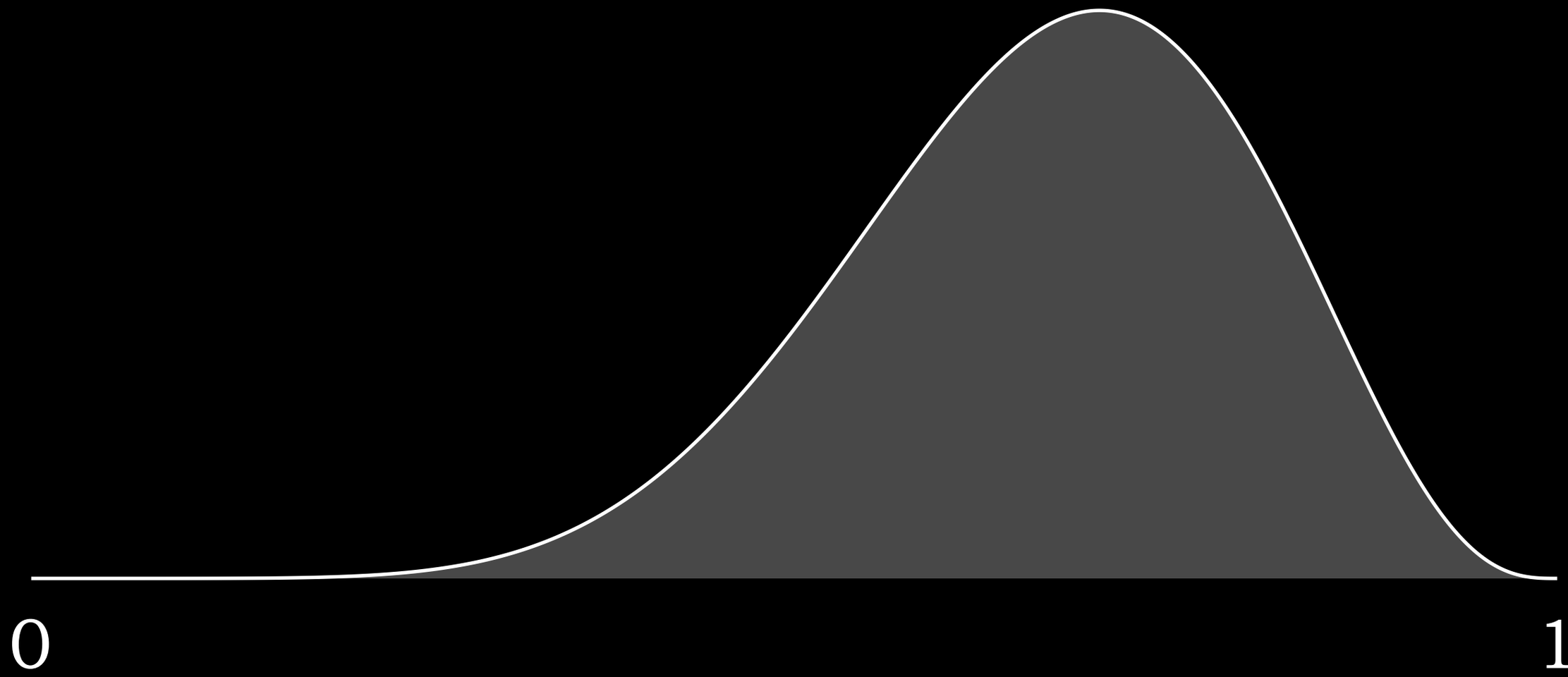
$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

MLE: choose parameters that maximize this expression.

Minor problem: there is no analytic solution.

Remember: likelihood is monotonically non-decreasing!

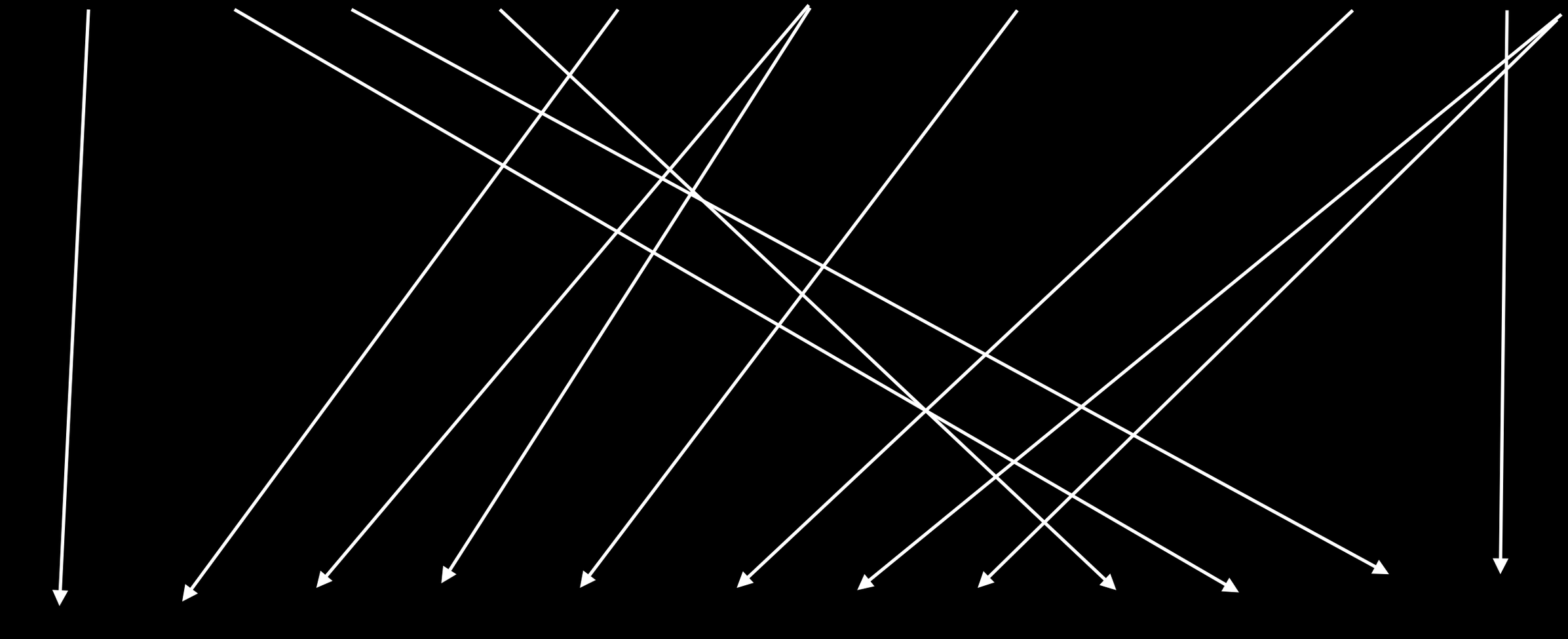
... and, likelihood is ***convex*** for this model:



# 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

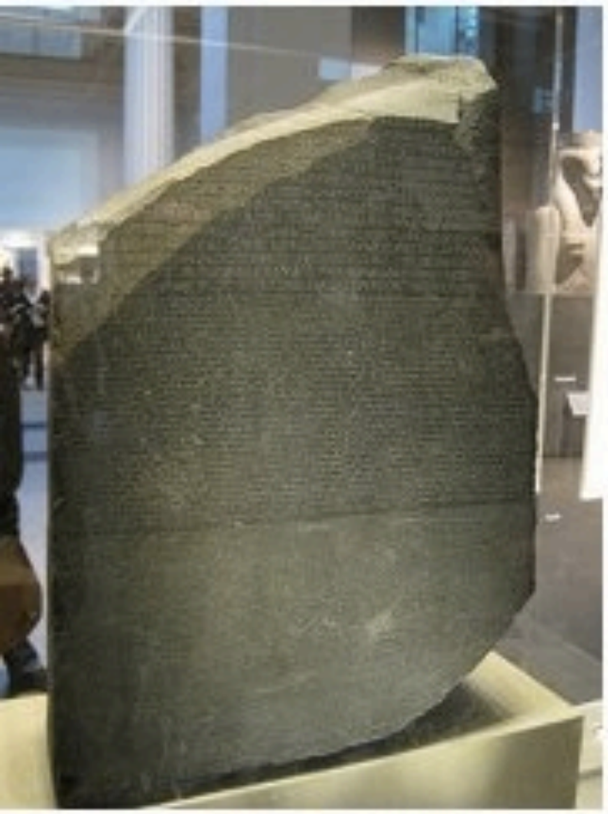


However , the sky remained clear under the strong north wind .

What are some things this model doesn’t account for?

# Summary

- We can formulate learning as an optimization problem: choose parameters that optimize some function, such as likelihood.
- Supervised: maximum likelihood.
  - Beware of overfitting.
- Unsupervised: expectation maximization.
- Many, many, many other algorithms.
- Thursday: learning better translation models.
  - Also: first “Language in 10 minutes”



*The Rosetta Stone*
*(image by calotype46)*

---

**Challenge 1**



# Alignment : Challenge Problem 1

Aligning words is a key task in data-driven machine translation. We start with a large **parallel corpus** where the translation of each sentence is known. So the input consists of sentence pairs like this one:

*le droit de permis passe donc de $ 25 à $ 500 . — we see the licence fee going up from $ 25 to $ 500 .*

It is relatively easy to obtain data in this form, though not completely trivial. This particular example is from the Canadian Hansards, proceedings of government meetings that are required to be published in both English and French. To align the sentences we can exploit document boundaries and structural cues arising from the fact that the documents are translations of each other, such as the fact that translated sentences will be in the same order, of similar length, and so forth. The problem is that we don't know the word-to-word correspondences in each sentence. That's where you come in: **your challenge is to write a program that aligns the words automatically**. For the sentence above, you might output the following correspondences:

*le — the, droit — fee, permis— license, passe—going, passe—up, donc—from, $ — $, 25 — 25, à — to, $ — $, 50 — 50*

Some words might be unaligned (e.g. *we* and *see*), while some words might be aligned to multiple words in the corresponding sentence (e.g. *passe* is aligned to *going up*). Sometimes words aren't exact translations of each other. That's ok: even experienced bilinguals will sometimes disagree on the best alignment of a sentence. But while word alignment doesn't capture every nuance, it is very useful for learning translation models or bilingual dictionaries.

## Getting Started

Run this command:

```
git clone https://github.com/alopez/dreamt.git
```