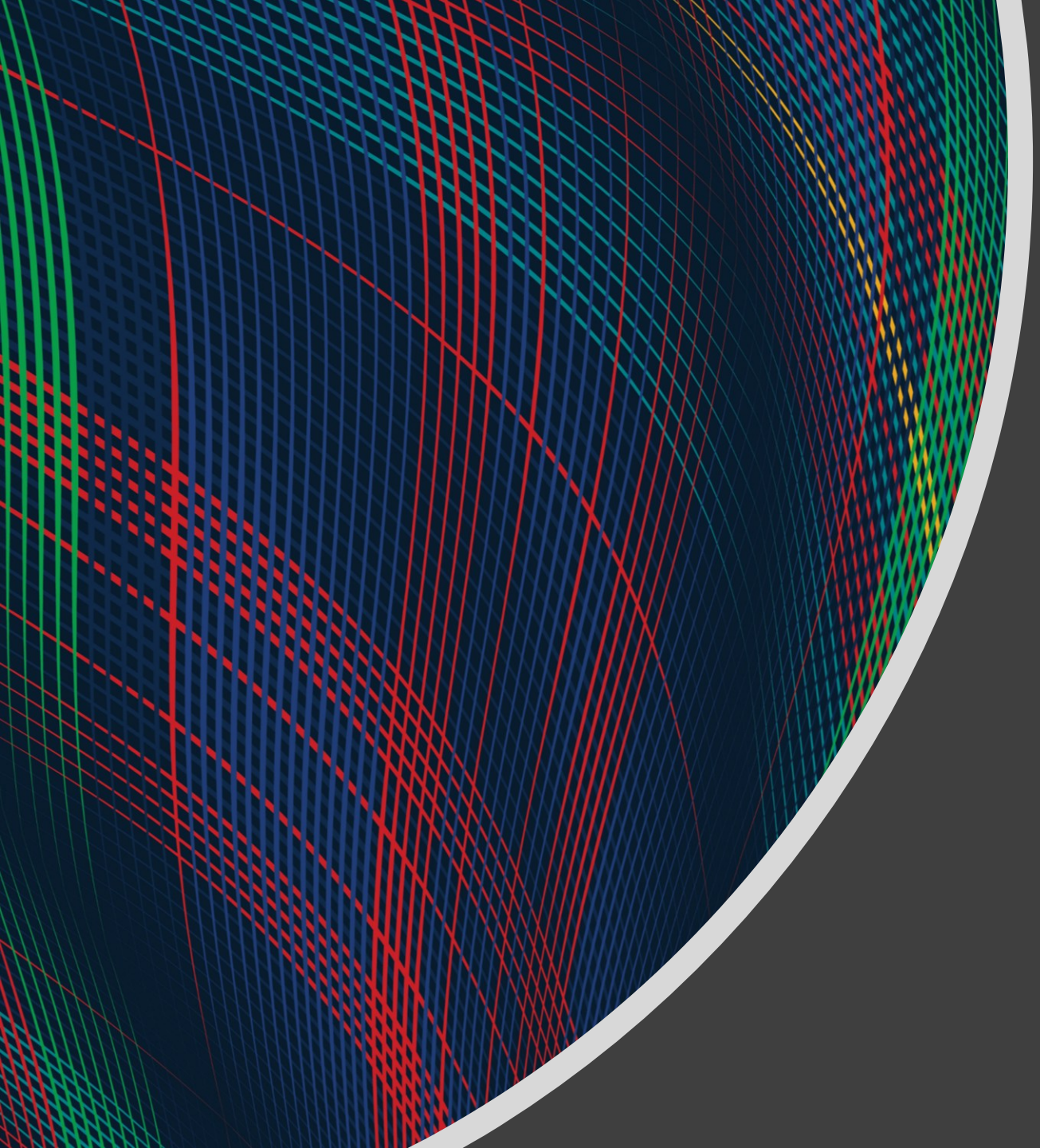


07-280
AI & ML I

NLP: N-gram LMs

Instructors:
Pat Virtue & Nihar Shah

Natural Language Processing (NLP)

Last time

MLE

NLP

- Tokenization
- N-grams Worksheet
- Language Models (LMs)

Today

Language Models

- N-gram LMs
- Training Data
- Sampling (Decoding) (Generation)

Feature Engineering for NLP

Feature Learning

- Word Embeddings
- Image and Audio Embeddings

What is a language model (LM)?

A language model is a model that approximates the joint probability distribution of a sequence of words (tokens) for a given language (usually trained on a tokenized corpus):

$$p(w_1, w_2, \ldots, w_T)$$

$p(w_1, w_2, w_3, w_4, \ldots, w_{10})$

$p(a, a, a, a, a \ldots a)$

$|V| = 8$

$|V|^T$

LM Applications: Speech Recognition

Example: Speech Recognition

“artificial ……”

Find most probable next word given “artificial” and the audio for second word.

Which second word gives the highest probability?

Break down problem

n-gram probability * audio probability

$P(\textbf{limb} \mid \text{artificial, audio})$

$P(\textbf{limb} \mid \text{artificial}) * P(\text{audio} \mid \textbf{limb})$

$P(\textbf{intelligence} \mid \text{artificial, audio})$

$P(\textbf{intelligence} \mid \text{artificial}) * P(\text{audio} \mid \textbf{intelligence})$

$P(\textbf{flavoring} \mid \text{artificial, audio})$

$P(\textbf{flavoring} \mid \text{artificial}) * P(\text{audio} \mid \textbf{flavoring})$

LM Applications: Text Generation

Given a LM that can model the joint distribution any 10-token sequence in the vocabulary, $p(w_1, w_2, \ldots, w_{10})$, how can we get the probability of a 6th token, given 5 previous tokens

$$p(w_6 \mid w_{1:5}) = \frac{p(w_{1:5}, w_6)}{p(w_{1:5})} \rightarrow p(w_{1:6})$$

$$p(w_{1:5}) = \sum_{w_6} p(w_{1:6})$$

$$=$$

$$p(w_{1:6}) = \sum_{w_{10}} \sum_{w_9} \sum_{w_8} \sum_{w_7} p(w_{1:10})$$

Language Models

N-grams

Poll 1

What is the probability of $p(mouse \mid with, a)$?

https://www.cs.cmu.edu/~pvirtue/AIS/ngrams/find-ngrams.html

$$\frac{N(\text{'with', 'a', 'mouse'})}{N(\text{'with', 'a', \_\_})}$$

Search for N-grams

Type tokens (words and/or punctuation) in the search box to find them in the text and see which word comes after them.

i am sam . i am sam . sam i am . that sam-i-am . that sam-i-am ! i do not like that sam-i-am do you like green eggs and ham i do not like them , sam-i-am . i do not like green eggs and ham . would you like them here or there ? i would not like them here or there . i would not like them anywhere . i do not like green eggs and ham . i do not like them , sam-i-am . would you like them in a house ? would you like them with a mouse ? i do not like them in a house . i do not like them with a mouse . i do not like them here or there . i do not like them anywhere . i do not like green eggs and ham . i do not like them , sam-i-am . would you eat them in a box ? would you eat them with a fox ? not in a box . not with a fox . not in a house . not with a mouse . i would not eat them here or there . i would not eat them anywhere . i would not eat green eggs and ham . i do not like them , sam-i-am . would you , could you , in a car ? eat them ! eat them ! here they are . i would not , could not , in a car . you may like them . you will see . you may like them in a tree ? i would not , could not in a tree . not in a car ! you let me be . i do not like them in a box . i do not like them with a fox i do not like them in a house i do not like them with a mouse i do not like them here or there . i do not like them

with a

FIND N-GRAMS

N = 3

19 results for with a

N-gram Worksheet

Computing N-grams

Unigram $p(w_t) = \frac{N(w_t)}{N}$ ← N = |Corpus|

Bigrams $p(w_t \mid w_{t-1}) = \frac{N(w_{t-1}, w_t)}{N(w_{t-1}, \bullet)}$

MLE to learn N-grams

$p(w_t \mid w_{t-1})$

What type of distribution?

- ~~Bernoulli~~
- Categorical
- ~~Gaussian (1-D)~~
- ~~Multivariate Gaus..~~

Normal

θ

Params

ϕ_1

ϕ_2

$\vdots$

ϕ_{dog} → [ ]

$\phi_{|V|}$

$w_t = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$|V| = 8$

$\sum_{k=1}^{|V|} \phi_k = 1$

MLE to learn N-grams

$$\sum_{k=1}^{8} \phi_k = 1$$

$$\vec{\phi} = \text{argmax} \prod p(w_t^{(i)} | \text{the}^{(i)}, \vec{\phi})$$

$$= \text{argmax} \sum_{i=1}^{N} \log p(w_t^{(i)} | \text{the}, \vec{\phi})$$

$$\log \phi_{dog} + \log \phi_{dog} + \log \phi_{dog} + \log \phi_{zoo} \cdots$$
$$+ \log \phi_{cat} + \log \phi_{dog} + \log \phi_{cat} + \log \phi_{zoo}$$

$$= \text{argmax} \; 4 \log \phi_{dog} + 2 \log \phi_{zoo} + 2 \log \phi_{cat}$$

$$\ell(\vec{\phi}; \mathcal{D})$$

$$\frac{\partial \ell}{\partial \vec{\phi}} = 0$$

$$\frac{N(w_t | \text{the})}{N(\text{the})}$$

```
[ 'the', 'dog', 'ran', '.', 'the',
 'dog', 'ate', '.', 'the', 'dog',
 'ran', 'the', 'zoo', '.', 'the',
 'cat', 'ate', 'the', 'dog', '!',
 'the', 'cat', 'ran', 'the',
 'zoo', '.' ]
```

N-gram Language Model

If we have unigram and bigram probabilities, do we have a language model??

$p(w_t)$ $p(w_t \mid w_{t-1})$

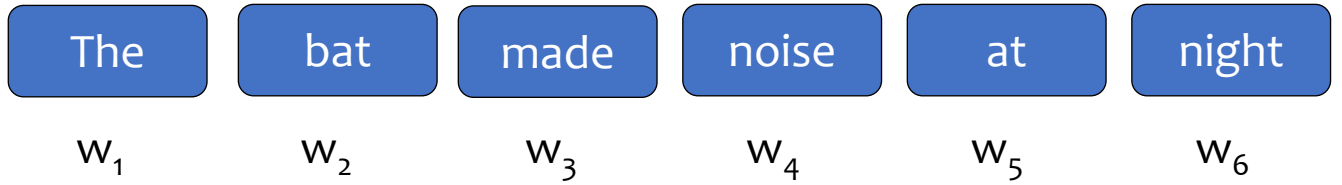
$p(w_1, w_2, w_3, w_4, w_5)$

the dog ate the zoo

$= p(w_1)\, p(w_2 \mid w_1)\, p(w_3 \mid w_2)\, p(w_4 \mid w_3)\, p(w_5 \mid w_4)$

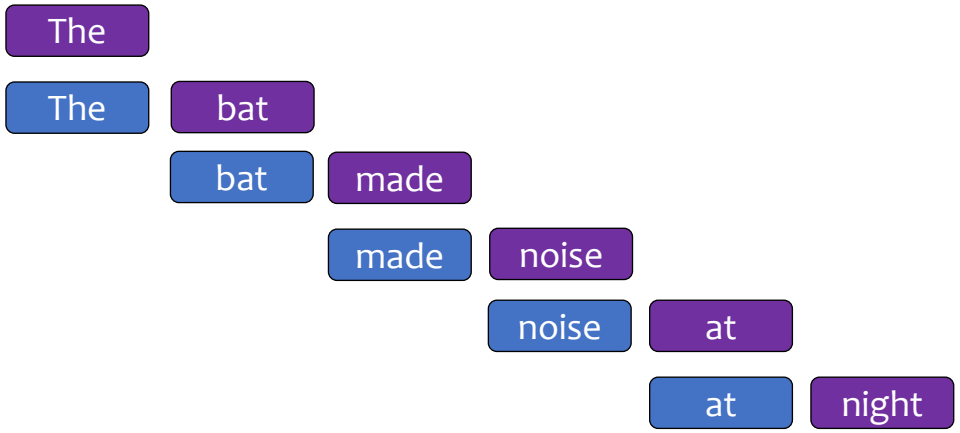
n-Gram Language Model

Question: How can we **define** a probability distribution over a sequence of length T?



n-Gram Model (n=2) $$p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_{t-1})$$

$p(w_1, w_2, w_3, \ldots, w_6) =$

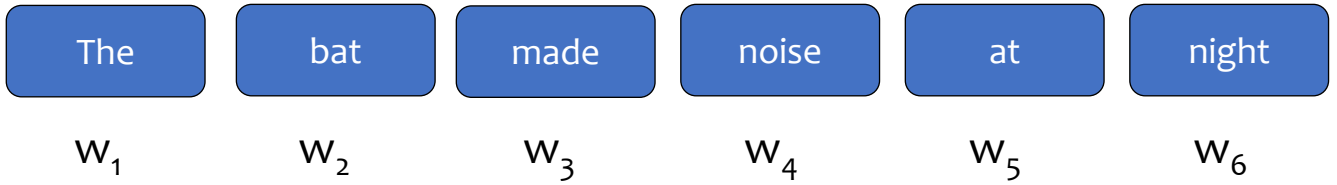


$p(w_1)$
$p(w_2 \mid w_1)$
$p(w_3 \mid w_2)$
$p(w_4 \mid w_3)$
$p(w_5 \mid w_4)$
$p(w_6 \mid w_5)$

$p(w_1)$
$p(w_2 \mid w_1)$
$p(w_3 \mid w_1, w_2)$
$p(w_4 \mid w_1, w_2, w_3)$

n-Gram Language Model

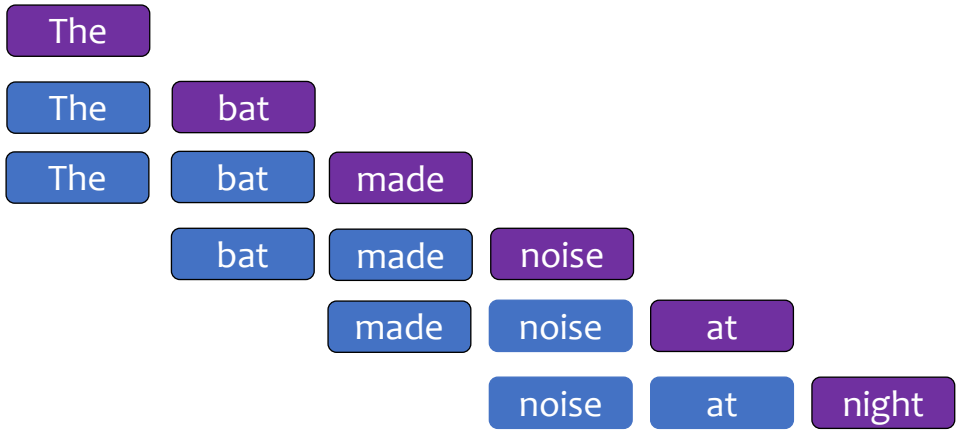
Question: How can we **define** a probability distribution over a sequence of length T?



n-Gram Model (n=3)

$$p(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} p(w_t \mid w_{t-1}, w_{t-2})$$

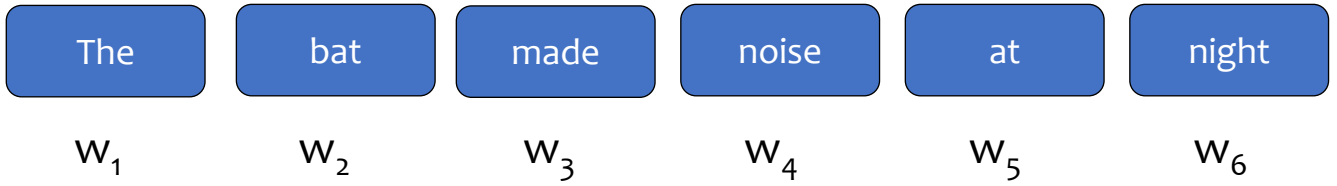
$p(w_1, w_2, w_3, \ldots, w_6) =$



$p(w_1)$
$p(w_2 \mid w_1)$
$p(w_3 \mid w_2, w_1)$
$p(w_4 \mid w_3, w_2)$
$p(w_5 \mid w_4, w_3)$
$p(w_6 \mid w_5, w_4)$

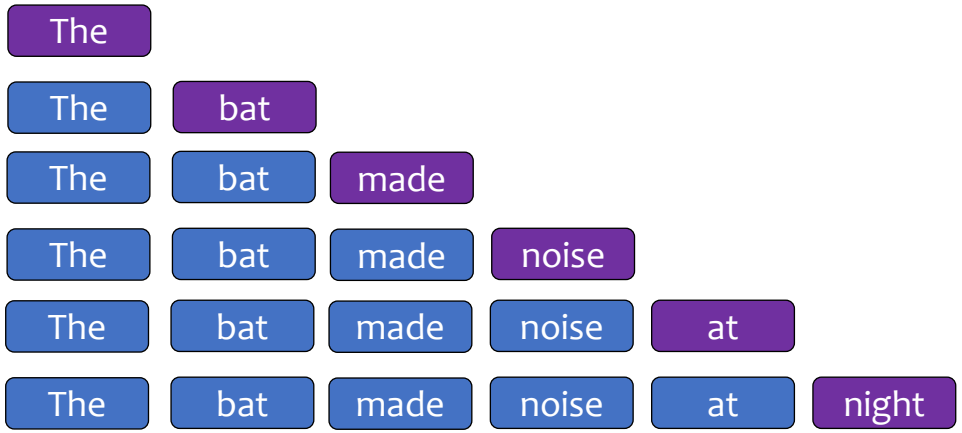
The Chain Rule of Probability

Question: How can we **define** a probability distribution over a sequence of length T?



w/o assumption

$$p(w_1, w_2, w_3, \dots, w_6) =$$



$$p(w_1)$$
$$p(w_2 \mid w_1)$$
$$p(w_3 \mid w_2, w_1)$$
$$p(w_4 \mid w_3, w_2, w_1)$$
$$p(w_5 \mid w_4, w_3, w_2, w_1)$$
$$p(w_6 \mid w_5, w_4, w_3, w_2, w_1)$$

Markov Chains

N-gram Markov assumption

First order Markov Chain

Markov chain

$X_1 \rightarrow X_2 \rightarrow \ldots X_{t-1} \rightarrow X_t \rightarrow X_{t+1} \rightarrow \ldots X_T$

write joint as product of ~~$p(X_t | X_{t-1})$~~ $p(X_{t+1} | X_t)$

Assumption: 果 $X_{t-1} \perp\!\!\!\perp X_{t+1} \mid X_t$

$X_{t+1} \perp\!\!\!\perp X_1 \mid X_t$

$p(X_{t+1} | X_t, X_{1:t-1}) = p(X_{t+1} | X_t)$ X_2'

LM Training Data

N-gram Training

Where do the n-gram probabilities come from?

Google n-grams demo

N-gram Training Data

Unsupervised

Self-supervised learning (auto-regressive)

Example: Jane Austen, *Pride and Prejudice*

Vanity and pride are different things, though the words are often used synonymously. A person may be proud without being vain. Pride relates more to our opinion of ourselves, vanity to what we would have others think of us.

N-gram Training Data

Self-supervised learning (auto-regressive)

Example: Jane Austen, *Pride and Prejudice*

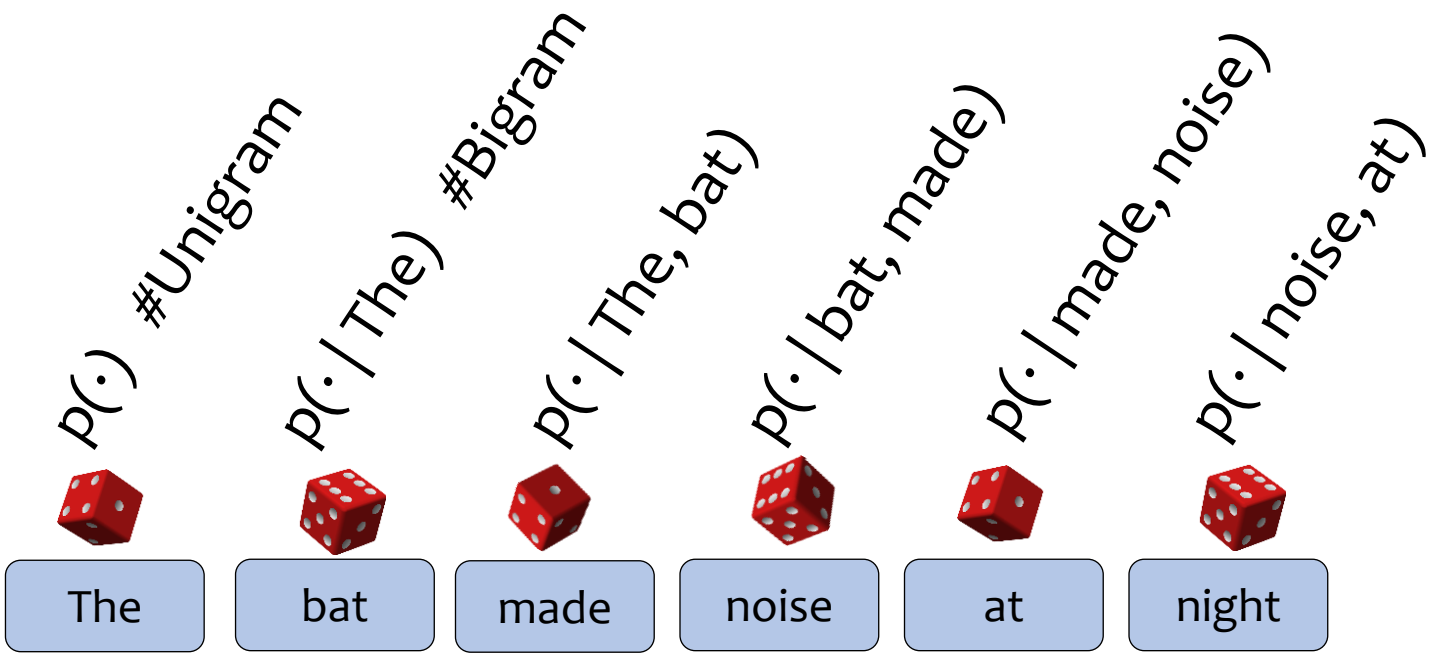
Vanity and pride are different things, though the words are often used synonymously. A person may be proud without being vain. Pride relates more to our opinion of ourselves, vanity to what we would have others think of us.

Sampling from a Language Model

Sampling from a Language Model

How do we sample from a language model?

1. Treat each probability distribution like a (50k-sided) weighted die
2. Pick the die corresponding to $p(w_t \mid w_1, \ldots, w_{t-1})$
3. Roll that die and generate whichever word w_t lands face up
4. Repeat



Slide: CMU, Matt Gormley

N-gram Examples

Random samples from language model trained on Shakespeare:

n=1: "in as , stands gods revenge ! france pitch good in fair hoist an what fair shallow-rooted , . that with wherefore it what a as your . , powers course which thee dalliance all"

n=2: "look you may i have given them to the dank here to the jaws of tune of great difference of ladies . o that did contemn what of ear is shorter time ; yet seems to"

n=3: "believe , they all confess that you withhold his levied host , having brought the fatal bowels of the pope ! ' and that this distemper'd messenger of heaven , since thou deniest the gentle desdemona ,"

n=7: "so express'd : but what of that ? 'twere good you do so much for charity . i cannot find it ; 'tis not in the bond . you , merchant , have you any thing to say ? but little"

This is starting to look a lot like Shakespeare… because it is Shakespeare

Slide: CMU, Zico Kolter

N-gram LM Issues

What are the downsides of N-grams for language models?

Sampling from a Language Model

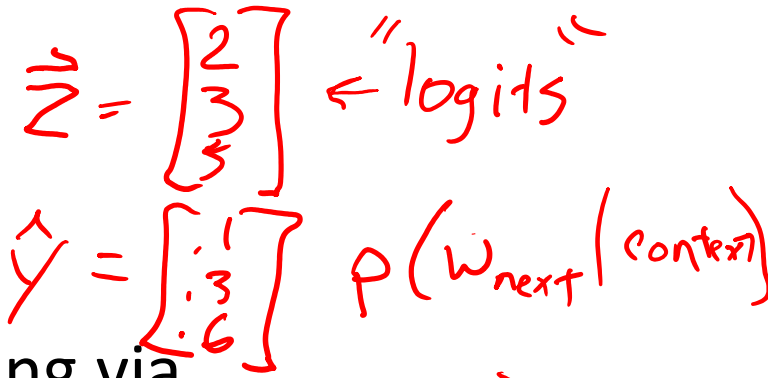
Variations

- Greedy (argmax) versus sampling

- Top-K

- Temperature

Sampling from a Language Model



Temperature sampling

Allows us to adjust the level randomness during sampling via hyperparameter T (Temperature):

- T= 1.0: Use the original probabilities (default)
- T > 1.0: "Softer" distribution → more random, diverse outputs
- T < 1.0: "Sharper" distribution → more predictable, focused outputs
- T → 0: Effectively becomes greedy (always pick the most likely token)

Adjusted probability:

$$P_{temperature}(w_i) = \frac{e^{\log P(w_i)/T}}{\sum_j e^{\log P(w_i)/T}}$$