

# Lecture reviews — Week 04

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle  
Faculté I&C

# Purpose of these lecture reviews

- ▶ Improve/deepen your learning
- ▶ Answer your questions
- ▶ Save you practice/revision time

Why are these sessions not recorded?

1. the intention is to have *appropriate/adapted/personalized* face-to-face interaction
2. recording them would lead to an extra 2 hours/week video lecture (which is too much *passive* content)

# Content

1. Big picture:  
What did you retain? What keypoints do you remember?
2. Questions?
3. More examples

# Week 4 keypoints

- Words / tokens (def, pay attention, ...)
- n-grams
  - ↳ char
  - ↳ tokens : language models
- learn/train : parameters :  $P(x_1 \dots x_n)$
- usage :  $\prod_{i=1}^n P(x_i | x_{i-1} \dots x_{i-n})$
- learning: MLE vs smoothing (add  $\alpha$ )
- DoV forms

# Week 4 keypoints

- ▶ Words vs. tokens
- ▶  $n$ -gram models
- ▶ MLE and add-one smoothing are bad (in NLP)
- ▶ Language Identification
- ▶ Out-of-Vocabulary forms:
  - ▶ OoV forms do matter
  - ▶ 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors

usage vs training

$$\prod_{i=1}^n P(X_i | X_{i-(n-1)} \dots X_{i-1})$$

Diagram illustrating the usage of the joint probability distribution. A green arrow above the expression indicates the length of the sequence is  $n$ . A red arrow below the conditioning variables indicates the length of the conditioning sequence is  $n-1$ . A blue bracket underneath the entire expression indicates it represents the joint probability of the sequence.

$$P(X_{i-(n-1)} \dots X_{i-1} X_i)$$

---


$$\boxed{P(X_{i-(n-1)} \dots X_{i-1} X_i)} = \sum_x P(X_{i-(n-1)} \dots X_{i-1} x)$$

$$f_{\theta}(x)$$

↓

$$P(X_{i-(n-1)} \dots X_i)$$

Diagram illustrating the training process. A red function  $f_{\theta}(x)$  is shown above a blue arrow pointing down to the joint probability expression  $P(X_{i-(n-1)} \dots X_i)$ . A blue arrow below the expression indicates the length of the sequence is  $n$ .

# Week 4 keypoints

- ▶ Words vs. tokens
- ▶  $n$ -gram models
- ▶ MLE and add-one smoothing are bad (in NLP)
- ▶ Language Identification
- ▶ Out-of-Vocabulary forms:
  - ▶ OoV forms do matter
  - ▶ 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors

Questions?

# Week 4 review example

Take a random Wikipedia page (e.g. <https://en.wikipedia.org/wiki/ACVRL1>) and compare two phrases using 3-grams (of tokens).

For instance:

This gene encodes a type **1** receptor

and

This gene encodes a type **2** receptor

language model (of size 3)

most probable?

$$= A \cdot P(2 | a \text{ type}) \cdot P(\text{receptor} | \text{type } 2)$$

A

$$P(\text{This gene encodes}) \cdot P(a | \text{gene encodes})$$
$$\cdot P(\text{type} | \text{encodes } a) \cdot P(I | a \text{ type}) \cdot P(\text{receptor} | \text{type } I)$$



# Week 4 review example

Take a random Wikipedia page (e.g. <https://en.wikipedia.org/wiki/ACVRL1>) and compare two phrases using 3-grams (of tokens).

For instance:

*This gene encodes a type 1 receptor*

and

*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?

$$P(xyz)$$

$$P(xy)P(z|xy).$$

1. where do we start from  
2. tokenize

3. search  $P(\text{token}_i, \text{token}_{i+1}, \text{token}_{i+2})$

↳ This vs this

↓  
this | Beg of Sent

This  
≡ <BOS>|this

learning  $\langle \text{Bos} \rangle$  ~~serine~~ / theonomic

1st param:  $P(\langle \text{Bos} \rangle, \text{serine}, /)$

$$= \frac{1 + \alpha}{N + \alpha \cdot M}$$

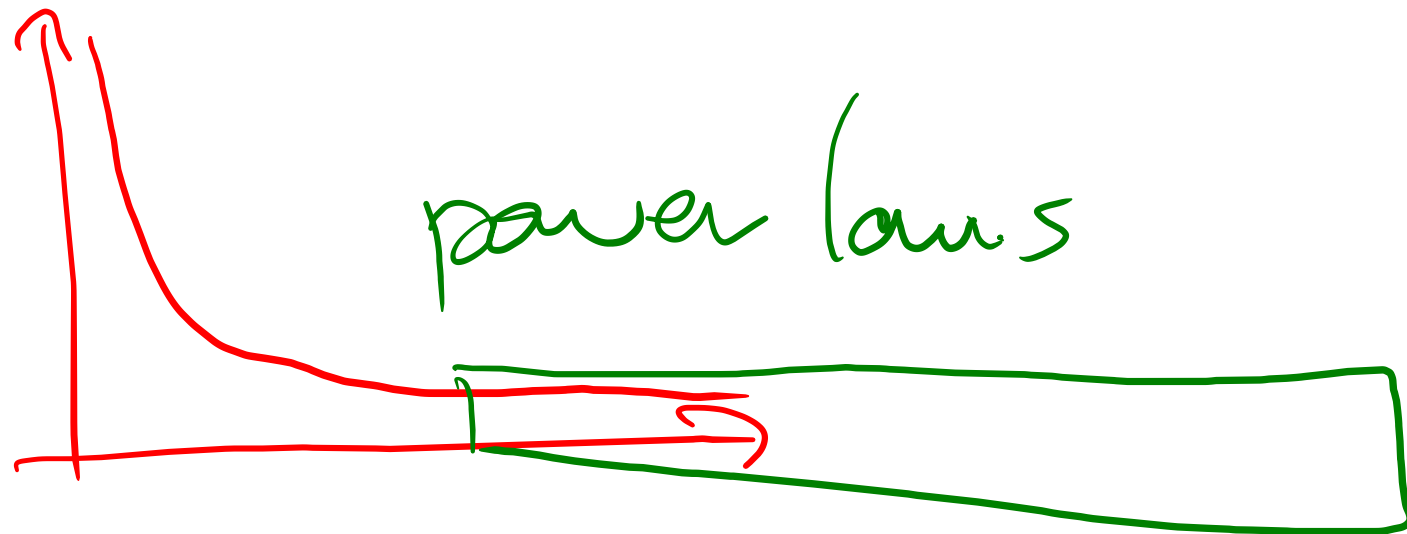
! ~~MLE?~~

$\downarrow$   
# 3-gram  
in corpus

$\searrow$   
# of possible  
3-grams  
(huge!)

MLE

are bad  
with



# Week 4 review example

Take a random Wikipedia page (e.g. <https://en.wikipedia.org/wiki/ACVRL1>) and compare two phrases using 3-grams (of tokens).

For instance:

*This gene encodes a type 1 receptor*

and

*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)

# Week 4 review example

Take a random Wikipedia page (e.g. <https://en.wikipedia.org/wiki/ACVRL1>) and compare two phrases using 3-grams (of tokens).

For instance:

*This gene encodes a type 1 receptor*

and

*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)
3. How to deal with upper-/lowercase? (e.g. “*This*”)

# Week 4 review example

Take a random Wikipedia page (e.g. <https://en.wikipedia.org/wiki/ACVRL1>) and compare two phrases using 3-grams (of tokens).

For instance:

*This gene encodes a type 1 receptor*

and

*This gene encodes a type 2 receptor*

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)
3. How to deal with upper-/lowercase? (e.g. “*This*”)
4. What estimates? (MLE? Smoothing?)