Lecture reviews — Week 04

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Purpose of these lecture reviews

- Improve/deepend 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

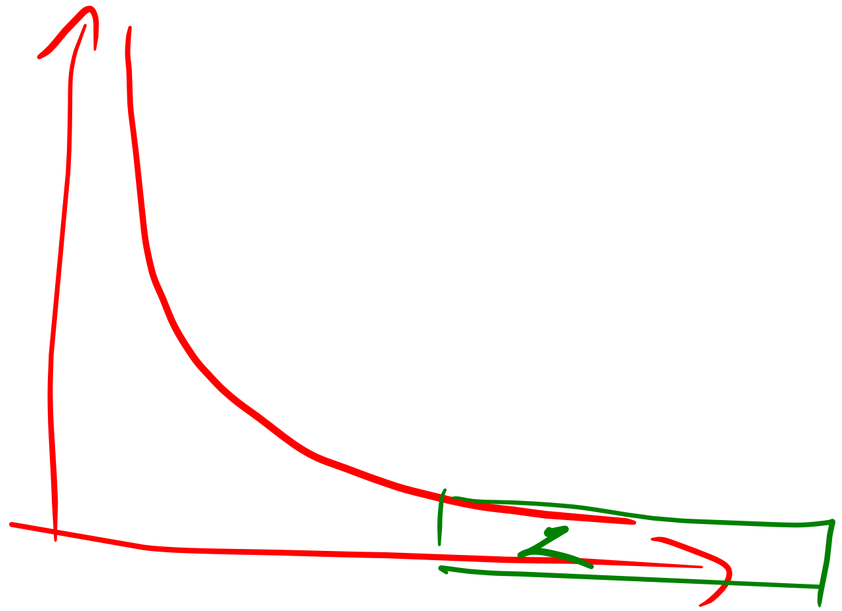
1. Big picture:
   What did you retain? What keypoints do you remember?
2. Questions?
3. More examples

Week 4 keypoints

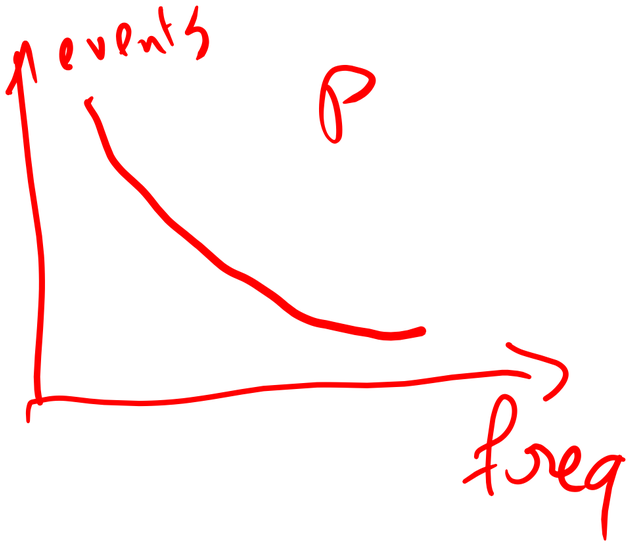
- what unit for NLP → token / word

- n-grams → chars / tokens
  - L.M.
  - Proba (current | preceed)
    - parameters
      - estimation : → MLE / → add α (Dirichlet prior)
      - OOV forms

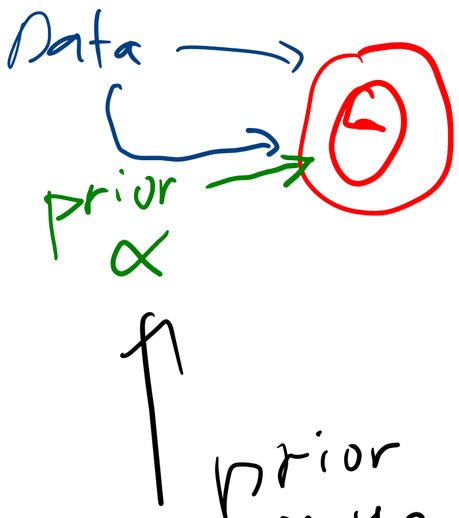
Week 4 keypoints



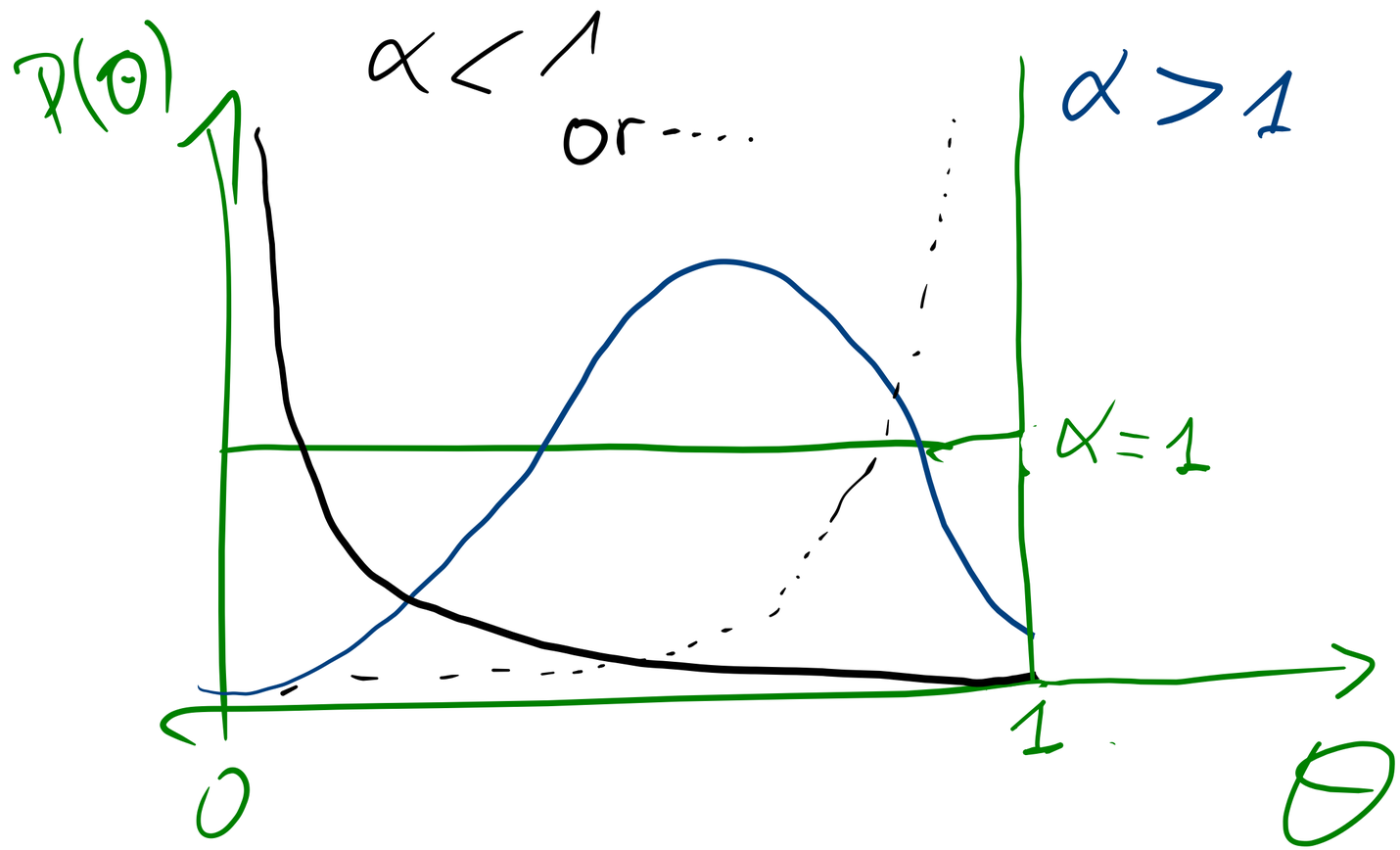
- Words vs. tokens
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)
- Language Identification
- Out-of-Vocabulary froms:
  - OoV forms do matter
  - 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors



$p_{\text{estimate}}(\text{params})$



prior on the prior



θ param

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

Week 4 keypoints

- Words vs. tokens
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)
- Language Identification
- Out-of-Vocabulary froms:
  - 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 I receptor
and
This gene encodes a type 2 receptor

P(This gene encodes) · P(a | gene encodes) · Ptype(encodes a) · .....

↳ from parameters:

$$\frac{P(\text{gene encodes a})}{\sum_x P(\text{gene encodes } x)}$$

1 param

other params

This ≡ this | <BOS>

estimation:

3-gram

$$P(\text{gene encodes a}) = \frac{1 + \alpha}{K - 2 + \alpha N^3}$$

Dirichlet prior

K tokens

N : # possible tokens

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 I 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?)

J.-C. Chappelier & M. Rajman