# Lecture Series on Predictive Language Models

El Mehdi ISSOUANI

Post-doc: Laboratory of Applied Mathematics of Compiègne (LMAC)
University of Technology of Compiègne (UTC)

Thesis obtained in June 2023 in MODAL'X at University Paris Nanterre

Thursday $6^{th}$ February, 2025

Lecture 8: Mathematical Formalization of POS Tagging: Maximum Entropy and Extension to Empirical Likelihood

# Outline

1. Tokenization
2. Part Of Speech Tagging (**POS tagging**)

**Tokenization examples**

*[He called Mr. Green at 2 p.m. in St. Louis, Mr. White did not answer. He then left him a voice mail message.]*

- Sentence tokenization

  *[He called Mr. Green at 2 p.m. in St. Louis, Mr. White did not answer.]*
  *[He then left him a voice mail message.]*

- Word tokenization

  *[He], [then], [left], [him], [a], [voice], [mail], [message], [.]*

1. NLP Tasks
   - Tokenization
   - Part-of-Speech tagging

2. Mathematical Formalization of POS Tagging

3. Maximum entropy principle (MaxEnt)

4. Penalized generalized empirical likelihood (PGEL)

## POS tagging examples

| Time | flies | like | an | arrow | . |
|---|---|---|---|---|---|
| ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| NN | VB | IN | DT | NN | . |

| Fruit | flies | like | a | banana | . |
|---|---|---|---|---|---|
| ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| JJ | NNS | VB | DT | NN | . |

VB: **Verb** - DT: **Determiner** - NN: **Noun singular** - IN: **Preposition** - NNS: **Noun plural** - JJ: **Adjective**

# Notations and Vocabulary

- We denote a sentence $s$ with $N_s$ words

$$s = w_1, \ldots, w_i, \ldots, w_{N_s}$$

  where $w_i$ represents the $i$th word and $N_s \in \mathbb{N}^*$ is random.
- $t_i$ is the tag of the $i$th word with $t_i \in \mathrm{T} = \{NN, NNS, VB, \ldots\}$ representing the tagset of finite size (example: 7, 36, 87)
- $n$: Number of words in the whole dataset $n = \sum_{s \in \mathcal{S}} N_s$

$$\begin{array}{c|cccccc}
 & w_1 & w_2 & \ldots & w_{N-1} & w_N & . \\
 & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\hline
sent \longrightarrow & \text{Time} & \text{flies} & \text{like} & \text{an} & \text{arrow} & . \\
 & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\
tags \longrightarrow & \text{NN} & \text{VB} & \text{IN} & \text{DT} & \text{NN} & . \\
\hline
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
 & t_1 & t_2 & \ldots & t_{N-1} & t_N & .
\end{array}$$

# Mathematical Models for POS Tagging

- Goal : we aim to obtain the best tag sequence

$$(t_1^*, ..., t_N^*) = \underset{(t_1,...,t_N)\in \mathrm{T}^{\mathrm{N}}}{\arg\max} \left[p\left(t_1, ..., t_N | w_1, ..., w_N\right)\right]$$

- Denote $x_1, ..., x_N$ the contexts of each word.

$$\text{Example:} \quad x_i = [w_{i-2}, w_{i-1}, \mathrm{w_i}, w_{i+1}, w_{i+2}]$$

- We now can write

$$p(t_1, ..., t_N | w_1, ..., w_N) = \prod_{i=1}^{N} p\left(t_i \,|\, x_i\right)$$

- Goal : we aim to obtain the best tag sequence

$$(t_1^*, ..., t_N^*) = \underset{(t_1,..., t_N)\in \mathrm{T}^{\mathrm{N}}}{\arg\max} \left[p\left(t_1, ..., t_N | w_1, ..., w_N\right)\right]$$

- Denote $x_1, ..., x_N$ the contexts of each word.

$$\text{Example:} \quad x_i = [w_{i-2}, w_{i-1}, \mathrm{w_i}, w_{i+1}, w_{i+2}]$$

- We now can write

$$p(t_1, ..., t_N | w_1, ..., w_N) = \prod_{i=1}^{N} p\left(t_i \,|\, x_i\right)$$

# Feature-based models

- Definition: A feature encodes the binary link between the tag and the context (or word environment).

  - Example: *"Time flies like an arrow."*

| $x_2 =$ | Time | flies | like |
|---|---|---|---|

$$\downarrow$$

| $t_2 =$ | | VB | |
|---|---|---|---|

| features: | ... | ... | (flies, NNS) | (flies, VB) | ... | (Time, VB) | ... |
|---|---|---|---|---|---|---|---|
| $f(x_2, \mathrm{NNS})$ | 0 | ... | 1 | 0 | ... | 0 | 0 |
| $f(x_2, \mathrm{VB})$ | 0 | ... | 0 | 1 | ... | 1 | 0 |
| $f(x_2, \mathrm{NN})$ | 0 | ... | 0 | 0 | ... | 0 | 0 |
| $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

# Maximum entropy principle (MaxEnt)

A method to infer a measure $p(z)$ defined on a given set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ under some constraint $\mathcal{P}$ (Csiszar (1996)[5], Hanson (2012)[6]).

$$p^* = \arg\max_{p \in \mathcal{P}} \{\mathcal{H}(p)\} = \arg\max_{p \in \mathcal{P}} \left\{ - \int p(z) \log p(z) \mathfrak{l}(dz) \right\},$$

$$\text{where} \quad \mathcal{P} = \left\{ p : \int f_k(z) p(z) \mathfrak{l}(dz) = \mu_k, k = 1, \ldots, q \right\}.$$

# Link to divergences

If we have access to a given default distribution $p_0 \in \mathcal{P}$

$$\begin{aligned} p^* = \arg\max_{p \in \mathcal{P}} \{\mathcal{H}(p)\} = & \arg\max_{p \in \mathcal{P}} \{\mathcal{H}(p_0) - D(p, p_0)\} \\ = & \arg\min_{p \in \mathcal{P}} \{D(p, p_0)\} \end{aligned}$$

$$\text{where} \quad D(p, p_0) = \int \left[ p(z) \log \frac{p(z)}{p_0(z)} - p(z) + p_0(z) \right] \mathfrak{l}(dz).$$

Consider $\lambda_i$'s the Kuhn & Tucker coefficients in the optimization program with each $\lambda_i$ corresponding to a constraint $\mu_i$,

$$p^*(z) = \frac{\exp\left(\sum_{k=1,..,q} \lambda_k f_k(z)\right)}{\int \exp\left(\sum_{k=1,..,q} \lambda_k f_k(u)\right) \mathfrak{l}(du)}.$$

.

# MaxEnt solution

Consider $\lambda_i$'s the Kuhn & Tucker coefficients in the optimization program with each $\lambda_i$ corresponding to a constraint $\mu_i$,

$$p^*(z) = \frac{\exp\left(\sum_{k=1,..,q} \lambda_k f_k(z)\right)}{\int \exp\left(\sum_{k=1,..,q} \lambda_k f_k(u)\right) \mathfrak{l}(du)}.$$

$\longrightarrow$ Notice that $p(z)$ doesn't depend on constraints $\mu_k$.

By putting $z = (x, t)$ we can rewrite

**Log-linear models, the join probability is given by**

$$p^*(z) = p^*(x,t) = p^*(t|x)p^*(x) = \frac{\exp\left(\sum_{k=1,..,q} \lambda_k f_k\left(x,t\right)\right)}{\int \exp\left(\sum_{k=1,..,q} \lambda_k f_k\left(u,v\right)\right) \mathfrak{l}(du,dv)},$$

**Conditional probabilities**

$$p^*(t|x) = \frac{\exp\left(\sum_{k=1,..,q} \lambda_k f_k\left(x,t\right)\right)}{\int \exp\left(\sum_{k=1,..,q} \lambda_k f_k\left(x,t'\right)\right) \mathfrak{l}'(dt')}.$$

# Generalized empirical likelihood - Keziou (2003, 2006) [7, 3]

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall \ y \in \mathbb{R}.$$

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall\ y \in \mathbb{R}.$$

- Consider the following assumptions for the function $\varphi$.

# Generalized empirical likelihood - Keziou (2003, 2006) [7, 3]

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall \ y \in \mathbb{R}.$$

- Consider the following assumptions for the function $\varphi$.

**H1** $\varphi$ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

# Generalized empirical likelihood - Keziou (2003, 2006) [7, 3]

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall \ y \in \mathbb{R}.$$

- Consider the following assumptions for the function $\varphi$.

**H1** $\varphi$ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

**H2** $\varphi$ is twice differentiable on a neighborhood of 0 ;

# Generalized empirical likelihood - Keziou (2003, 2006) [7, 3]

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall\ y \in \mathbb{R}.$$

- Consider the following assumptions for the function $\varphi$.

**H1** $\varphi$ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

**H2** $\varphi$ is twice differentiable on a neighborhood of 0 ;

**H3** (renormalization) $\varphi(0) = 0$ and $\varphi^{(1)}(0) = 0$, $\varphi^{(2)}(0) > 0$, which implies that $\varphi$ has an unique minimum at zero ;

- Consider $(\mathcal{Z}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. For every convex function $\varphi$, its Fenchel-Legendre transform is given by

$$\varphi^*(z) = \sup_{z \in \mathbb{R}} \{zy - \varphi(z)\}, \ \forall \ y \in \mathbb{R}.$$

- Consider the following assumptions for the function $\varphi$.

**H1** $\varphi$ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

**H2** $\varphi$ is twice differentiable on a neighborhood of 0 ;

**H3** (renormalization) $\varphi(0) = 0$ and $\varphi^{(1)}(0) = 0$, $\varphi^{(2)}(0) > 0$, which implies that $\varphi$ has an unique minimum at zero ;

**H4** $\varphi$ is differentiable on $d(\varphi)$, that is to say differentiable on $int\{d(\varphi)\}$, with right and left limits on the respective endpoints of the support of $d(\varphi)$, where $int\{.\}$ is the topological interior.

$$I_{\varphi^*}(\mathbb{Q},\mathbb{P}) = \begin{cases} \int_{\mathcal{Z}} \varphi^*\left(\frac{d\mathbb{Q}}{d\mathbb{P}} - 1\right) d\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else.} \end{cases}$$

**Theorem** (Borwein & Lewis (1992)[2], Keziou (2003)[7], Bertail (2007)[1])

Let $\varphi$ be a function satisfying assumptions **H1-H3**. If the following qualification constraint holds,

$$\mathrm{Qual}(\mathbb{P}) : \begin{cases} \exists \mathbb{T} \in \mathcal{M}, \mathbb{T}f = \mu \text{ and} \\ \inf d(\varphi^*) \ < \ \inf_{\mathcal{Z}} \frac{d\mathbb{T}}{d\mathbb{P}} \ \leq \ \sup_{\mathcal{Z}} \frac{d\mathbb{T}}{d\mathbb{P}} \ < \ \sup d(\varphi^*) \ \ \mathbb{P}-a.s., \end{cases}$$

then, we have the dual equality:

$$\inf_{\mathbb{Q}\in\mathcal{M}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}) |\ (\mathbb{Q}-\mathbb{P})f = \mu\} = \sup_{\lambda\in\mathbb{R}^q} \left\{\lambda'\mu - \int_{\mathcal{Z}} \varphi(\lambda' f) d\mathbb{P}\right\}$$

# $\varphi^*$-discrepancy and Dual Theorem

$$I_{\varphi^*}(\mathbb{Q},\mathbb{P}) = \begin{cases} \int_{\mathcal{Z}} \varphi^*\left(\frac{d\mathbb{Q}}{d\mathbb{P}} - 1\right) d\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else.} \end{cases}$$

**Theorem (Borwein & Lewis (1992)[2], Keziou (2003)[7], Bertail (2007)[1])**

Let $\varphi$ be a function satisfying assumptions **H1-H3**. If the following qualification constraint holds,

$$\mathrm{Qual}(\mathbb{P}) : \begin{cases} \exists \mathbb{T} \in \mathcal{M}, \mathbb{T}f = \mu \text{ and} \\ \inf d(\varphi^*) \ < \ \inf_{\mathcal{Z}} \frac{d\mathbb{T}}{d\mathbb{P}} \ \leq \ \sup_{\mathcal{Z}} \frac{d\mathbb{T}}{d\mathbb{P}} \ < \ \sup d(\varphi^*) \ \ \mathbb{P} - a.s., \end{cases}$$

then, we have the dual equality:

$$\beta_n(\mu) = \inf_{\mathbb{Q}\in\mathcal{M}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}) |\ (\mathbb{Q}-\mathbb{P})f = \mu\} = \sup_{\lambda\in\mathbb{R}^q} \left\{\lambda'\mu - \int_{\mathcal{Z}} \varphi(\lambda' f) d\mathbb{P}\right\}$$

# GEL solution

If $\varphi$ satisfies **H4**, then the infimum on the left hand side at $\mathbb{Q}^*$ is given by

$$p^* = \mathbb{Q}^* = (1 + \varphi^{(1)}(\lambda^{*\prime} f))\mathbb{P}.$$

$$\text{Define the empirical measure:} \quad \mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

- Example of **Kullback divergence** for measures:

It's the case where $\varphi_0(x) = -x - \log(1-x)$ and $\varphi_0^*(x) = x - \log(1+x)$.

## Kullback divergence

$$I_{\varphi_0^*}(\mathbb{Q}, \mathbb{P}) = K(\mathbb{Q}, \mathbb{P}) = -\int_{\mathcal{Z}} \log\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right) d\mathbb{P} + \int_{\mathcal{Z}} (d\mathbb{Q} - d\mathbb{P}),$$

$$p_i^* = \frac{1}{n(1 - \lambda^{*\prime}(Z_i - \mu))} \text{ and } \beta_n(\mu) = -1 - \sum \frac{1}{n}\log(np_i^*) + \sum p_i^*.$$

$$\text{where,} \quad p_i^* = \frac{1}{n(1 - \lambda^{*\prime}(Z_i - \mu))}.$$

# Examples of relative entropy and $\chi^2$ divergences

- **Relative entropy:** The particular case of $\varphi_1(x) = e^x - 1 - x$ whose convex conjugate is given by $\varphi_1^*(x) = (x+1)\log(1+x) - x$. Then,

**Relative entropy**

$$I_{\varphi_1^*}(\mathbb{Q},\mathbb{P}) = \begin{cases} \int_{\mathcal{Z}} \frac{d\mathbb{Q}}{d\mathbb{P}} \log(\frac{d\mathbb{Q}}{d\mathbb{P}}) d\mathbb{P} - \int_{\mathcal{Z}} (d\mathbb{Q} - d\mathbb{P}) & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else.} \end{cases}$$

And that the optimal weights are given by

$$p_i^* = \frac{1}{n} \exp\left(\lambda^{*\prime} f(Z_i,\mu)\right), \quad \text{where } \lambda^* \underset{n\to\infty}{\sim} -S_n^{-2} \overline{f}_n$$

- **Program solution with $\chi^2$ divergence:**

$$\beta_n(\mu) = n \bar{f}_n' S_n^{-2} \bar{f}_n \quad \text{where } \bar{f}_n = \frac{1}{n} \sum_{i=1}^{n} f(Z_i,\mu).$$

# Penalizing the dual program

- Remember that for the POS Tagging: $n \ll q$
- **Penalizing the dual** (Chang et al. (2018) [4], Shi (2016) [8])

$$P_n(\mu,\lambda) = \mathbb{P}_n\left( -\lambda'\left(f\left(x,t\right)-\mu\right) - \varphi(\lambda'\left(f\left(x,t\right)-\mu\right))\right) - \frac{1}{2}\|\lambda\|_R^2,$$

In the case of Relative entropy divergence:

$$\Longrightarrow P_n(\mu,\lambda) = 1 + \frac{1}{n}\sum_{i=1}^{n}\left(-\exp(\lambda'\left(f\left(x_i,t_i\right)-\mu\right)) - \frac{1}{2}\|\lambda\|_R^2\right).$$

When $R = \rho_n \mathrm{I}_q$ , then we have asymptotically for $\lambda$ close to 0,

$$P_n(\mu,\lambda) \approx \frac{1}{n}\sum_{i=1}^{n}\left(-\lambda'\left(f\left(x_i,t_i\right)-\mu\right)) - \frac{1}{2}\lambda'(S_n^2 + \rho_n \mathrm{I}_q)\lambda\right)$$

# The explicit expression of $\lambda^*$ for POS Tagging

The maximum is attained at

$$\lambda_n^* \underset{\infty}{\sim} -\left(S_n^2 + \rho_n \mathrm{I}_q\right)^{-1} \mathbb{P}_n(f - \mu) \quad (\text{recall that } \mathbb{P}_n = \frac{1}{n} \sum_i \delta_{Z_i})$$

$$\underset{\infty}{\sim} -\left(S_n^2 + \rho_n \mathrm{I}_q\right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left(f\left(x_i, t_i\right) - \mu\right)$$

In the penalized case we can see that the optimal weights depend on $\mu$

## Without using the maximum of likelihood (Issouani (2023))

$$\hat{p}\left(t_i | x\right) = \frac{e^{-(\bar{f}_n - \mu)'\left(S_n^2 + \rho_n \mathrm{I}_q\right)^{-2}(f(x, t_i) - \mu)}}{\sum_{t_k \in \mathcal{T}} e^{-(\bar{f}_n - \mu)'\left(S_n^2 + \rho_n \mathrm{I}_q\right)^{-2}(f(x, t_k) - \mu)}},$$

where we recall that $S_n^2 = \frac{1}{n} \sum_i^n \left(f(x_i, t_i) - \mu\right)\left(f(x_i, t_i) - \mu\right)'$.

# Application for POS Tagging

- **PennTreebank corpus:** 3914 sentences having a total of 100676 tokens (12408 tokens without repetitions) and 36 tags.
  - We estimate $\mu$ using the empirical mean estimated using the entire initial dataset
  - Then, we split the dataset into a training set (75% of the initial dataset) and a test set

- **Results:**
  - Estimation accuracy of 98% (over the training sample)
  - Prediction accuracy of 95% (on average over the test samples).

# Perspectives

- The impact of divergence choice on probabilities.
- Choose another norm than $L_2$ for the dual penalizing.

Patrice Bertail, Hugo Harari-Kermadec, and Denis Ravaille.
φ-divergence empirique et vraisemblance empirique generalisee.
*Annales d'Economie et de Statistique*, pages 131–157, 2007.

Jonathan M Borwein and Adrian S Lewis.
Partially finite convex programming, part ii: Explicit lattice models.
*Mathematical Programming*, 57:49–83, 1992.

Michel Broniatowski and Amor Keziou.
Minimization of ϕ-divergences on sets of signed measures.
*Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006.

Jinyuan Chang, Cheng Yong Tang, and Tong Tong Wu.
A new scope of penalized empirical likelihood with high-dimensional estimating equations.
*The Annals of Statistics*, 46(6B):3185–3216, 2018.

Imre Csiszár.
Maxent, mathematics, and information theory.
In *Maximum entropy and Bayesian methods*, pages 35–50. Springer, 1996.

Kenneth M Hanson and Richard N Silver.
*Maximum Entropy and Bayesian Methods: Santa Fe, New Mexico, USA, 1995 Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, volume 79.
Springer Science & Business Media, 2012.

Amor Keziou.
Dual representation of $\varphi$-divergences and applications.
*Comptes Rendus Mathematique*, 336(10):857–862, 2003.

Zhentao Shi.
Econometric estimation with high-dimensional moment equalities.
*Journal of Econometrics*, 195(1):104–119, 2016.

Thank you for your attention!