Penalized Generalized Empirical Likelihood in High Dimension : Application to POS-Tagging

El Mehdi Issouani, Patrice Bertail, Thierry Dumont, Melanie Zetlaoui

May 12, 2025

Abstract In this paper, we combine methods used in penalized generalized empirical likelihood (GEL) frameworks with feature extraction techniques that enable projecting textual data into numerical spaces. We discuss some recent techniques used in GEL when the number of features gets very large in comparison to the sample size. We relate this approach to the Maximum Entropy (MaxEnt) principle used in several tasks of Natural Language Processing (NLP), in particular in POS-Tagging. Since the features belong to a large dimensional space, we propose a penalization method based on the dual representation of the original problem: this yields to an explicit approximation of conditional probabilities of tags given the context. This method considerably reduces computational costs. As a byproduct, for different GEL methods, we obtain the corresponding POS-Tagging classifiers generalizing the MaxEnt method.

1 Introduction

1.1 Some references on Generalized Empirical Likelihood

Empirical likelihood (EL) was first introduced by Thomas and Grunkemeier (1975) [45] to provide improved confidence intervals using the Kaplan-Meier estimator in survival analysis. Extensions of this approach to survey sampling were explored by Hartley and Rao (1968) [18]. Owen (1988) [35] developed a general framework for empirical likelihood in nonparametric inference. Between 1988 and 1990, Owen further generalized Wilk's theorem (1938) [46], demonstrating that the quantity $-2\log(\mathcal{R})$ asymptotically follows a χ^2 distribution in a nonparametric context, with R representing the likelihood ratio (see Owen (1988) [35], Owen et al. (1990) [34]).

The empirical log-likelihood ratio can be interpreted as the minimization of the Kullback divergence between the empirical distribution $\mathbb{P}_n$ of the observed data and a probability measure $\mathbb{Q}$ dominated by $\mathbb{P}_n$, subject to linear or nonlinear constraints defined by the statistical model. Other pseudo-metrics besides the Kullback divergence have been considered by Owen et al. (1990) [34] and several other authors. For example, choosing relative entropy has given rise to "Entropy econometrics" in econometrics (Golan et al. (1996) [16]). Related work appears in probabilistic literature concerning divergences and entropy methods (see Leonard (2001) [25, 26, 27], Gamboa and Gassiat (1996) [15]). Generalizations of empirical likelihood based on the Cressie-Read discrepancy have also been developed. These generalizations, called "generalized empirical likelihood," have been studied in econometrics by Newey and Smith (2004) [32], although they typically lose certain desirable likelihood properties, such as Bartlett-correctability. Bertail et al. (2014) [4] demonstrated that Owen's empirical likelihood approach for estimating means can be extended under mild conditions to any regular convex statistical divergence or φ^*-discrepancy (where φ^* is a regular convex function), and for general Hadamard-differentiable functionals (see Bertail et al. (2007) [6] and Bertail et al. (2015) [5]). They call this method "empirical energy minimizers" by analogy to the theoretical probabilistic literature on the

subject (see Leonard (2001) [25, 26, 27] and the references therein).

1.2 POS-Tagging

The main purpose of this paper is to propose some new classification methods which are time and computationally inexpensive in the framework of Natural Language Processing (NLP). We recall a few notions of POS-Tagging.
POS-Tagging assigns a part of speech (grammatical category, gender, or number) to each word in a sentence. Given an *input* (e.g., a sentence and a tagset), the goal is to predict the most suitable tag for each word using computational methods. A more detailed presentation with examples is provided in Section 3.
In this framework, the generalized empirical method is closely related to the maximum entropy (MaxEnt) method used in POS-Tagging, through the concept of "dual likelihood" introduced by Mykland (1995) [31]. Specifically, selecting a particular divergence naturally leads to a dual likelihood function. In the special case of entropy, this dual likelihood matches precisely the likelihood function used by the MaxEnt method described later. This connection suggests possible extensions: first, the creation of new types of likelihood functions, and second, the development of procedures that can better handle high-dimensional problems common in text analysis.

2 Generalized empirical likelihood and MaxEnt models

2.1 A brief overview of empirical likelihood

Let $Z_1, ..., Z_n$ be independent identically distributed variables following $\rightsquigarrow \mathbb{P} \in \mathcal{P}$ (where $\mathcal{P}$ is a convex set of probability), with Z_i taking values on a space $\mathcal{X}$ defined on a probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$. We are interested in constructing a confidence region for the functional parameter $\theta = \mathbf{T}(\mathbb{P})$ defined on ζ, taking values in $\mathbb{R}^d$. In the following, we define $\mathbb{P}_n$ the empirical probability measure as follows

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}.$$

Owen (1990) [34] has shown that $\mathbb{P}_n$ is the NPMLE of $\mathbb{P}$ (Non Parametric Maximum Likelihood Estimate). Thus the NPMLE of $\mathbf{T}(\mathbb{P})$ is then its emprical counterpart $\hat{\theta}_n = \mathbf{T}(\mathbb{P}_n)$, called a statistical functional. Many statisticians (since Von Mises, see Serfling (1980) [42]) have been interested in deriving the asymptotic properties of $\hat{\theta}_n$ using differentiability assumptions on $\mathbf{T}$ via Taylor expansion (the delta method).

The empirical likelihood ratio evaluated at θ is defined by

$$\mathcal{R}_n(\theta) = \sup_{\mathbb{Q}_n \in \mathcal{P}_n} \left\{ \prod_{i=1}^{n} \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}(Z_i), \mathbf{T}(\mathbb{Q}_n) = \theta \right\},$$

where

$$\mathcal{P}_n = \left\{ \mathbb{Q}_n = \sum_{i=1}^{n} p_{i,n} \delta_{Z_i}, p_{i,n} \geq 0, \sum_{i=1}^{n} p_{i,n} = 1 \right\}$$

.

The log-likelihood ratio is thus

$$\log(\mathcal{R}_n(\theta)) = \sup_{(p_{i,n})_{i \leq n}} \left\{ \sum_{i=1}^{n} \log\left(\frac{p_{i,n}}{\frac{1}{n}} \right), \mathbf{T}\left(\sum_{i=1}^{n} p_{i,n} \delta_{Z_i} \right) = \theta, \sum_{i=1}^{n} p_{i,n} = 1 \right\}.$$

A better way to see this problem from a probabilistic point of view is to consider the formula above as the minimization of the Kullback distance I_K between $\mathbb{Q}_n$ and $\mathbb{P}_n$, where

$$I_K(\mathbb{Q},\mathbb{P}) = \begin{cases} -\int \log(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}})\mathrm{d}\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ \infty & \text{else} \end{cases}$$

under the two constraints, on the parameter and the probabilities $p_{i,n}$.

For instance, $\mathbf{T}(\mathbb{P})$ may be the unique solution of some estimating equations $E_{\mathbb{P}} g(Z, \mathbf{T}(\mathbb{P})) = 0$ (see Qin and Lawless (1994) [38]) where for each fixed parameter $\mathbf{T}(\mathbb{P})$, g is a measurable function defined from $\mathcal{X}$ to $\mathbb{R}^d$, $d \geq 1$. These equations will also include marginal constraints (that is constraints independent of the parameter $\mathbf{T}(\mathbb{P})$ incorporating some knowledge of the data: see an application on large datasets of this kind of idea in Crepet et al. (2009) [13]). In this case, the constraint becomes $E_{\mathbb{Q}_n} g(Z,\theta) = 0 = \sum p_{i,n} g(Z_i,\theta)$ and the empirical likelihood boils down to the convex maximization program

$$\mathcal{R}_n(\theta) = \sup_{p_{i,n}, i=1,\ldots,n} \left\{ \begin{array}{c} \frac{\prod_{i=1}^n p_{i,n}}{1/n^n} \text{ under} \sum_{i=1}^n p_{i,n} g(Z_i,\theta) = 0 \\ \sum_{i=1}^n p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\}.$$

Some standard results in convex optimization theory give conditions for this problem to have a solution and also allow to obtain a dual representation of this problem. This is precisely the dual representation that generates the MaxEnt model used in NLP as explained below.

2.2 A general view of empirical likelihood

Consider a measured space $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. Working on a space of signed measures will be essential for applications to ensure the existence of solutions of the original optimization program. Let g be a measurable function defined from $\mathcal{X}$ to $\mathbb{R}^d$, $d \geq 1$. For any measure $m \in \mathcal{M}$, we write $mg = \int g\mathrm{d}m$. In the following, we consider φ, a convex function whose support $\mathrm{dom}(\varphi) = \{x \in \mathbb{R},\ \varphi(x) < \infty\}$, is assumed to be non-void (that is φ is proper). We denote respectively $\inf\{\mathrm{dom}(\varphi)\}$ and $\sup\{\mathrm{dom}(\varphi)\}$, the extremes of this support. For every convex function φ, its convex dual or Fenchel-Legendre transform is given by

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

Recall that φ^* is then a semi-continuous inferior (s.c.i.) convex function. We define by $\varphi^{(i)}$ the derivative of order i of φ when it exists. From now on, we will assume the following assumptions for the function φ.

A1 φ is strictly convex and $\mathrm{dom}(\varphi)$ contains a neighborhood of 0 ;

A2 φ is twice differentiable on a neighborhood of 0 ;

A3 (renormalization) Assume $\varphi(0) = 0$ and $\varphi^{(1)}(0) = 0$, $\varphi^{(2)}(0) = 1$, which implies that φ has an unique minimum at zero and $\varphi(x)$ behaves like $x^2/2$ at 0 ; For the divergences of interest (see Table 3 for common divergences with their domain), it is always possible to renormalize the function φ in such a way.

A4 φ is differentiable on $\mathrm{dom}(\varphi)$, that is to say, differentiable on $\mathrm{int}\{\mathrm{dom}(\varphi)\}$, with right and left limits on the respective endpoints of the support of $\mathrm{dom}(\varphi)$, where $\mathrm{int}\{.\}$ is the topological interior.

A5 φ is twice differentiable on $\mathrm{dom}(\varphi) \cap \mathbb{R}^+$ and, on this domain, the second order derivative of φ is bounded from below by a constant $\varphi_{\min} > 0$.

Let φ satisfies the assumptions **A1, A2, A3**. Then, the Fenchel dual transform φ^* of φ also satisfies these assumptions. The φ^*-discrepancy I_{φ^*} between $\mathbb{Q}$ and $\mathbb{P}$, where $\mathbb{Q}$ is a signed measure and $\mathbb{P}$ a positive measure, is defined as follows

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

For details on φ^*-discrepancies or divergences and some historical comments, see Liese and Vajda (1987) [28], Leonard (2001) [25, 26, 27].

Our primary interest in φ^*-discrepancies comes from the following duality representation, derived from results by Borwein and Lewis (1991) [7] on convex integral functionals (see also Rockafellar (1968) [41]). The following result, presented by Bertail et al.(2015) in [6], is a simplified version of the duality result established by Borwein and Lewis (1991) [7]. Additional dual representations and detailed topological analyses of the problem are provided by Keziou (2003) [22] and Broniatowski and Keziou (2006) [10].In the following we denote by $\inf_{\mathcal{X}} \frac{d\mathbf{T}}{d\mathbb{P}}$ and $\sup_{\mathcal{X}} \frac{d\mathbf{T}}{d\mathbb{P}}$ respectively the lower point and the upper point of the support of the corresponding Radon-Nikodym density.

Proposition 2.1. *(see Bertail et al. (2015) [5]) Let $\mathbb{P} \in \mathcal{M}$ be a probability measure with finite support, and let f be a measurable function on $(\mathcal{X}, \mathcal{A}, \mathcal{M})$. Consider a convex function φ satisfying assumptions **A1**–**A3**. Suppose the following qualification condition holds:*

$$Qual(\mathbb{P}) : \begin{cases} \exists \mathbf{T} \in \mathcal{M},\ \mathbf{T}g(.,\theta) = T_0 \text{ and} \\ \inf\{\mathrm{dom}(\varphi^*)\} < \inf_{\mathcal{X}} \frac{\mathrm{d}\mathbf{T}}{\mathrm{d}\mathbb{P}} \ \leq\ \sup_{\mathcal{X}} \frac{\mathrm{d}\mathbf{T}}{\mathrm{d}\mathbb{P}} < \sup\{\mathrm{dom}(\varphi^*)\}, \end{cases}$$

then the following dual equality holds:

$$\inf_{\mathbb{Q}\in\mathcal{M}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}) |\ (\mathbb{Q}-\mathbb{P})g(.,\theta) = T_0\} = \sup_{\lambda\in\mathbb{R}^d} \left\{\lambda^\top T_0 - \mathbb{P}\varphi(\lambda^\top g(.,\theta))\right\}. \tag{2}$$

*s Moreover, if φ satisfies condition **A4**, then the supremum on the right hand side of (2) is reached at some point λ^* and the infimum on the left hand side at $\mathbb{Q}^*$ is given by*

$$\mathbb{Q}^* = (1 + \varphi^{(1)}(\lambda^{*\top} g(.,\theta)))\mathbb{P}.$$

In addition, similar results hold when the number of constraints becomes large or even infinite; see Leonard (2001) [25, 26, 27] and Gamboa and Gassiat (1996) [15].

The same kind of results also holds when the number of constraints goes to infinity or even is infinite, see Leonard (2001) [25, 26, 27] and Broniatowki and Keziou (2012) [11] for some applications to a continuum of moment constraints.

Let $Z, Z_1, ...Z_n$ be i.i.d. r.v.'s defined on $\mathcal{X}$ with common probability measure $\mathbb{P} \in \mathcal{M}$. Assume in addition that $g(Z,\theta)$ is such that $\Sigma_d = \mathbb{E}\left(g(Z,\theta)g(Z,\theta)^\top\right)$ exists and is positive definite. This rules out the so-called over-identified case (in the econometric literature).

For a given φ, we define, by analogy to the empirical likelihood problem, the quantity

$$\beta_n(\theta) = n \inf_{\mathbb{Q}\in\mathcal{M}_n} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n)\} \quad \text{with } \mathcal{M}_n = \{\mathbb{Q} \in \mathcal{M} \mid \mathbb{Q} \ll \mathbb{P}_n,\ \mathbb{E}_{\mathbb{Q}} g(Z,\theta) = 0\}.$$

For $\mathbb{Q}$ in $\mathcal{M}_n$, the constraints can be rewritten as $(\mathbb{Q}-\mathbb{P}_n)g(.,\theta) = -\mathbb{P}_n g(.,\theta)$. Using the result of Equation (2) or the results of Broniatowki and Keziou (2006) [10], we get the dual

representation

$$\beta_n(\theta) := n \inf_{\mathbb{Q}\in\mathcal{M}_n} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n),\ (\mathbb{Q}-\mathbb{P}_n)g(.,\theta) = -\mathbb{P}_n g(.,\theta)\}$$
$$= n \sup_{\lambda\in\mathbb{R}^d} \mathbb{P}_n \left(-\lambda^\top g(.,\theta) - \varphi(\lambda^\top g(.,\theta))\right). \tag{3}$$

Notice that $-x-\varphi(x)$ is a strictly concave function and that the function $\lambda \to \lambda^\top g$ is also concave. The parameter λ can be simply interpreted as the Kuhn-Tucker coefficient associated with the original optimization problem. From this representation of $\beta_n(\theta)$, we can now derive the usual properties of the empirical likelihood and its generalization. In the following, we will also use the notations

$$\overline{g}_n = \mathbb{P}_n g\,(.,\theta) = \frac{1}{n}\sum_{i=1}^n g(Z_i,\theta)\ ;\ S_n^2 = \frac{1}{n}\sum_{i=1}^n g(Z_i,\theta)g(Z_i,\theta)^\top \text{ and } S_n^{-2} = (S_n^2)^{-1}.$$

Since the dimension d is fixed, S_n^2 is always invertible for n large enough when Σ_d is positive definite.

The following results give an explicit approximation of the optimal Kuhn and Tucker coefficient λ_n^* and recall the standard asymptotic distribution of $\beta_n(\theta)$, already obtained in Owen (2001) [36] and Harari (2006) [17]. Our main contribution here is an explicit control of the Kuhn-Tucker coefficient, which will be useful for constructing new, quick classifiers.

Theorem 2.1. *Under the assumptions **A1**–**A5**. Assuming that Σ_d is positive definite, we have*

$$\lambda_n^* = -S_n^{-2}\bar{g}_n + o_{\mathbb{P}}\left(n^{-1/2}\right)$$

Moreover, we have

$$2\beta_n(\theta) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2(d).$$

In Table 3, we review the form of the weights obtained at the optimal value of the Kuhn & Tucker coefficients for different φ-divergences.

2.3 Penalizing the dual likelihood in large dimension

The preceding results and the asymptotic validity of generalized empirical likelihood essentially hold when d the number of constraints (equal for us to the dimension of the parameter θ) is fixed and small compared to n. McKeague and von Keilegom (2009) [19] have studied the validity of empirical likelihood when d depends on n and such that $d << n^{1/3}$. They show that empirical likelihood still works: this can be explained by the fact that in that case, the empirical variance automatically computed by the internal optimization program is still a convergent estimator of the true variance. However, as noticed by several authors, the method fails when the number of constraints tends to be too big, in particular when it is of the same size as n, see Lahiri (2012) [23] and Bartolucci (2007) [1].

Several propositions have emerged to treat large dimension problems with generalized empirical likelihood. We may classify them into three classes (or combinations of the three methods).

(i) **Enlarge-the-margin methods** : by this, we mean that instead of the original empirical likelihood problem, allow for some flexibility or some pertubations of the original constraints. This can be done either by adding one or several points to the data which do not have exactly the correct mean (see Chen, Variyath and Abraham (2008), Emerson and Owen (2009)). Or this can be done by replacing the original constraints by some inequality constraints with respect to some norm $\|.\|_R$ defined by $\|x\|_R = x^\top Rx$, where R is possibly random allowing for some flexibility in the constraints. This leads to a relaxed empirical likelihood version

$$\mathcal{R}_n^{pen}(\theta) = \sup_{(p_{i,n})_{i\leq n}} \left\{ \begin{array}{c} n^n \prod_{i=1}^n p_{i,n} \text{ under } \left\|\sum_{i=1}^n p_{i,n} g(Z_i,\theta)\right\|_R \leq \delta_n \\ \sum_{i=1}^n p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\} \tag{4}$$

where δ_n is a margin to be calibrated (possibly depending on the data).

(ii) **Penalize the empirical likelihood** either on the primal form or the dual form. It is well known in the convex literature that program (4) may also be rewritten

$$log(\mathcal{R}_n^{pen}(\theta)) = \sup_{p_{i,n}, i=1,\ldots,n} \left\{ \begin{array}{c} \sum_{i=1}^n \log(p_{i,n}) - \ C_n(\delta_n)||\sum_{i=1}^n p_{i,n} g(Z_i,\theta)||_R \\ \sum_{i=1}^n p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\}$$

which may be interpreted as a penalized version of the original program. Such penalizations have been studied in Bartollucci (2007) [1] and Lahiri and Mukhopadhyay (2011) [23] when $g(Z_i,\theta) = Z_i - \theta,\ Z_i = (Z_{i,1}, ...Z_{i,d})^\top \in \mathbb{R}^d$. The proposition of Bartolluci (2007) corresponds to the choice $R = \hat{S}_n^{-2}$ and $\|x\|_R = x^\top \hat{S}_n^{-2} x$, $C_n(h) = n/2h^2$, and $\hat{S}_n^2$ is the sample covariance matrix

$$\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z}_n)(Z_i - \overline{Z}_n)^\top, \quad \text{with } \overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i.$$

Notice that this proposition may cause problems when d is bigger than n, since in that case the sample covariance matrix is not full rank and thus not invertible. The proposition of Lahiri and Mukhopadhyay (2011) [23] in a more general dependent framework (the $(Z_i)_{i=1,\ldots,n}$ may be weak mixing or with long range dependence) corresponds to

$$R = \text{diag}(\hat{\sigma}_j^{-2})_{j=1,\ldots,d},$$

with $C(h) = h$, where we use

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^n (Z_{i,j} - \bar{Z}_{j,n})^2$$

the marginal empirical variances and $Z_{i,j}$ and $\bar{Z}_{j,n}$ the jth component of Z_i and $\bar{Z}_n$ respectively. Notice that for a very large dimension, this proposal is expected to work better since $\hat{S}_n^2$ is singular if $d > n$.

Another proposition is to penalize the empirical likelihood in its dual form (see Mykland [31]) for an introduction to dual likelihood. Penalized version in the dual form has been recently studied by Otsu (2007) [33], and Chang et al. (2018) [12]. The most important results have been obtained by Shi (2016) [43] who proved that, for empirical likelihood with a correct penalization, the number of constraints may be as large as $o(\exp(n^{1/3}))$.

For generalized empirical likelihood, this corresponds to studying a penalized version of the dual program of the form

$$\gamma_n(\theta,\lambda) = \mathbb{P}_n\left(-\lambda' g(.,\theta) - \varphi(\lambda' g(.,\theta))\right) - \frac{1}{2}||\lambda||_R^2, \tag{5}$$

and the corresponding optimisation problem

$$\gamma_n^*(\theta) = \sup_{\lambda\in\mathbb{R}^d} (\gamma_n(\theta,\lambda)), \tag{6}$$

which is clearly linked to the proposition of Bartolucci (2007) [1] and Lahiri and Mukhopadhyay (2012) [23] by duality consideration. We will not investigate here the relations between the different dual formulations : however this would be clearly of interest in particular when one

uses the ℓ_1 or the ℓ_∞ norms or a combination of these norms with ℓ_2 (elastic net) instead of a simpler ℓ_2 norm.

When $R = \rho_n.I$, for some positive constant ρ_n, note that in the χ^2 case the dual problem (5) and for the choice of à ℓ_2 penalization, the optimization program becomes

$$\gamma_n^*(\theta) = \sup_{\lambda \in \mathbb{R}^d} \mathbb{P}_n \left\{ -\lambda^\top g(.,\theta) - \frac{(\lambda^\top g(.,\theta))^2}{2} - \frac{\rho_n}{2} \lambda^\top \lambda \right\}$$

and the solution of this program is simply the regularized Hotteling statistics

$$\frac{1}{2} \mathbb{P}_n g(.,\theta)^\top \left[\mathbb{P}_n g(.,\theta) g(.,\theta)^\top + \rho_n I \right]^{-1} \mathbb{P}_n g(.,\theta)$$

which is a regularized form of the T^2 Hotelling statistics (with no centering).

Consider $\mu = [\mu_j]_{\{j=1,\ldots,d\}}$ the eigenvalues of the empirical covariance matrix $S_n^2 = \mathbb{P}_n g(.,\theta) g(.,\theta)^\top$. Define the vector

$$\frac{\mu}{\mu + \rho_n} = \left\{ \frac{\mu_j}{\mu_j + \rho_n} \right\}_{j=1,\ldots,d}$$

and the so-called effective dimensions

$$\left\| \frac{\mu}{\mu + \rho_n} \right\|_1 = \sum_{j=1}^d \left| \frac{\mu_j}{\mu_j + \rho_n} \right| \text{ and } \quad \left\| \frac{\mu}{\mu + \rho_n} \right\|_2 = \sqrt{\sum_{j=1}^d \left(\frac{\mu_j}{\mu_j + \rho_n} \right)^2}.$$

The following result is a penalized version of Theorem 2.1 when the dimension $d \geq n$. It shows that a standardized version of the penalized generalized empirical likelihood is asymptotically normal. See also Peng and Schick (2018) [37] for similar results for empirical likelihood in a slightly different large dimension framework. We also give an approximation of the optimal value of λ, which will prove useful to construct our predictors for POS-Tagging.

Theorem 2.2. *Under the assumptions **A1**–**A5**. Assume that Σ_d is positive definite and that its largest eigenvalue is bounded, then we have*

$$\lambda_n^* = -\left(S_n^2 + \rho_n I \right)^{-1} \bar{g}_n + o_{\mathbb{P}}\left(n^{-1/2} \rho_n \right)$$

Moreover, if we have $\left\| \frac{\mu}{\mu+\rho_n} \right\|_2 \xrightarrow[n\to\infty]{} \infty$ as $d > n$ goes to ∞, then we have

$$\frac{2\gamma_n^*(\theta) - \left\| \frac{\mu}{\mu+\rho_n} \right\|_1}{\sqrt{2 \left\| \frac{\mu}{\mu+\rho_n} \right\|_2^2}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

When the dimension $d << n$, Bertail et al. (2008) [3] have shown that one can choose $\rho_n = 0$ and can get some exact exponential bounds for this quantity. Issouani et al. (2024) [21] investigate conditions to obtain exponential bounds in this large dimension framework by choosing an adequate value for the penalty ρ_n, assuming few moment conditions on g. The asymptotic behavior of the χ^2 or GMM (Generalized Moment Method) case when there is an infinite number (or a continuum) of constraints has been treated by several authors in the econometric literature, see for instance Carasco and Florens (2000), using some Tikhonov regularization of the operator S_n^2. The condition on the maximum eigenvalue of Σ_d simply means that the components of the function g can not be too strongly correlated and the condition on

$\left\| \frac{\mu}{\mu+\rho_n} \right\|_2$ is automatically satisfied when the penalization is small.

In the following section, we will apply our results to the framework of NLP. We will focus essentially on the simpler case of POS-Tagging. Still, the same kind of ideas may be applied in more complex classification tasks such as text (complex/simple) classification and simplification (see [20]).

3 Application to POS-Tagging

POS-Tagging assigns each word in a sentence its part of speech, such as grammatical category, gender, or number. Given an *input* (e.g., a sentence and a tagset), the goal is to predict the best tag for each word using a computational tool. The main challenge is to resolve ambiguities by selecting the correct tag based on context. This task is particularly suited for evaluating prediction accuracy, as it allows straightforward verification of correct tags. Tagging is more efficient with methods that take into account the local context (see Subsection 2.1 of Issouani's phD thesis [20]).

3.1 A brief overview of POS-Tagging

Advantages of POS-Tagging POS-Tagging is a simple task (often linear in processing time) with many applications. As Feldman (2010) [14] noted, it is crucial for linguistic research, enabling the analysis of constructions in large corpora [30]. POS information can also serve as a basis for syntactic parsing by providing grammatical context, serving as a preprocessor to speed up parsing. Additionally, it supports word-sense disambiguation, text production, and morphological generation by modeling POS sequences. This knowledge is essential for tasks like extracting key words (e.g., verbs) for text summarization or simplification.

Tagsets and Examples There exists several tagsets such as Penn Treebank, Brown and British national corpora. The POS-tags used below are taken from the tagset Penn Treebank Corpus, proposed at the University of Pennsylvania that includes 36 tags, see Figure 3.1. Recall that there is also a collection called *universal tagset*, which just says if the word is a **pron, noun, verb, det, adj, adv or punct**. So it contains 7 tags in total.

CC	Coord. conjunction	**NNS**	Noun, plural	**UH**	Interjection
CD	Cardinal digit	**NNP**	Proper noun, sing.	**VB**	Verb, base form
DT	Determiner	**NNPS**	Proper noun, plur.	**VBD**	Verb, past tense
EX	Existential there	**PDT**	Predeterminer	**VBG**	Verb, gerund
FW	Foreign word	**POS**	Possessive ending	**VBN**	Verb, past participle
IN	Prep./Subord. conj.	**PRP**	Personal pronoun	**VBP**	Verb, present sing.
JJ	Adjective	**PRP$**	Poss. pronoun	**VBZ**	Verb, 3rd sing. pres.
JJR	Adj., comparative	**RB**	Adverb	**WDT**	wh-determiner
JJS	Adj., superlative	**RBR**	Adj., comparative	**WP**	wh-pronoun
LS	List marker	**RBS**	Adj., superlative	**WP$**	Poss. wh-pronoun
MD	Modal	**RP**	Particle	**WRB**	wh-adverb
NN	Noun, singular	**TO**	to		

Figure 1: PennTreebank tagset. See A.Taylor & al. (2003) *[44]*

POS-Tagging examples Here are two examples of pos-tagged sentences:

"I saw a girl with a telescope." ; "The grand jury commented on a number of other topics."

I	saw	a	girl	with	a	telescope	.
↓	↓	↓	↓	↓	↓	↓	↓
PRP	VBD	DT	NN	IN	DT	NN	.

The	grand	jury	commented	on	a	number	of	other	topics	.
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
DT	JJ	NN	VBD	IN	DT	NN	IN	JJ	NNS	.

Existing approaches There are now numerous systems for the automatic assignment of parts of speech tagging, employing many different machine learning methods. Among recent top-performing methods are Hidden Markov Models (Brants (2000) [8]), maximum entropy approaches (Ratnaparkhi et al. (1996) [40]), and transformation-based learning (Brill (1994) [9]). An overview of these and other approaches can be found in Manning and Schoetze (1999) [29] (Chapter 10). Notice that Ratnaparkhi's thesis (1998) [39] contains different NLP tasks like sentence boundary detection (tokenization), POS-Tagging, and parsing where he used ideas similar to the maximum entropy method.

Mathematical modeling POS-Tagging task can be considered as a ***classification*** problem, where the goal is to estimate a function $h : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ which maps an object $x \in \boldsymbol{\mathcal{X}}$, where $\mathcal{X}$ is some abstract measurable space (texts, sentence, sequence of words) equipped with a σ-algebra, to its correct class $y \in \boldsymbol{\mathcal{Y}}$. That is to say, we consider a classifier of the form

$$\begin{cases} h : & \boldsymbol{\mathcal{X}} \longrightarrow \boldsymbol{\mathcal{Y}} \\ & x \longmapsto y. \end{cases}$$

3.2 Penalized MaxEnt method applied to POS-Tagging

In the following, we apply the ideas of generalized empirical likelihood developed in section 2.2 to the POS-Tagging problem described in 3.1.

Let us remember that we have at our disposal a corpus $C = \{(w_i, t_i)\}_{i=1\ldots n}$. We transform C to a new dataset $D = \{(x_i, t_i)\}_{i=1\ldots n}$ where x_i's are the contexts of w_i's. Each x_i represents a new vector containing the current word w_i and the surrounding information (including words, punctuation, affixes of the current word, etc.). We intend to estimate $p(t_i|x_i)$ using the generalized empirical likelihood framework (Bertail (2006) [2]) equipped with relative entropy divergence. For two p.d. $\mathbf{p} = (p_1, ..., p_n)$ and $\mathbf{q} = (q_1, ..., q_n)$, recall that D_E is given by :

$$D_E(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

Features are functions f that encode the local textual information: they take as input a pair, the context and its tag, and return a high-dimensional binary vector $f(x_i, t_i)$. We will show in Subsection 4.1, how this feature functions are constructed practically according to a given dictionary. Now, assume that we observe n i.i.d. random vectors $Z_i = f(x_i, t_i)$ having some distribution $\mathbb{P}$. In this framework, the multidimensional function $g(Z, \theta)$ used before is given by the vector $f(t, x) - \theta = (f_j(t, x) - \theta_j)_{j=1\ldots,d}$, where $\theta = (\theta_j)_{j=1,\ldots,d}$ is the expectation of $f(x, t)$. As before we assume that the variance $V(Z_i)$ exists.

3.2.1 Relative entropy and MaxEnt problem ($d < n$)

It is easy to check that the dual of the initial minimization problem is given by

$$\sup_{\lambda\in\mathbb{R}^d}\left\{1-\frac{1}{n}\sum_{i=1}^{n}e^{\lambda^\top(f(x_i,t_i)-\theta)}\right\}. \tag{7}$$

From these, we see that the optimal weights $p_i^* = p^*(x_i,t_i)$, $i=1,...,n$ satisfy

$$p^*(x_i)\,p^*(t_i|x_i) = \frac{1}{n}e^{\lambda^{*\top}(f(x_i,t_i)-\theta)}.$$

Since $\sum_{k=1}^{T} p^*(t_k \mid x_i) = 1$, it follows that

$$p^*(t_i|x_i) = \frac{\exp(\lambda^{*\top}\theta)\cdot\exp(-\lambda^{*\top}f(x_i,t_i)))}{\exp(\lambda^{*\top}\theta)\cdot\sum_{k=1}^{T}\exp(-\lambda^{*\top}f(x_i,t_k)))}.$$

This justifies the use of formula (7) in the MaxEnt program.

Finally, we obtain

$$p^*(t_i \mid x_i) \quad = \quad \frac{e^{-\lambda^{*\top}f(x_i,t_i)}}{\sum_{k=1}^{T}e^{-\lambda^{*\top}f(x_i,t_k)}} = \frac{e^{-\sum_{j=1}^{K}\lambda_j^*\cdot f_j(x_i,t_i)}}{\sum_{k=1}^{T}e^{-\sum_{j=1}^{K}\lambda_j^*\cdot f_j(x_i,t_k)}} \tag{8}$$

This shows that minimizing the Relative entropy divergence between the desired distribution and the multinomial distribution gives the same solution as the one obtained when maximizing the likelihood of a log-linear model based on the features.

Moreover, we get the predictive probability of t_i given x using the estimate

$$p^*(t_i|x) = \frac{e^{-\lambda^{*\top}f(x,t_i)}}{\sum_{t_i\in\mathcal{T}}e^{-\lambda^{*\top}f(x,t_i)}}$$

We know from the duality results exposed before that the optimal value of λ is asymptotically given by $\lambda^* = -S_n^{-2}(\bar{f}_n-\theta)$ up to $o(n^{-1/2})$. Unfortunately, this quantity depends on θ. It may be estimated in two different ways according to the context we are interested in.

- Method 1 : Either estimate θ by estimating the log-linear model considered in the MaxEnt method or by using the method proposed in Quin and Lawless(1994) [38] that is, find the value of $\hat{\theta}$ which realizes

$$\inf_{\theta\in\mathbb{R}^d}\sup_{\lambda\in\mathbb{R}^d}\left\{1-\frac{1}{n}\sum_{i=1}^{n}e^{\lambda^\top(f(x_i,t_i)-\theta)}\right\} \tag{9}$$

  This will yield asymptotically $\hat{\lambda}^* = -\hat{S}_n^{-2}(\bar{f}_n-\hat{\theta})$ with

$$\hat{S}_n^2 = \frac{1}{n}\sum_{i}^{n}\left(f(x,t_i)-\hat{\theta}\right)\left(f(x,t_i)-\hat{\theta}\right)^\top.$$

  This yields an asymptotic expression for the conditional probability given by

$$\hat{p}(t_i|x) = \frac{e^{-(\bar{f}_n-\hat{\theta})^\top\hat{S}_n^{-2}f(x,t_i)}}{\sum_{t_k\in\mathcal{T}}e^{-(\bar{f}_n-\hat{\theta})^\top\hat{S}_n^{-2}f(x,t_k)}}.$$

  The advantage of this expression is that it does not require the computational optimization used for the log-linear model proposed in the MaxEnt method.

- Method 2 : In some situations, for a given context, we can observe another corpus and compute an estimator $\tilde{\theta}$ of θ. In that case, we can use directly this estimator to get the predictive probability

$$\tilde{p}\left(t_i|x\right) = \frac{e^{-(\bar{f}_n-\tilde{\theta})^\top \tilde{S}_n^{-2} f(x,t_i)}}{\sum\limits_{t_k\in\mathcal{T}} e^{-(\bar{f}_n-\tilde{\theta})^\top \tilde{S}_n^{-2} f(x,t_k)}}$$

with

$$\tilde{S}_n^2 = \frac{1}{n}\sum_i^n \left(f(x,t_i)-\tilde{\theta}\right)\left(f(x,t_i)-\tilde{\theta}\right)^\top.$$

Since the number of tags is relatively limited in POS tagging, such computation is almost immediate.

3.3 The penalized version of MaxEnt ($d \geq n$)

A natural question is to propose a method adapted to large dimension constraints in the framework of NLP. For this, consider the problem of generalized empirical likelihood with an L2 penalization.

The penalized empirical divergence in the relative entropy case is

$$\gamma_n(\theta,\lambda) = \mathbb{P}_n\left(-\lambda^\top\left(f\left(x,t\right)-\theta\right) - \varphi(\lambda^\top\left(f\left(x,t\right)-\theta\right))\right) - \frac{1}{2}||\lambda||_R^2.$$

and becomes

$$\gamma_n(\theta,\lambda) = 1 + \frac{1}{n}\sum_{i=1}^n\left(-\exp(\lambda^\top\left(f\left(x_i,t_i\right)-\theta\right)) - \frac{1}{2}||\lambda||_R^2\right).$$

Notice that when $R=\rho_n I$, then this quantity becomes asymptotically for λ close to 0 (as expected),

$$\gamma_n(\theta,\lambda) \approx \frac{1}{n}\sum_{i=1}^n\left(-\lambda^\top\left(f\left(x_i,t_i\right)-\theta\right)) - \frac{1}{2}\lambda^\top(S_n^2+\rho_n I)\lambda\right)$$

whose maximum is attained at

$$\lambda_n^* = -(S_n^2+\rho_n I)^{-1}\frac{1}{n}\sum_{i=1}^n\left(f\left(x_i,t_i\right)-\theta\right) = -(S_n^2+\rho_n I)^{-1}\mathbb{P}_n(f-\theta)$$

yielding the value at the optimum

$$\frac{1}{2}\mathbb{P}_n(f-\theta)^\top(\mathbb{P}_n(f-\theta)(f-\theta)^\top+\rho_n I)^{-1}\mathbb{P}_n(f-\theta),$$

as in χ^2 case (for which the expression was exact).

In the penalized case, we see that the optimal weights depend on θ and are given by

$$\hat{p}\left(t_i|x\right) = \frac{e^{-(\bar{f}_n-\theta)^\top\left(S_n^2+\rho_n I_d\right)^{-2}(f(x,t_i)-\theta)}}{\sum\limits_{t_k\in\mathcal{T}} e^{-(\bar{f}_n-\theta)^\top(S_n^2+\rho_n I_d)^{-2}(f(x,t_k)-\theta)}},$$

where we recall that $S_n^2 = \frac{1}{n}\sum_i^n\left(f(x_i,t_i)-\theta\right)\left(f(x_i,t_i)-\theta\right)^\top$ and $\bar{f}_n = \frac{1}{n}\sum_{i=1}^n f(x_i,t_i)$.

As explained in Paragraph 3.2.1, it is possible to have an estimator of θ based on another corpus and to obtain a plug-in version of this quantity. Thus, the problem will essentially be to have an adequate value for the penalization parameter.

4 Practical implementation of Penalized GEL on the Penn-Treebank corpus

4.1 Preparation of the database

In the following, we will build a POS-tagger, which is based on a penalized maximum entropy principle, that takes as input a sentence, and assigns a grammatical class (or POS-tag) to each word in this sentence, using the "penalization" ideas developed above. To accomplish this, we use the Penn-Treebank corpus, which uses a tagset containing a total of 46 tags, 36 grammatical tags (verbs, nouns, prepositions, etc.), and 10 punctuation tags (comma, closing brackets, etc.). More precisely, the version of the corpus that we are using, is the one included in the Python *nltk* package. It contains 3914 sentences, which represent 100676 tokens (here single words) or 12408 tokens without repetitions. We extract randomly (several times) a sample of size $N = 10000$ from the 100676 initial tokens for memory capacity reasons.

To prepare the database, the first step is to construct two functions:

(1) a context function that takes a tagged sentence in the form of (t_i, w_i) pairs (tag, word), $i = 1$ to the size of the sentence (in term of number of words), as input and returns the same sentence but in the form of (t_i, x_i) pairs (tag, context), where x is a context vector that contains information about the word w as well as its neighboring words within the sentence where it was observed. The information we have retained includes the two words preceding the central word w, the two words following w, whether w is the beginning or end of a sentence, and whether it is a number.

The following table 1 provides an example of transforming the following tagged sentence into (tag, context) pairs instead of (tag, word) pairs.

Pierre	Vinken	,	61	years	old	,	will	join	the	board
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
NNP	NNP	,	CD	NNS	JJ	,	MD	VB	DT	NN

as	a	nonexecutive	director	Nov.	29	.
↓	↓	↓	↓	↓	↓	↓
IN	DT	JJ	NN	NNP	CD	.

Tag ($\mathbf{t_i}$)	**Context ($\mathbf{x_i}$)**								
	w_{i-2}	w_{i-1}	$\mathbf{w_i}$	w_{i+1}	w_{i+2}	Digit	Capital	w_0	w_∞
NNP	0	0	Pierre	Vinken	,	0	1	1	0
NNP	0	Pierre	Vinken	,	61	0	1	0	0
,	Pierre	Vinken	,	61	years	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
VB	,	wil	join	the	board	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
NNP	nonexecutive	director	Nov.	29	.	0	0	0	0
CD	director	Nov.	29	.	0	0	1	0	0
.	Nov.	29	.	0	0	0	0	0	1

Table 1: An example of (Tag, Context) pairs

- w_i represents the central word

- w_{i-1} and w_{i+1} represent respectively the preceding word and the following word by one position
- w_{i-2} and w_{i+2} represent respectively the preceding word and the following word by two positions
- *Digit* and *Capital* check if the central word is a digit or if it starts with a capital respectively
- w_0 is equal to 1 if the central word is a starting word (the first word of the sentence), 0 otherwise.
- w_∞ is equal to 1 if the central word is an ending word (the last word of the sentence) and 0 otherwise.

For instance the pair $(w_{10}, t_{10}) = (\text{the}, \text{DT})$ is transformed into $(t_{10}, x_{10}) = (\text{DT}, x_{10})$ where $x_{10} = (\text{will}, \text{join}, \text{the}, \text{board}, \text{as}, 0, 0, 0, 0)$.

(2) a feature function that takes a context and returns a high-dimensional binary vector. Each component of this vector (or feature) equals 1 if the condition is satisfied and 0 otherwise. To accomplish this, we construct a dictionary of central words, a dictionary of words one position before the central word, a dictionary of words two positions before the central word, and so on. We concatenate these five dictionaries. For a context x_i, we retrieve the "previous word" information. If this word appears in the dictionary of previous words, then the context feature vector will have zero components everywhere except for the position of the word. The conditions are of the form w_i = a particular word from the dictionary of central words, w_{i-1} = a particular word from the dictionary of previous words, etc. That is we create as many dichotomic variables as there are possible sequences of 5 words and select only the one that occurs more than a given threshold.

For example, if the dictionary of words corresponding to two positions before the central word contains 35 words, and the current context being examined contains information w_{i-2} that appears at the 4th position of this dictionary, then the feature vector block corresponding to the words two positions before the central word for this context will be of the form:

$$(0, 0, 0, 1, 0, 0, \ldots, 0, 0)$$

Actually, the position of the 1 does not only indicate the presence of an information related to the word alone, but to the word combined with a tag, which means that there may be two positions (two features) for the same word but with a different tag. Therefore, it should be understood that the features are functions of the pair $f(x_i, t_i)$ and not just functions of the context $f(x_i)$. In the example of POS-Tagging given in ([20]), we see that the word "flies" can have two possible tags (NN and VB). So, for this same word, there will be two different features in the block of the central word, one that activates only when the central word of the context x_i=flies and t_i =NN, and a second feature that activates when the central word of the context x_i =flies and t_i =VB.

features	...	...	...	(flies, NN)	(flies, VB)	...	...	...	...	...
	↓			↓	↓				↓	↓
f(flies, NN)	0	...	0	1	0	0	0	...	0	0
f(flies, VB)	0	...	0	0	1	0	0	...	0	0

We also construct some features that look at pairs (tag, suffix) where the suffix represents the last three or the last two letters of the central word of a given context. We also perform a

filter-based selection of the features to only keep those that are observed more than ten times (since it was sufficient in our case to get good performance, but the filter with a threshold equal to 10 can be modified in other cases).

With the dataset now expressed as a collection of $(t_i, f(t_i, x_i))$ pairs, it provides the basis for model estimation.

4.2 Results

After estimating μ, by the empirical mean

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} f\left(t_i, x_i\right)$$

on the entire initial dataset, we split the dataset into two parts:

- a training sample $\{(t_i, f(t_i, x_i))\}_{i \in \{1,\dots,n\}}$ (representing 75% of the initial dataset)
- a test sample $\{(t_i, f(t_i, x_i))\}_{i \in \{1,\dots,n_0\}}$ (25% of the initial dataset).

On the training dataset, we calculate the empirical mean $\bar{f}_n = \frac{1}{n} \sum_{i=1}^{n} f\left(t_i, x_i\right)$ and the empirical covariance matrix of the features

$$S_n^2 = \frac{1}{n} \sum_{i}^{n} \left(f(x_i, t_i) - \hat{\mu}_N\right) \left(f(x_i, t_i) - \hat{\mu}_N\right)'$$

.

To calibrate the penalty parameter ρ we use the method introduced by Ledoit and Wolf (2000) [24] and further developed in [21]. The main idea is to use a penalized estimator of the empirical covariance matrix (which precisely appears in our expressions of the conditional probabilities of tags) and to choose the estimated version of the penalty that minimizes the Frobenius norm between this estimator and the true covariance matrix. Recall that the modified Frobenius scalar product and its associated norm are defined, for any compatible matrices or vectors A and B, by

$$\langle A, B \rangle = \frac{Tr\left(AB^\top\right)}{d}, \quad \|A\|^2 = \langle A, A \rangle = \frac{Tr\left(AA^\top\right)}{d}.$$

The optimal estimated penalty is then given by

$$\rho_n = \frac{\hat{\beta}_n^2 \hat{\sigma}_n^2}{\hat{\alpha}_n^2} \quad \text{where} \quad \hat{\sigma}_n^2 = \left\langle S_n^2, I_d \right\rangle \; ; \quad \hat{\delta}_n^2 = \left\| S_n^2 - \hat{\sigma}_n^2 I_d \right\|^2 \; ; \quad \hat{\alpha}_n^2 = \hat{\delta}_n^2 - \hat{\beta}_n^2$$

$$\text{with } \bar{\beta}_n^2 = \frac{1}{n^2} \sum_{i=1}^{n} \left\| f(x_i, t_i)(f(x_i, t_i))' - S_n^2 \right\|^2 \text{ and } \hat{\beta}_n^2 = \min\left(\bar{\beta}_n^2, \hat{\delta}_n^2\right).$$

This allows us to estimate the conditional probabilities of each tag given the context x as follows

$$\forall t_i \in \mathcal{T}, \quad \hat{p}\left(t_i | x\right) = \frac{e^{-(\bar{f}_n - \hat{\mu}_N)' \left(S_n^2 + \rho_n I_d\right)^{-2} (f(x, t_i) - \hat{\mu}_N)}}{\sum\limits_{t_k \in \mathcal{T}} e^{-(\bar{f}_n - \hat{\mu}_N)' (S_n^2 + \rho_n I_d)^{-2} (f(x, t_k) - \hat{\mu}_N)}}.$$

Once these probabilities are obtained, the tag assigned to the input context is the one for which the estimated conditional probability is the highest :

$$\forall x_i \in \text{Training set}, \quad \hat{t}_i = \underset{t_k \in \mathcal{T}}{\operatorname{argmax}} \left\{\hat{p}\left(t_k, x_i\right)\right\}.$$

We then estimate the model's error on the test sample by the number of misclassifications :

$$\text{Error} = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{1}_{\{t_i \neq \hat{t}_i\}} \Leftrightarrow \text{Precision} = \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{1}_{\{t_i = \hat{t}_i\}} = 1 - \text{Error}.$$

The same procedure is repeated 10 times on different random samples of size N=10000 drawn without replacement from the initial dataset containing 100676 entries (tag-context pairs). Finally, we achieve an estimation accuracy of 98% (on average over the different training samples) and a prediction accuracy of 95% (on average over the test samples).

The estimated conditional probabilities
Here is an example of the values of the estimated conditional probabilities for the following sentence which is an observed sentence among the training set:

$$(w_1, \ldots, w_{18}) = \text{Pierre Vinken, 61 years} \ldots \text{Nov. 29.}$$

Let's consider $(x_1, \ldots, x_{18})$ the corresponding contexts to each word $(w_i)_{i=1,\ldots,18}$, i.e.:

Pierre	Vinken	,	61	years	...	nonexecutive	director	Nov.	29	.
↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕
$(x_1$	x_2	x_3	x_4	x_5	...	x_{14}	x_{15}	x_{16}	x_{17}	$x_{18})$

The table 2 gives the conditional probabilities of a tag given a context. It gives also the predicted POS-tag for each context which is simply the tag with the highest conditional probabilities. The probability values are rounded.

Tag t	**Conditionnal probabilities** $\mathbb{P}(t\|x_i)$						
	x_1	x_2	x_3	...	x_{16}	x_{17}	x_{18}
NN	0.21	0,02	0,017	...	0.41	0	0
NNS	0	0	0	...	0	0	0
NNP	0.76	0.95	0	...	0.4	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
VBD	0	0	0	...	0	0	0
Predicted POS-tag	NNP	NNP	,	...	NN	CD	.

Table 2: Estimated conditional probabilities

We also constructed a similar model that classifies input text into simple or complex versions (see Chapter 4 of Issouani's thesis [20].

Appendix

Appendix 1: New classifiers

Most divergences used in practice (Kullback, relative entropy, χ^2 and Hellinger among others) are specific cases of the Cressie-Read divergence (see Csiszár 1967). Table 3 (partly taken from Bertail et al. (2007) [6]) displays the corresponding convex functions φ_α and their domains, along with the optimal conditional probability expressions $\hat{p}(t|x)$ arising from each divergence. For the general Cressie-Read family, the domain depends on α, as illustrated by the examples.

Divergences	α	**Convex generator $\boldsymbol{\varphi_\alpha}$**		Classifiers $\hat{p}(t\|x)$	Weights $q_i^* = q^*(Z_i)$
		$\varphi_\alpha^*(x)$ and $\varphi_\alpha(x)$	Domaine		
relative entropy	1	$\varphi_\alpha^*(x) = (x+1)\log(x+1) - x$	$]-1, +\infty]$	$\frac{\exp\left(-\lambda^{*\top}(f(x,t)-\theta)\right)}{\sum_{t'}\exp\left(-\lambda^{*\top}(f(x,t')-\theta)\right)}$	$\frac{1}{n}\exp\left(\lambda^{*\top}g(Z_i,\theta)\right)$
		$\varphi_\alpha(x) = e^x - 1 - x$	$\mathbb{R}$		
Kullback	0	$\varphi_\alpha^*(x) = x - \log(1+x)$	$]-1, +\infty[$	$\frac{(1-\lambda^{*\top}(f(x,t)-\theta))^{-1}}{\sum_{t'}(1-\lambda^{*\top}(f(x,t')-\theta))^{-1}}$	$\frac{1}{n\left(1-\lambda^{*\top}g(Z_i,\theta)\right)}$
		$\varphi_\alpha(x) = -\log(1-x) - x$	$]-\infty, 1[$		
Hellinger	0.5	$\varphi_\alpha^*(x) = 2(\sqrt{x+1}-1)^2$	$]-1, +\infty[$	$\frac{(2-\lambda^{*\top}(f(x,t)-\theta))^{-2}}{\sum_{t'}(2-\lambda^{*\top}(f(x,t')-\theta))^{-2}}$	$\frac{4}{n\left(2-\lambda^{*\top}g(Z_i,\theta)\right)^2}$
		$\varphi_\alpha(x) = \frac{x^2}{2-x}$	$]-\infty, 2[$		
$\boldsymbol{\chi^2}$	2	$\varphi_\alpha^*(x) = \frac{x^2}{2}$	$\mathbb{R}$	$\frac{1+\lambda^{*\top}(f(x,t)-\theta)}{\sum_{t'}(1+\lambda^{*\top}(f(x,t')-\theta))}$	$\frac{1}{n}\left(1+\lambda^{*\top}g(Z_i,\theta)\right)$
		$\varphi_\alpha(x) = \frac{x^2}{2}$	$\mathbb{R}$		
Cressie-Read	α	$\varphi_\alpha^*(x) = \frac{(1+x)^\alpha - \alpha x - 1}{\alpha(\alpha-1)}$	$-$	$\frac{(1+(\alpha-1)\lambda^{*\top}(f(x,t)-\theta))^{\frac{1}{\alpha-1}}}{\sum_{t'}(1+(\alpha-1)\lambda^{*\top}(f(x,t')-\theta))^{\frac{1}{\alpha-1}}}$	$\frac{1}{n}\left[1+(\alpha-1)\lambda^{*\top}g(Z_i,\theta)\right]^{\frac{1}{\alpha-1}}$
		$\varphi_\alpha(x) = \frac{[(\alpha-1)x+1]^{\frac{\alpha}{\alpha-1}} - \alpha x - 1}{\alpha}$	$-$		

Table 3: Divergence, domain, optimal weights q_i^* and associated optimal classifiers $\hat{p}(t|x)$.

Appendix 2: Proofs

The following lemma is taken from Owen (2001) (11.2 and 11.3, pg 225. [36]).

Lemma Let $Z = (Z_i)_{i=1}^n$ be a sequence of independent and identically distributed random variables, and for every $n \in \mathbb{N}$, define $\|Z\|_\infty = \max_{i=1,\ldots,n} |Z_i|$. If $\mathbb{E}\left[Z_1^2\right] < \infty$, then $\|Z\|_\infty = o\left(n^{1/2}\right)$ and $\frac{1}{n}\sum_{i=1}^n |Z_i|^3 = o\left(n^{1/2}\right)$, in probability.

Primal and dual program
Recall that he penalized generalized empirical φ-divergence is given by

$$\gamma_n(\theta, \lambda) = \mathbb{P}_n \left(- \lambda^\top g\left(., \theta\right) - \varphi(\lambda^\top g\left(., \theta\right)) \right) - \frac{1}{2} ||\lambda||_R^2. \tag{10}$$

For sake of simplicity, we choose here, for some $\rho_n > 0$, $R = \rho_n I$ and $||\lambda||_R^2 = \rho_n \lambda^\top \lambda$. This expression becomes in the relative entropy case ($\varphi(x) = e^x - 1 - x$) as follows

$$\gamma_n(\theta, \lambda) = 1 + \frac{1}{n} \sum_{i=1}^n \left(- \exp\left(\lambda^\top \left(g\left(Z_i, \theta\right)\right)\right) - \frac{\rho_n}{2} ||\lambda||_2 \right).$$

Optimisation according to λ
The derivative with respect to λ on the right-hand side of Equation (10) is zero for $j \in \{1, \ldots, d\}$, yielding the conditions

$$\forall j \in \{1, \ldots, d\} \quad 0 = \sum_{i=1}^n g_j\left(Z_i, \theta\right) \left[1 + \varphi^{(1)}\left(\lambda^\top g(Z_i, \theta)\right)\right] + \rho_n \lambda_j.$$

As in Owen's book ([36]), consider $V_i = g(Z_i, \theta)$, so that V_i is centered and has covariance matrix denoted by $\Sigma_d = \mathbb{E}\left[g(Z_1, \theta) g(Z_1, \theta)^\top\right] = \mathbb{E}\left[V_1 V_1^\top\right]$ and define the function g as

$$G(\lambda) = \frac{1}{n} \sum_{i=1}^n V_i \left[1 + \varphi^{(1)}\left(\lambda^\top V_i\right)\right] + \rho_n \lambda = 0. \tag{11}$$

Now let λ_n^* be the solution of $G(\lambda_n^*) = 0$. Define $\lambda_n^* = \|\lambda_n^*\| \xi_n$ where ξ_n is a unit vector $\|\xi_n\|_2 = 1$. It follows that

$$0 = \xi_n^\top G(\lambda_n^*) \Longrightarrow -\xi_n^\top \bar{V}_n = \frac{1}{n} \sum_{i=1}^n \xi_n^\top V_i \cdot \varphi^{(1)}\left(\lambda_n^{*\top} V_i\right) + \rho_n \, \xi_n^\top \lambda_n^*.$$

Consider a point t_i within the interval $\left[0, \xi_n^\top V_i\right]$. A Taylor expansion of $\varphi^{(1)}$ around zero gives

$$\varphi^{(1)}\left(\|\lambda_n^*\| \xi_n^\top V_i\right) = \|\lambda_n^*\| \xi_n^\top V_i \cdot \varphi^{(2)}\left(\|\lambda_n^*\| t_i\right),$$

with $t_i \in \left[0, \xi_n^\top V_i\right]$. Using the fact that $\varphi^{(2)}$ is bounded below by m, we have

$$\begin{aligned}
-\xi_n^\top \bar{V}_n &= \|\lambda_n^*\| \frac{1}{n} \sum_{i=1}^n \left(\xi_n^\top V_i\right)^2 \cdot \varphi^{(2)}\left(\|\lambda_n^*\| t_i\right) + \rho_n \, \xi_n^\top \lambda_n^*, \\
&\geq m \|\lambda_n^*\| \left(\frac{1}{n} \sum_{i : \xi_n' V_i \geq 0}^n \left(\xi_n^\top V_i\right)^2 + \rho_n \right), \\
&\geq m \|\lambda_n^*\| \rho_n.
\end{aligned}$$

The Central Limit Theorem implies

$$-\xi_n^\top \bar{V}_n = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\right).$$

Thus,

$$\|\lambda_n^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}/\rho_n\right).$$

Now, we define

$$\tilde{\lambda}_n = \lambda_n^* + \left(S_n^2 + \rho_n I\right)^{-1} \bar{V}_n, \text{ where } S_n^2 = \frac{1}{n}\sum_{i=1}^n V_i V_i^\top.$$

Then, performing a first-order Taylor expansion of $\varphi^{(1)}$ around zero in $G(\lambda_n^*)$ yields

$$0 = \varphi^{(2)}(0)(S_n^2 + \rho_n I)\tilde{\lambda}_n + \frac{1}{n}\sum_{i=1}^n V_i \alpha_{i,n}$$

where, uniformly in i, we have

$$\|\alpha_{i,n}\| \leq B\left|\lambda_n^{*\top} V_i\right| \leq B\,\|\lambda_n^*\|\,\|V\|_\infty = o_{\mathbb{P}}(1)$$

since $\|V\|_\infty = \max_{i=1,\ldots,n}\|V_i\| = o_{\mathrm{P}}\left(n^{1/2}\right)$ by Lemma 2.2. Finally, since $S_n^2 + \rho_n I$ is bounded below by ρ_n and $\bar{V}_n - \theta = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\right)$, we have

$$\tilde{\lambda}_n = o_{\mathbb{P}}\left(n^{-1/2}\rho_n\right).$$

This means that

$$\lambda_n^* = -\left(S_n^2 + \rho_n I\right)^{-1}\bar{V}_n + o_{\mathbb{P}}\left(n^{-1/2}\rho_n\right)$$

Now we have the expansions (of order 2 of φ)

$$\begin{aligned}
\gamma_n(\theta, \lambda_n^*) &= -n\cdot\lambda_n^{*\top}\bar{V}_n - \sum_{i=1}^n \varphi\left(\lambda_n^{*\top} V_i\right) - \frac{1}{2}\|\lambda_n^*\|_R^2, \\
&= -n\cdot\lambda_n^{*\top}\bar{V}_n - \sum_{i=1}^n\left(\frac{\left(\lambda_n^{*\top} Y_i\right)^2}{2}\varphi^{(2)}(0) + \frac{\rho_n}{2}\lambda_n^{*\top}\lambda_n^* + \tilde{\alpha}_{i,n}\right), \\
&= -n\cdot\lambda_n^{*\top}\bar{V}_n - \frac{n}{2}\left(\lambda_n^{*\top}\left(S_n^2 + \rho_n I\right)\lambda_n^*\right) - \sum_{i=1}^n \tilde{\alpha}_{i,n}\,, \\
&= -n\lambda_n^{*\top}\bar{V}_n - \sum_{i=1}^n \tilde{\alpha}_{i,n} - \frac{n}{2}\left[\tilde{\lambda}_n^\top(S_n^2 + \rho_n I)\tilde{\lambda}_n - 2\tilde{\lambda}_n^\top\bar{V}_n\right. \\
&\quad \left. + \bar{V}_n^\top(S_n^2 + \rho_n I)^{-1}\bar{V}_n\right], \\
&= n\bar{V}_n^\top(S_n^2 + \rho_n I)^{-1}\bar{V}_n - \sum_{i=1}^n \tilde{\alpha}_{i,n} \\
&\quad -\frac{n}{2}\tilde{\lambda}_n^\top(S_n^2 + \rho I)\tilde{\lambda}_n - \frac{n}{2}\bar{V}_n^\top(S_n^2 + \rho_n I)^{-1}\bar{V}_n, \\
&= \frac{n}{2}\bar{V}_n^\top(S_n^2 + \rho_n I)^{-1}\bar{V}_n - \frac{n}{2}\tilde{\lambda}_n^\top(S_n^2 + \rho_n I)^{-1}\tilde{\lambda}_n - \sum_{i=1}^n \tilde{\alpha}_{i,n}.
\end{aligned}$$

We have $\|\tilde{\alpha}_{i,n}\| \leq \widetilde{B}\left|\lambda_n^{*\top} V_i\right|^3$, for some constant $\widetilde{B} > 0$, in probability.

This leads to:

$$\begin{aligned}\left\|\sum_{i=1}^{n}\tilde{\alpha}_{i,n}\right\| &\leq \widetilde{B}^3\left\|\lambda_n^*\right\|^3\sum_{i=1}^{n}\left\|V_i\right\|^3, \\ &= \mathcal{O}_{\mathbb{P}}\left(n^{-3/2}\right)\cdot n\cdot o_{\mathbb{P}}\left(n^{1/2}\right), \\ &= o_{\mathbb{P}}(1).\end{aligned}$$

It follows that

$$\gamma_n(\theta,\lambda_n^*,\rho_n)=\frac{n}{2}\bar{V}_n^{\top}(S_n^2+\rho_n I)^{-1}\bar{V}_n+o_{\mathbb{P}}(1)$$

Notice that this is exactly the expression one would obtain with a penalized χ^2 divergence (without any remainder).

Consider $\mu=[\mu_j]_{\{j=1,\ldots,d\}}$ the eigenvalues of the empirical covariance matrix $S_n^2=\mathbb{P}_n g(.,\theta)g(.,\theta)^{\top}$. Define

$$\forall\rho_n>0,\quad \frac{\mu}{\mu+\rho_n}=\left\{\frac{\mu_j}{\mu_j+\rho_n}\right\}_{j=1,\ldots,d}$$

and the so-called effective dimensions

$$\left\|\frac{\mu}{\mu+\rho_n}\right\|_1=\sum_{j=1}^{d}\left|\frac{\mu_j}{\mu_j+\rho_n}\right| \text{ and } \quad \left\|\frac{\mu}{\mu+\rho_n}\right\|_2=\sqrt{\sum_{j=1}^{d}\left(\frac{\mu_j}{\mu_j+\rho_n}\right)^2}.$$

Now follow the same arguments as in [37], since we assume that the largest eigenvalue of the true covariance matrix is bounded and that $\left\|\frac{\mu}{\mu+\rho_n}\right\|_2\xrightarrow[n\to\infty]{}\infty$ as $d\geq n$ goes to ∞. It follows from their Theorem 1 that

$$\frac{2\gamma_n^*(\theta)-\left\|\frac{\mu}{\mu+\rho_n}\right\|_1}{\sqrt{2\left\|\frac{\mu}{\mu+\rho_n}\right\|_2^2}}\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}(0,1).$$

References

[1] F. Bartolucci. A penalized version of the empirical likelihood ratio for the population mean. *Statistics & probability letters*, 77(1):104–110, 2007.

[2] P. Bertail. Empirical likelihood in some semiparametric models. *Bernoulli*, 12(2):299–331, 2006.

[3] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Exponential bounds for multivariate self-normalized sums. *Electronic Communications in Probability*, 13:628–640, 2008.

[4] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Empirical φ-divergence minimizers for hadamard differentiable functionals. In *Topics in Nonparametric Statistics*, pages 21–32. Springer, 2014.

[5] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Empirical phi-discrepancies and quasi-empirical likelihood: exponential bounds. *ESAIM: Proceedings and Surveys*, 51:212–231, 2015.

[6] P. Bertail, H. Harari-Kermadec, and D. Ravaille. φ-divergence empirique et vraisemblance empirique generalisee. *Annales d'Economie et de Statistique*, pages 131–157, 2007.

[7] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

[8] T. Brants. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[9] E. Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, AAAI '94, pages 722–727, Menlo Park, CA, USA, 1994. American Association for Artificial Intelligence.

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

[11] M. Broniatowski and A. Keziou. Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, 142(9):2554–2573, 2012.

[12] J. Chang, C. Y. Tang, and T. T. Wu. A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185–3216, 2018.

[13] A. Crépet, H. Harari-Kermadec, and J. Tressou. Using empirical likelihood to combine data: application to food risk assessment. *Biometrics*, 65(1):257–266, 2009.

[14] A. Feldman and J. Hana. *A resource-light approach to morpho-syntactic tagging*. Brill, 2010.

[15] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25(1):328–350, 1997.

[16] A. Golan, G. Judge, and D. Miller. Maximum entropy econometrics, robust estimation with limited. 1996.

[17] H. Harari-Kermadec. *Vraisemblance empirique généralisée et estimation semi-paramétrique*. PhD thesis, ENSAE ParisTech, 2006.

[18] H. O. Hartley and J. Rao. A new estimation theory for sample surveys. *Biometrika*, 55(3):547–557, 1968.

[19] N. L. Hjort, I. W. McKeague, and I. Van Keilegom. Extending the scope of empirical likelihood. *The Annals of Statistics*, 37(3):1079–1111, 2009.

[20] E. M. Issouani. *Modèles et algorithmes de simplification automatique de textes*. PhD thesis, université Paris, 2023.

[21] E. M. Issouani, P. Bertail, and E. Gautherat. Exponential bounds for regularized hotelling's t2 statistic in high dimension. *Journal of Multivariate Analysis*, 203:105342, 2024.

[22] A. Keziou. Dual representation of φ-divergences and applications. *Comptes Rendus Mathematique*, 336(10):857–862, 2003.

[23] S. N. Lahiri and S. Mukhopadhyay. A penalized empirical likelihood method in high dimensions. *The Annals of Statistics*, 40(5):2511–2540, 2012.

[24] O. Ledoit and M. Wolf. A well conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2000.

[25] C. Léonard. Convex conjugates of integral functionals. *Acta Mathematica Hungarica*, 93(4):253–280, 2001.

[26] C. Léonard. Minimization of energy functionals applied to some inverse problems. *Applied mathematics and optimization*, 44(3):273–297, 2001.

[27] C. Léonard. Minimizers of energy functionals. *Acta Mathematica Hungarica*, 93(4):281–325, 2001.

[28] F. Liese and I. Vajda. *Convex statistical distances*, volume 95. Teubner, 1987.

[29] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[30] W. D. Meurers. On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of german. *Lingua*, 115(11):1619–1639, 2005.

[31] P. A. Mykland. Dual likelihood. *The Annals of Statistics*, pages 396–421, 1995.

[32] W. K. Newey and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.

[33] T. Otsu. Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis*, 98(10):1923–1954, 2007.

[34] A. Owen et al. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[35] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[36] A. B. Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

[37] H. Peng and A. Schick. Asymptotic normality of quadratic forms with random vectors of increasing dimension. *Journal of Multivariate Analysis*, 164:22–39, 2018.

[38] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *the Annals of Statistics*, 22(1):300–325, 1994.

[39] A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, Philadelphia, PA, USA, 1998. AAI9840230.

[40] A. Ratnaparkhi et al. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA, 1996.

[41] R. Rockafellar. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

[42] R. Serfling. Approximation theorems of. *Mathematical Sratistics*, 1980.

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

[44] A. Taylor, M. Marcus, and B. Santorini. The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer, 2003.

[45] D. R. Thomas and G. L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70(352):865–871, 1975.

[46] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.