Maximum Likelihood Set for Estimating a Probability Mass Function

Bruno M. Jedynak
bruno.jedynak@jhu.edu
Département de Mathématiques,
Université des Sciences et Technologies de Lille, France,
and Center for Imaging Science,
The Johns Hopkins University, Baltimore, MD, U.S.A.

Sanjeev Khudanpur
khudanpur@jhu.edu
Department of Electrical and Computer Engineering,
The Johns Hopkins University, Baltimore, MD, U.S.A.

Revised January 7, 2005

Abstract

We propose a new method for estimating the probability mass function (pmf) of a discrete and finite random variable from a small sample. We focus on the observed *counts* — the number of times each value appears in the sample — and define the Maximum Likelihood Set (MLS) as the set of pmf's that put more mass on the observed counts than on any other set of counts possible for the same *sample size*. We characterize the MLS in detail in this article. We show that the MLS is a "diamond"-shaped subset of the probability simplex $[0,1]^k$ bounded by at most $k \times (k-1)$ hyper-planes, where k is the number of possible values of the random variable. The MLS always contains the empirical distribution, as well as a family of Bayesian

estimators based on a Dirichlet prior, particularly the well known Laplace estimator. We propose to select from the MLS the pmf that is "closest" to a fixed pmf that encodes prior knowledge. When using Kullback-Leibler distance for this selection, the optimization problem comprises finding the minimum of a convex function over a domain defined by linear inequalities, for which standard numerical procedures are available. We apply this estimate to language modeling using Zipf's law to encode prior knowledge, and show that this method permits obtaining state of the art results while being conceptually simpler than most competing methods.

1 Introduction

Let p be a probability mass function (pmf) over a set $\{1, \ldots, k\}$ of finite cardinality. This may represent a set of numerical values for a quantitative variable or a set of indices for a qualitative variable. The later situation is often qualified as non-metric, as will be the case in Section 4, where the indices will refer to words of the English vocabulary.

Suppose that we observe n samples $x_1, \ldots, x_n$, that are independent and identically distributed with common pmf p, which is unknown and needs to be estimated from the observed samples. Prior information may be available about p and, in particular, a specific estimate, or an estimate of a certain form, may be preferred when $n = 0$.

For the case when $n >> k$, a very satisfactory answer is the empirical distribution or type $\hat{p}$, namely:

$$\hat{p}(X = i) \;=\; \hat{p}_i \;=\; \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}(x_t = i) \;\equiv\; \frac{n_i}{n}, \qquad i \in \{1, \ldots, k\}, \quad (1)$$

where $\mathbf{1}(\cdot)$ is an indicator function and, hence, n_i is the number of times the value i is observed in the sample.

When n is small, the pioneering work of Laplace (for $k = 2$) has lead to the well known Bayesian estimates as alternatives to the type. During World War II, while working on cracking German cryptographic systems, Jack Good and Alan Turing invented a method for regularizing the type (Good, 1953; Orlitsky, Santhanam and Zhang, 2003). In their case, $k = 26$ was the number of letters in the Latin alphabet, and $n \approx 100 - 1000$. In section 4, we consider a case where k is the number of words in the English

vocabulary, which is set to about 10^5, and the training sample is $n \approx 10^6$ words. Many "smoothing" techniques, most being variations on the Good-Turing idea, have been compared for such a case by Chen and Goodman (1996) and Chen and Rosenfeld (1999). Excellent empirical performance is obtained by using Good-Turing like estimators. With the exception of the Bayesian estimates, however, there is often only a heuristic justification, and no principled derivation of the estimation formulae.

There have, of course, been numerous studies of the pmf estimation problem since Laplace, and it is not our intention to present a comprehensive survey of the literature here, which begins at least as far back as (Lidstone, 1920), and continues to be an active area of investigation of numerous recent publications (Ristad, 1995; Poschel et al, 2003).

We propose the following new method for estimating p. We consider the *counts* — the number of times each value appears — and define the Maximum Likelihood Set (MLS) as the set of probability mass functions that put more mass on the observed counts than on any other set of counts possible for the given n. In a second step, an element is chosen from this set. It can be the one with maximum entropy, or another based on available prior information. This view of the problem, we believe, is very natural. Indeed, so much so that when we first arrived at this view, we expected that someone, in the time of Laplace or thereafter, had already investigated it. We have however not found any evidence of this in the literature.

1.1 The Empirical Distribution

The empirical distribution, or *type*, of a sample $x_1, \ldots, x_n$, as briefly mentioned earlier, is

$$\hat{p} = \left(\frac{n_1}{n}, \ldots, \frac{n_k}{n} \right), \text{ with } n = \sum_{i=1}^{k} n_i \,, \tag{2}$$

where $n_i, 1 \leq i \leq k$, are the counts, that is the number of times the value i appeared in the sample. We write $\mathcal{P}^k$ the set of pmfs over a set of cardinality k and $\mathcal{P}_n^k$ the set of types with denominator n over a set of cardinality k. The probability, under $p \in \mathcal{P}^k$, of observing $x_1, \ldots, x_n$ is

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{k} p_i^{n_i} \,, \tag{3}$$

where n_i are the counts as above. The right hand side of (3), viewed as a function of the pmf p is called the likelihood function, and may be rewritten as

$$\prod_{i=1}^{k} p_i^{n_i} = 2^{-n(D(\hat{p},p)+H(\hat{p}))}, \tag{4}$$

where

$$D(p,q) = \sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i}, \tag{5}$$

with $0 \log_2 \frac{0}{q} = 0$ and $p \log_2 \frac{p}{0} = \infty$ for $p > 0$, is the *Kullback-Leibler* distance of p from q, and

$$H(p) = -\sum_{i=1}^{k} p_i \log_2 p_i, \tag{6}$$

with $0 \log_2 0 = 0$, is the Shannon entropy of p.

It is clear from (4) that the type $\hat{p}$ is a *sufficient statistic* for estimating p. Also note that $\hat{p}$ is the maximum likelihood estimate of p, i.e. the choice of p for which the likelihood (3) of $x_1, \ldots, x_n$, is maximum. Indeed, $D(\hat{p}, p) \geq 0$, with equality iff $p = \hat{p}$ (cf e.g. Cover and Thomas (1991)).

For k fixed and $n \to \infty$, the type is a strongly consistent and efficient estimate of the pmf. However, the type may not be the best possible estimate for finite n. For example, one may have prior information about the true distribution that is captured in the type only for very large n. There is also a more structural objection: when k is large, there might be many values $1 \leq i \leq k$, for which $p_i << \frac{1}{n}$. In this case, with high probability, we will observe $n_i = 0$. Hence, low probability events tend to be underestimated and high probability events overestimated by $\hat{p}$. One manifestation of this effect is that the expected entropy of the type underestimates the entropy of the original pmf. Indeed,

$$E\left[H(\hat{p})\right] = -E\left[\sum_{i=1}^{k} \hat{p}_i \log \frac{\hat{p}_i}{p_i} p_i\right] = -E\left[D(\hat{p},p)\right] + H(p) \leq H(p).$$

In Section 2, we therefore construct a set of pmfs that contains the type as well as other pmfs that are close to it. In particular, it contains pmfs with larger entropy than the type. We will then choose an estimate from this set based on available prior knowledge.

1.2 Bayesian Estimates

Bayesian analysis offers an alternative to maximum likelihood estimation. The Dirichlet family, indexed by a parameter β, is a family of prior distributions over pmf's given by

$$\pi_\beta(p) = \frac{1}{Z(\beta)} \prod_{i=1}^{k} p_i^{\beta-1}, \qquad p \in \mathcal{P}^k, \quad \beta \in \mathbb{R}, \tag{7}$$

where $Z(\beta)$ is a normalizing constant. Note that for $\beta = 1$, (7) reduces to the *uniform distribution* over $\mathcal{P}^k$. Now, if the Bayesian cost function is quadratic, that is,

$$L(p, q) = \sum_{i=1}^{k} (p_i - q_i)^2, \tag{8}$$

then, the Bayesian estimate corresponding to the Dirichlet prior is the posterior expectation of p given $x_1, \ldots, x_n$, which can be shown to be

$$\hat{p}_\beta(i) = \frac{n_i + \beta}{n + \beta k}, \qquad \forall\, 1 \leq i \leq k. \tag{9}$$

This is often referred to as an "add-β rule". The special case of $\beta \to 0$ yields the maximum likelihood estimate $\hat{p}$, and $\beta = 1$ the so called Laplace rule (cf, e.g. Lidstone (1920)). Estimators with $\beta = 0.5$ and $\beta = \frac{1}{k}$ have also been considered; see Nemenman, Shafee and Bialek (2002) and references therein. Note that all such estimators with $\beta > 0$ assign a strictly positive mass to every value in $\{1, \ldots, k\}$, and they all converge to the type as $n \to \infty$.

We will see that the the set from which we will choose our estimate contains all add-β rules in (9) for $0 \leq \beta \leq 1$.

1.3 Minimax Estimates

An alternative to Bayesian analysis is minimax analysis where one seeks an estimate that would be optimal in the worst case over the underlying model and in average over the observations. More precisely, if p is the underlying model and q an estimate of p, one builds the functional

$$R(q) = \sup_{p=(p_1,\ldots,p_k)} \sum_{n_1,\ldots,n_k; \sum_{i=1}^k n_i = n} \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} L(p, q), \tag{10}$$

For the quadratic cost (8) as well as for the standardized quadratic cost

$$L(p,q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{p_i}, \tag{11}$$

the minimum of $R(q)$ is achieved by an add-β rule, with $\beta = k^{-1}\sqrt{n}$, (Steinhaus, 1957), and $\beta = 0$, (Olkin and Sobel, 1979), respectively.

1.4 Maximum Entropy Estimates

Maximum entropy estimation is another standard solution to data sparseness. Instead of estimating $\hat{p}$, the maximum entropy method first estimates $\hat{p}(A_j) = \hat{a}_j$ for *select sets* $A_j \subset \{1, \ldots, k\}$, for which we have sufficient evidence in the n samples. Fixing the probability of some subsets of $\{1, \ldots, k\}$ in this manner typically under-specifies the pmf of interest, leading to a set $\mathcal{M}$ of *admissible* pmfs

$$\mathcal{M} = \left\{ p \in \mathcal{P}^k \,:\, p(A_j) = \hat{a}_j,\ j = 1, \ldots, J \right\}, \tag{12}$$

in which the estimate $\hat{p}$ is but one member. From this admissible set, the pmf with the highest Shannon entropy is then chosen as the estimate of p. It is well known (see Berger *et al.* (1996)) that the pmf with the maximum entropy has an exponential form:

$$\hat{p}_{\mathrm{ME}}(i) = \frac{1}{Z(\Lambda)} \exp\left\{ \sum_{j=1}^{J} \lambda_j \mathbf{1}(i \in A_j) \right\}, \qquad \forall\, 1 \le i \le k, \tag{13}$$

where the parameters $\Lambda = (\lambda_1, \ldots, \lambda_J)$ are chosen to satisfy the constraints of (12).

It can be shown that for every i, as long as at least one $p \in \mathcal{M}$ satisfies $p_i > 0$, it follows that $\hat{p}_{\mathrm{ME}}(i) > 0$. Thus the maximum entropy estimate is inherently smooth.

There are several heuristics but few principles for selecting the sets A_j or even J. In language modeling, some A_j's are typically singleton, specifying, for instance, the probability of words that have been seen sufficiently often in the sample; some A_j's may contain all words which can take on a certain grammatical part of speech, e.g. adjectives; some A_j's may overlap with others; etc. Therefore, while maximum entropy estimation eliminates the

need for some of the ad hoc assumptions made by other techniques, it leaves open the problem of selecting the sets used to define $\mathcal{M}$.

Another weakness of the classical maximum entropy method, as pointed out by others, is that the specification of $\mathcal{M}$ via equality constraints leads to an *ad hoc* choice for any candidate A_j — either one must constrain its probability to be *exactly* $\hat{a}_j$, or leave it completely unconstrained. This is unsatisfactory. For instance, if one were considering as candidate sets A_j all singleton sets $\{v\}$ then the naive act of including all of them in the definition of $\mathcal{M}$ leads to $\mathcal{M} = \{\hat{p}\}$. On the other hand, leaving out all i for which, say, $n_i = 1$ from the definition of $\mathcal{M}$ may result in an estimate under which $n_i > 0$ and $n_{i'} = 0$, but $\hat{p}_{\mathrm{ME}}(i) = \hat{p}_{\mathrm{ME}}(i')$. Maximum entropy estimation has therefore been proposed with *inequality* constraints (cf e.g. Khudanpur (1995) and Kazama and Tsujii (2003)):

$$\mathcal{M} = \{p \,:\, a_j \leq p(A_j) \leq b_j,\ j = 1, \ldots, J\}\,. \tag{14}$$

To the best of our knowledge, there has not been much discussion in the literature of a principled way to make the choice of a_j and b_j, particularly of a way that depends only on the observed sample, and not on other ad hoc assumptions about p.

Yet another variation on maximum entropy consists of minimizing a functional of the form

$$\sum_{j=1}^{J} \mu_j d\left(p(A_j), \hat{a}_j\right) - H(p)\,, \tag{15}$$

where $d(.,.)$ is some metric of deviation from the constraints of (12), and the parameters $\mu = (\mu_1, \ldots, \mu_J)$ are estimated, usually, from held-out data. Yet another way to relax the constraints in (12) is to note, using convex duality (Berger, Della Pietra and Della Pietra, 1996), that the parameters Λ that satisfy the constraints are exactly the parameters for which the model of (13) assigns maximum likelihood to the observed sample. One may then choose a *penalized likelihood* approach with a regularizing function of Λ. Still, several parameters need to be estimated from held-out data in either case. Several such methods are compared in Chen and Goodman (1996) for the estimation of bigram and trigram language models.

In Section 2, we will seek to provide a principled way of relaxing the linear equality constraints in maximum entropy estimation.

1.5 Good-Turing and Other Held-Out Methods

In Jelinek (1998), page 258, the author asks "how much larger a probability should be assigned to an event observed once than to one not observed at all, or, in general, whether the ratio of probabilities of events observed n and m times, respectively, should really be n/m".

Considering pmfs that put more mass on the observed counts than on any others, which we do in Section 2, will lead to one answer to this question, namely (24). The Good-Turing and other held-out methods answer the question in a different way.

The basic idea is to divide the data into two parts. The first part, called the development set, is used for the collection of counts $\{n_i\}$. The second part, called the held-out set, is used to estimate additional parameters. A typical structure is as follows:

$$\tilde{p}_i = \begin{cases} \alpha \times \frac{n_i}{n} & \text{if } n_i > M, \\ q_i & \text{if } n_i \leq M, \end{cases} \tag{16}$$

where the (usually small) threshold M, and "smoothed" probability estimates q_i, $i = 0, \ldots, M$, are the additional parameters.

The Good-Turing estimate (Good, 1953; Orlitsky, Santhanam and Zhang, 2003; McAllester and Schapire, 2000) is obtained by setting

$$q_i = \frac{r_{n_i+1}}{r_{n_i}} \frac{n_i + 1}{n}, \qquad i \in \{1, \ldots, k\}, \tag{17}$$

where r_c is the number of symbols $j \in \{1, \ldots, k\}$ whose count $n_j = c$. Thus q_i for a symbol i depends not just on its count n_i and n, but on the counts of all other symbols.

Note that if $n_i > n_j$, it is *not* necessarily true that $q_i \geq q_j$, though this frequently holds in practice for symbols with very small counts. In other words, q_i may not respect the rank-ordering implied by the empirical counts $\{n_i\}$, particularly for symbols with large counts. For this reason, the threshold M is often chosen to be small enough so as not to have this undesirable effect. E.g., in language modeling, M is typically chosen to be 10 or less, depending on n. The parameter α is then computed so that $\tilde{p}_i$ sums to unity.

The Good-Turing estimate performs remarkably well for pmf on words. However, its derivation is somewhat ad hoc and unsatisfactory.

2 The Maximum Likelihood Set

One of the simplest and driving ideas in statistics is as follows : what we observe has to be fairly likely, otherwise we would not have observed it. One way to quantify this is to say that what we observe has to be more likely under the true pmf, than any other comparable event. Let's define the Maximum likelihood Set (MLS) as the set of pmf's that put more mass on the observed type than on any other type given n. Let $p = (p_1, \ldots, p_k)$ be a pmf over $\{1, \ldots, k\}$. The p-probability of observing the type $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ is

$$f(p, \hat{p}) = \frac{n!}{n_1! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i} \,. \tag{18}$$

The MLS, with these notations is defined as

$$\mathcal{M}(\hat{p}) = \{p \in \mathcal{P}^k \ : \ \forall\, \hat{q} \in \mathcal{P}_n^k, f(p, \hat{p}) \geq f(p, \hat{q})\} \,. \tag{19}$$

We will see in section 2.3 that this set always contains the type $\hat{p}$, which is the maximum likelihood estimate for p, and that it shrinks down to it as $n \to \infty$. For finite n, it contains pmf's that might reflect prior information such as "smoothness," or other desirable properties in a better way than the type, but still remaining "close" to the observed counts. Moreover, this set is a close convex subset of $\mathcal{P}^k$ opening the way to numerical optimization.

Using Stirling formulae, as well as (4), one can check that

$$f(p, \hat{p}) \doteq 2^{-D(\hat{p},p)}, \text{ where } u_n \doteq v_n \Leftrightarrow \lim_{n\to\infty} \frac{1}{n} \log \frac{u_n}{v_n} = 0 \,. \tag{20}$$

Hence, for n sufficiently large, the MLS associated with a type $\hat{p}$ is roughly

$$\left\{p \in \mathcal{P}^k \ : \ D(\hat{p}, p) \leq D(\hat{q}, p), \quad \forall\, \hat{q} \in \mathcal{P}_n^k\right\} \,, \tag{21}$$

leading to the loose description that the MLS is *the set of pmf's that are "closer" to the observed type than to any other.*

2.1 Characterization of the Maximum Likelihood Set

The MLS admits a simpler, though still implicit representation. Given the observed counts $(n_1, \ldots, n_k)$, define a neighborhood relationship on the set of types with denominator n: the neighbors of $(n_1, \ldots, n_k)$ are the types

obtained by changing a single sample from one value to another one. That is, assume that for a pair of indexes $1 \leq i, j \leq k$, we have $n_j > 0$ and $n_i < n$, then $(n'_1, \ldots, n'_k)$ defined by

$$n'_i = n_i + 1, \quad n'_j = n_j - 1, \quad \text{and} \quad n'_l = n_l \quad l \neq i \text{ or } j, \tag{22}$$

is a neighbor of $(n_1, \ldots, n_k)$.

If a pmf is in the MLS then it has to put more mass on the observed type than on any of its neighbors. It turns out that the converse is also true, which leads to the following result.

Proposition 1. *A pmf $p = (p_1, \ldots, p_k)$ on the set $\{1, \ldots, k\}$ belongs to the MLS $\mathcal{M}(\hat{p})$ associated with the counts $(n_1, \ldots, n_k)$ if and only if*

$$n_j p_i \leq (n_i + 1) p_j, \qquad \forall\, 1 \leq i \neq j \leq k, \tag{23}$$

or equivalently

$$\frac{\hat{p}_i}{\hat{p}_j + \frac{1}{n}} \leq \frac{p_i}{p_j} \leq \frac{\hat{p}_i + \frac{1}{n}}{\hat{p}_j}, \qquad \forall\, 1 \leq i \neq j \leq k, \tag{24}$$

where, by convention, $\frac{a}{0} = +\infty$ whenever $a > 0$.

The proof uses elementary algebra and is relegated to the Appendix.

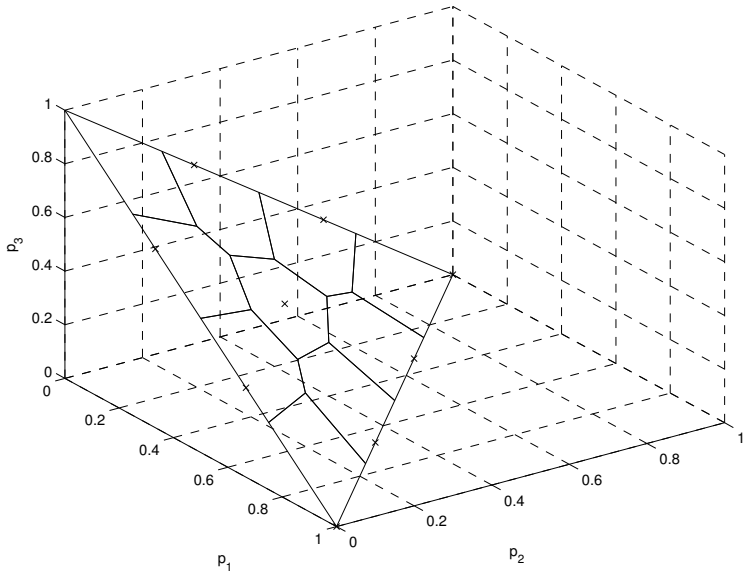
2.2 Motivating Examples

For $k = 2$, the MLS is

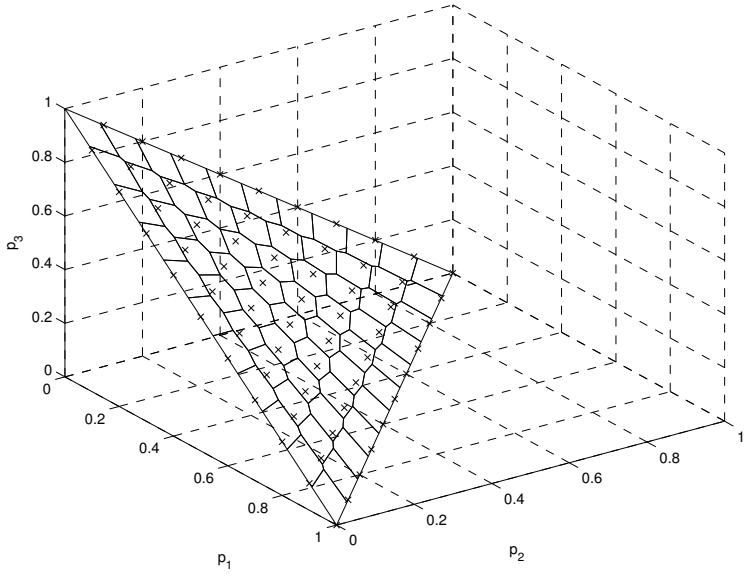
$$\mathcal{M}(\hat{p}) = \mathcal{M}\left(\left(\frac{n_1}{n}, 1 - \frac{n_1}{n}\right)\right) = \left\{p = (p_1, 1 - p_1); \frac{n_1}{n+1} \leq p_1 \leq \frac{n_1 + 1}{n+1}\right\}.$$

Note that this set contains the type and shrinks down to it as the number of samples goes to infinity. Beside the connection with Dirichlet priors mentioned in the introduction, the MLS in this case can be obtained through Bayesian estimation of a proportion with quadratic cost function and a Beta(α, β) prior distribution. It is the set of estimators corresponding to the prior parameters (α, β) satisfying $\alpha + \beta = 1$. See page 368 of Hogg and Craig (1995).

The MLSs for $k = 3$ are illustrated in Figure 1 for two different values of n. The MLSs are convex cells with linear boundaries. They have at most $k \times (k - 1)$ boundaries, one corresponding to each neighboring type.



A



B

Figure 1: Illustration of the Maximum Likelihood Sets for all the possible types for alphabet size $k = 3$. A: $n = 3$ samples. B: $n = 10$ samples. Each "cell" is an MLS comtaining exactly one type marked with a cross.

In order to select an estimate from the MLS, one could choose the pmf with maximum Shannon entropy. This choice will be motivated further in Section 3 below. We use it here to illustrate properties of the MLS set. For example, if the counts $(n_1, \ldots, n_k)$ are made of 0's and 1's only then the pmf selected is the *uniform* distribution over $\{1, \ldots, k\}$, since it is of maximum entropy over all pmf's over $\{1, \ldots, k\}$ and it is included in the MLS, as one can check from (23). In contrast, if there is one value, say the first one, that gets all the counts, then the selected estimate is, for $n > 0$,

$$p_1^* = \frac{n}{n+k-1}, \quad \text{and} \quad p_l^* = \frac{1}{n+k-1}, \quad \forall\, 1 < l \leq k. \tag{25}$$

If $n < k$, then note that $p_1^* \leq 0.5$, which stands in sharp contrast with the estimate $\hat{p}_1 = 1$ given by the type. Equation (25) is a direct consequence of the property (35).

2.3 Properties of the Maximum Likelihood Set

We now present some insightful and useful properties of the MLS.

Proposition 2. *Let $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ be a type. The elements $p = (p_1, \ldots, p_k)$ of the MLS $\mathcal{M}(\hat{p})$ defined by $\hat{p}$ satisfy the following.*

$$\hat{p} \ll p \quad i.e. \quad n_i > 0 \;\Rightarrow\; p_i > 0, \quad \forall\, 1 \leq i \leq k, \tag{26}$$

$$n_i < n_j \;\Rightarrow\; p_i \leq p_j \quad \forall\, 1 \leq i, j \leq k, \tag{27}$$

$$\frac{n}{n+k}\,\hat{p}_i \;\leq\; p_i \;\leq\; \hat{p}_i + \frac{1}{n} \quad \forall\, 1 \leq i \leq k, \tag{28}$$

$$\|p - \hat{p}\|_1 \;=\; \sum_{i=1}^{k} |p_i - \hat{p}_i| \;\leq\; \frac{2(k-1)}{n}, \quad \textit{and} \tag{29}$$

$$\hat{p} \;\in\; \mathcal{M}(\hat{p}) \tag{30}$$

but no other type with denominator n is an element of $\mathcal{M}(\hat{p})$.
If $x_1, \ldots, x_n$ are independent samples with common pmf $q \in \mathcal{P}^k$, then the MLS defined by their type $\hat{p}$ is such that

$$\sup_{p \in \mathcal{M}(\hat{p})} \|p - q\|_1 \to 0 \quad \textit{as} \quad n \to \infty \quad \textit{with probability } 1. \tag{31}$$

Proposition 2 is essentially a corollary of Proposition 1. Details of the proof are in the Appendix. Properties (26) and (27) are desirable for any estimate of the pmf generating $x_1, \ldots, x_n$. Properties (28) and (29) show how the elements of the MLS may deviate from the underlying type. Property (31) shows that for a fixed k, as n gets large, all the elements in the MLS get closer to the pmf generating the samples.

It is easy to see, by comparing (28) and (9), that the MLS contains the Bayesian estimates for $0 \leq \beta \leq 1$.

3 Selecting an Element from the Maximum Likelihood Set

Every pmf in the MLS satisfies a number of properties, as outlined above, that one would consider desirable in an estimate of the pmf generating the samples $x_1, \ldots, x_n$, and we advocate $\mathcal{M}(\hat{p})$ as an admissible set from which a particular pmf may be selected using secondary criteria. One such criterion is outlined next.

Proposition 3. *Let $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ be a type and $\mathcal{M}(\hat{p})$ its associated MLS. Let $q = (q_1, \ldots, q_k)$ be a pmf such that $\hat{p} << q$. Then, there exists a unique element $p^* \in \mathcal{M}(\hat{p})$ such that*

$$D(p^*, q) = \min_{p \in \mathcal{M}(\hat{p})} D(p, q)\,. \tag{32}$$

Note from (23) that $\mathcal{M}(\hat{p})$ is convex and closed in the Euclidean topology on $\mathcal{P}^k$. The existence of p^* therefore follows from Theorem 2.1 in (Csiszar, 1975) and the uniqueness follows from the convexity of $p \mapsto D(p, q)$.

The pmf q may be viewed as a means of incorporating a prior estimate in the estimation process. In the case when $n \gg k$, the MLS has a very small radius, and the choice of q has a negligible effect on the choice of p^*. In the limit as $n \to 0$, $p^* \to q$ by continuity. Therefor, in the small sample situation, the choice of q will greatly influence p^*.

One may choose for q the uniform pmf over $\{1, \ldots, k\}$. p^* is then the element of $\mathcal{M}(\hat{p})$ with maximum Shannon entropy. It has been argued by Nemenman, Shafee and Bialek (2002) that entropy might be the non-metric (categorical data) analog of smoothness. Other compelling arguments for this choice have been made by Jaynes (1994).

In a situation where one needs to estimate a conditional pmf $p(\cdot|y)$ and the marginal pmf $p(\cdot)$ is known, a viable prior estimate is $q(\cdot) = p(\cdot)$. See Jelinek (1998) for related smoothing methods in language modeling.

If one chooses a measure such as the Kullback-Leibler distance to select a pmf from the MLS, an additional satisfactory property of the selected pmf emerges.

Proposition 4. *Let $\mathcal{M}(\hat{p})$ be the MLS defined by the counts $(n_1, \ldots, n_k)$. For any pmf $q \gg \hat{p}$, the pmf*

$$p^* \quad = \quad \arg \min_{p \in \mathcal{M}(\hat{p})} D(p\|q) \tag{33}$$

has the "monotonicity" property:

$$n_i = n_j \quad \textit{and} \quad q_i \geq q_j \qquad \Rightarrow \qquad p_i^* \geq p_j^* \qquad \forall\, 1 \leq i \neq j \leq k. \tag{34}$$

Furthermore,

$$n_i = n_j \quad \textit{and} \quad q_i = q_j \qquad \Rightarrow \qquad p_i^* = p_j^* \qquad \forall\, 1 \leq i \neq j \leq k. \tag{35}$$

The proof is again relegated to the Appendix.

Every pmf $p \in \mathcal{M}(\hat{p})$ has been shown, via (27), to be *faithful to the evidence*. The monotonicity property (34) characterizes the selection rule of Proposition 3: if i is *a priori* more likely than j, then, *in the absence of evidence to the contrary*, it continues to be more likely under the selected p^*. The special case (35) has significant implications for the numerical computation of p^* as will be discussed in the following section.

Note that the Kullback-Leibler divergence of (32) is not the only "distance" one may use to select an pmf from the MLS. Any other function $D(\cdot, \cdot)$ with a projection theorem that guarantees the existence and uniqueness of p^* in (32), together with an algorithm that computes the projection, may be used. An obvious choice is the Euclidean distance, which leads to a standard quadratic programming problem.

3.1 Numerical Optimization Issues

The optimization problem (32) cannot, in general, be solved in closed-form and, in practice, requires a numerical procedure. The setting is known in numerical optimization literature as "general linearly constrained optimization"

(cf Fletcher (1981), Chapter 11, and Bazaraa, Sherali and Shetty (1993)). Stated briefly, one needs to minimize a convex function over a domain defined by linear inequalities such as (23). We minimize the Kullback-Leibler distance of (32) subject to p satisfying (23) using the numerical optimization package CFSQP developed by Lawrence, Zhou and Tits (1997).

The number of constraints specifying the MLS is $k(k-1)$. A typical language modeling situation requires a vocabulary of $k \approx 10^5$ words. Checking just once that a pmf is inside the domain therefore may in general require about 10^{10} operations! Fortunately, choosing q to be piecewise constant considerably reduces the dimensionality. To see this, consider the extreme situation where q is the uniform pmf. Two indexes $1 \leq i, j \leq k$ may be considered equivalent if $n_i = n_j$, and the optimization may be performed over the set of pmf's on $\{1, \ldots, k\}$ *modulo* this equivalence relation, thanks to (35). What is the number of indexes in this set? With n samples, it contains no more than $\sqrt{2n}$ indexes. This is therefore the "effective" k when q is uniform. For other pmfs q, the corresponding equivalence relation is $n_i = n_j$ together with $q_i = q_j$.

4 Language Modeling

Statistical language models are a key component in applications such as automatic speech recognition, machine translation, spelling correction, and document retrieval. Language modeling entails estimating a probability distribution over word-sequences, and this is typically done by modeling the sequence of words in a sentence by a finite memory Markov chain. An n-gram model is a set of conditional pmfs $P(w_n|w_1, \ldots, w_{n-1})$, one for every conditioning event. In applications such as document retrieval, where word-order is not of paramount importance and a bag-of-words representation is adequate, i.i.d. models, called unigram models, are used. In all cases, there is a need to estimate a pmf, marginal or conditional, on the vocabulary. In this section, we present experimental results for the estimation of unigram models.

If obtaining smooth estimates is the primary goal, one would naturally use the uniform distribution in the role of q in (32). We obtain empirical results for this (maximum entropy estimation) case as a first step. It should be clear to the reader, however, that all words are not equally likely even a priori, and it is known from several studies that the count n_i and the *rank* of

a word i, when the vocabulary is sorted in order of decreasing counts, has a roughly inverse relationship. The relationship, sometimes called Zipf's law, cf (Li, 1999), makes for a natural *prior estimate* q for estimating the unigram pmf. via (32). Specifically, we consider

$$q_{\text{Zipf}}(i) = \frac{\alpha(k)}{\text{rank}(i)}. \tag{36}$$

where $\alpha(k)$ is a normalizing constant. Empirical studies (Ha, Sicilia, Ming and Smith, 2002) show that this is a good initial estimate for unigrams. Note that α need not be computed, since it plays no role in the minimization of (32). The resulting estimate p^* in the MLS may then be interpreted as *the pmf supported by the evidence $x_1 \ldots, x_n$, which is closest to Zipf's law in the sense of K-L divergence.* This seems a plausible choice for language modeling.

A problem however remains, that for a given vocabulary, there is no *a priori* way of determining the rank-ordering of words. One could possibly use word-length to perform such ordering. We take a simpler approach and use the rank-ordering empirically observed in $x_1 \ldots, x_n$ to determine q. We make a further modification to break ties: all words which have the same count in $x_1 \ldots, x_n$ get a rank, namely the mean of the ranks spanned by those equal-count words. This latter modification results in an important numerical simplification. By assuming words with the same observed counts to have the same q-probability, we are assured that they will have the same p^* probability, reducing the number of free variables in the numerical optimization of (32) and indeed the specification of p^*. Without this modification, p^* would have up to $k-1$ free parameters, and in case of a most language models this is impractical.

We have conducted experiments on English text from the Wall Street Journal corpus, which contains articles from the general news and financial domain. A particular subset of this corpus, called the UPenn Treebank corpus (The Penn Treebank Project, 1992), has been widely used by many researchers in language modeling, and we use this for our experiments as well. The corpus is divided into sections, numbered 00 through 24. We use sections 00-20 as our training corpus, it contains 900K word tokens. Sections 21-22, containing 100K tokens are used variably as a training or a held-out corpus as needed, and finally sections 23-24, containing 100K tokens make up our test corpus. For the purpose of studying the variability of the estimates, we

divided sentences in sections 00-22 into 10 roughly equal parts, and results will be presented on these smaller corpora in the following.

We made a list of all seen words from sections 00-22 and augmented this vocabulary with a set of "unseen" words. The decision on how many unseen words to include is presently ad-hoc. We use a leave-one-out estimate of the number of unseen words by asking, for each x_t in $x_1 \ldots, x_n$, whether it would be an unseen word if the vocabulary were to be extracted from $\{x_1 \ldots, x_{t-1}, x_{t+1}, \ldots, x_n\}$, $t = 1, \ldots, n$. It is easy to see that this procedure yields $n_0 = n_1$; i.e. the number of unseen words is exactly equal to the number of words seen only once in the corpus. This procedure, while not theoretically satisfactory, is performed out of necessity.

We remark that the MLS of (19) is well defined even for an infinite vocabulary, and with a suitable prior estimate q, it may possible to let the vocabulary-size be unbounded for the estimate of (32) as well.

4.1 Empirical Results

The box at the top of Figure 2 illustrates, using crosses, the empirical pmf $\hat{p}$ obtained from sections 00-22, where the words have been (re)ordered along the abscissa in decreasing order of $\hat{p}_i$. Specifically, for $i = 1, \ldots, k_0$, the ordinate shows the logarithm (to the base 2) of

$$\frac{n_{\sigma(1)}}{n}, \ldots, \frac{n_{\sigma(k_0)}}{n}, \tag{37}$$

with $n_{\sigma(1)} \geq \ldots \geq n_{\sigma(k_0)}$. $k_0 = 37001$ is the number of distinct words seen in sections 00-22. The Zipf prior of (36) is shown in the same box using *dots*: it is a straight line with slope -1. A uniform prior would be a horizontal line on this plot. Finally, in the same box, the lower and upper bounds on each p_i in the MLS, per (28), are also illustrated using a *solid* and a *dashed* line respectively:

$$\left\{ \left(\log i, \log \frac{n_{\sigma(i)}}{n+k} \right) \quad \text{and} \quad \left(\log i, \log \frac{n_{\sigma(i)}+1}{n} \right), \quad 1 \leq i \leq k_0 \right\}, \tag{38}$$

where the number of words in the vocabulary $k = 52743$ is estimated using the procedure described above. Note that the envelope of the MLS has a trumpet-like shape. For large counts, the upper bound of the MLS is essentially indistinguishable from the type. The estimated pmf p^* may decrease the mass for these outcomes, but cannot increase it significantly. However,

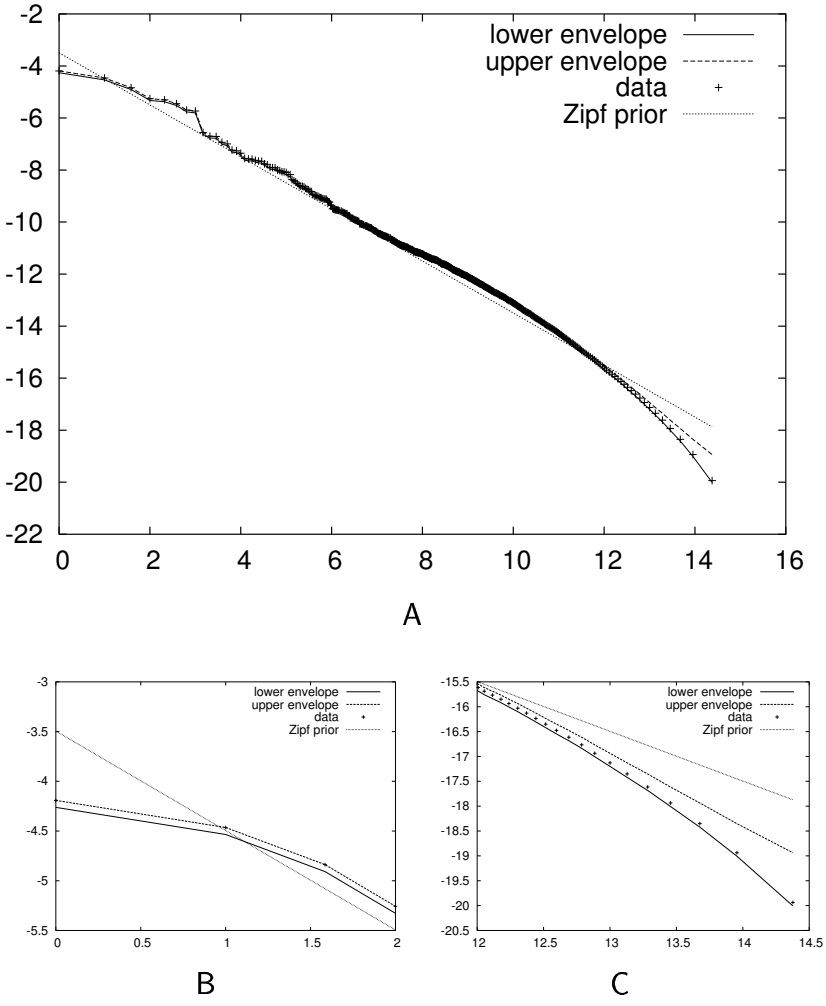


Figure 2: Plot of the empirical pmf from data, the Zipf prior and the lower and upper envelopes of the MLS on a log-log scale. A: Full range of observed counts. B: Zoom top left (≡ high counts). C: Zoom bottom right (≡ low counts).

for small counts, the envelope of the MLS has a flared bell shape showing the statistical variability of the corresponding probabilities, and that the type tends to underestimate rare events. Any pmf chosen from the MLS corresponds to a curve that lies between the upper and lower envelopes.

To measure the efficacy of an estimate $\tilde{p}$ of p, we compute the average codeword length (in bits) that the estimate $\tilde{p}$ achieves on the type $\hat{p}_T$ of the test set, that is

$$\ell(\tilde{p}) \;=\; \frac{1}{n_T}\sum_{t=1}^{n_T}\log\frac{1}{\tilde{p}(x_t)} \;=\; D(\hat{p}_T\|\tilde{p}) + H(\hat{p}_T)\,, \tag{39}$$

where n_T is the size of the test set, the x_t's are the words of the test set and $H(\cdot)$ the Shannon entropy.

Experimental results, for the Wall Street Journal data, along with standard deviations, when avaiblable, are shown in Table 1.

	$\hat{p}_\beta\ \beta=1$	$\hat{p}_\beta\ \beta=\frac{1}{2}$	$\hat{p}_\beta\ \beta=\frac{1}{k}$	$\hat{p}_{\rm GT}$
$\ell(\cdot)$	10.21	10.21	10.52	10.19
	p^* : q=unif	p^* : q=Zipf	p^* : q=$\hat{p}_{\rm GT}$	
$\ell(\cdot)$	10.21	10.20	10.19	

	$\hat{p}_\beta\ \beta=1$	$\hat{p}_\beta\ \beta=\frac{1}{2}$	$\hat{p}_\beta\ \beta=\frac{1}{k}$	$\hat{p}_{\rm GT}$
avg. $\ell(\cdot)$	10.58	10.42	11.31	10.37
std. dev.	0.017	0.017	0.036	0.016
	p^* : q=unif	p^* : q=Zipf	p^* : q=$\hat{p}_{\rm GT}$	
avg. $\ell(\cdot)$	10.58	10.40	10.37	
std. dev.	0.015	0.017	0.018	

Table 1: codeword length in bit for pmf estimates: Upper Table : $n = 10^6$ words. Lower Table : average and standard deviation over 10 training sets with $n = 10^5$ words. $\hat{p}_\beta$ is the add-β rule of (9). $\hat{p}_{\rm GT}$ is the Good-Turing estimate of (16) and (17). p^* is the MLS estimate of (32) with the prior q as indicated. "avg." stands for average and "std. dev." stands for standard deviations.

Looking at the average codeword lengths in Table 1, the reader unfamiliar with language modeling might be surprised to see how well the Good-Turing estimate (fifth column) performs compared to the add-β rules. Three MLS-derived estimates are presented. In the first of the latter, we have used

the uniform pmf as a prior. The estimate thus obtained has comparable performance with the add-1 rule but not as good as the add-$\frac{1}{2}$ rule for the smaller training set. Next, using a Zipf prior, we increase the performance to outperform all add-β rules considered so far and come closer to the GT estimate. Third, we use the GT estimate itself as a prior! We then get an average codeword length that is indistinguishable from the GT estimate. In our experiments, the GT estimate has never been inside the MLS. We have thus shown empirically that there exist pmfs that are "closer" to the empirical pmf than to any other type whose codeword lengths are undisdinguishable from those of the GT estimate. Furthermore, unlike the GT estimate, these pmfs are guaranteed not to contradict the observed counts in the data.

Note as an aside that the "effective-k" for numerical optimization is about 600 for $n = 10^6$ and about 180 when $n = 10^5$ for all priors used.

5 Conclusion

We have proposed a new method for estimating a probability mass function from a sample: we consider the observed counts; the Maximum Likelihood Set is defined as the set of pmf's that put more mass on the observed counts than on any other set of counts; the closest element from the MLS to a prior estimate in the Kullback-Leibler sense is then selected.

The MLS is an "admissible set" for estimating a pmf that has the following properties: it is built from first principles, it is strongly consistent (31) and faithfull to the evidence (26)(27).

The way we select a pmf from the MLS permits to encode domain specific information in a very natural way as demonstrated with the Zipf law for language modeling. Moreover, it is practical as it entails minimizing a convex function over a domain defined by linear inequalities. This is a classical problem in numerical analysis, with known solutions. This way of incorporating domain information is a novel alternative to Bayesian or minimax methods.

Experiments with pmf's on English words show that the proposed method is competitive with state of the art methods.

Acknowledgments

The authors thank Ali Yazgan for his valuable assistance in the use of the CFSQP package and in conducting most of the empirical studies in Section 4.1.

This research was partially supported by the National Science Foundation via Grant No ITR-0225656 and IIS-9982329, ARO DAAD19/-02-1-0337 and general funds from the Center for Imaging Science at The Johns Hopkins University.

Finally, we are grateful to the anonymous referees, who gave several insightful suggestions toward the improvement of this paper.

A Appendix

Proof of Proposition 1. First, we establish that if $p \in \mathcal{M}(\hat{p})$ then p satisfies (23). Towards this end, for any i and any $j \neq i$ such that $n_j > 0$, let

$$\hat{q} = \left(\frac{n_1}{n}, \ldots, \frac{n_i + 1}{n}, \ldots, \frac{n_j - 1}{n}, \ldots, \frac{n_k}{n} \right). \tag{40}$$

By definition, $f(p, \hat{p}) \geq f(p, \hat{q})$, and hence

$$\begin{aligned} & \frac{n!}{n_1! \cdots n_i! \cdots n_j! \cdots n_k!} \prod_l p_l^{n_l} \\ & \quad \geq \quad \frac{n!}{n_1! \cdots (n_i+1)! \cdots (n_j-1)! \cdots n_k!} p_i^{n_i+1} p_j^{n_j-1} \prod_{l \neq i,j} p_l^{n_l} \\ \frac{1}{n_j} p_j & \geq \frac{1}{n_i + 1} p_i. \end{aligned}$$

The property (23) follows. If $n_j = 0$, then (23) follows trivially.

Next, we establish that if p satisfies (23) then $p \in \mathcal{M}(\hat{p})$. Towards this end, again, let

$$\hat{q} = \left(\frac{\tilde{n}_1}{n}, \ldots, \frac{\tilde{n}_k}{n} \right) \tag{41}$$

be an empirical pmf associated with any other set of counts $(\tilde{n}_1, \ldots, \tilde{n}_k)$ for an n-length sample. We construct a sequence of pmf's $\hat{q}^{(0)}, \ldots, \hat{q}^{(n)}$ such that

$$\hat{q}^{(0)} = \hat{q}, \qquad f(p, \hat{q}^{(0)}) \leq f(p, \hat{q}^{(1)}) \leq \ldots \leq f(p, \hat{q}^{(n)}) \qquad \text{and} \qquad \hat{q}^{(n)} = \hat{p}. \tag{42}$$

In particular, we begin with $\hat{q}^{(0)}$ defined by the counts

$$\left(n_1^{(0)}, \ldots, n_k^{(0)}\right) = (\tilde{n}_1, \ldots, \tilde{n}_k)\,, \tag{43}$$

and, for $m = 1, \ldots, n$,

- if $\hat{q}^{(m-1)} = \hat{p}$, then we set $\hat{q}^{(m)} = \hat{q}^{(m-1)}$,
- otherwise, choose i and j such that $n_i^{(m-1)} > n_i$ and $n_j^{(m-1)} < n_j$, and define $\hat{q}^{(m)}$ by the counts

  $$n_i^{(m)} = n_i^{(m-1)} - 1, \quad n_j^{(m)} = n_j^{(m-1)} + 1, \quad \text{and} \quad n_l^{(m)} = n_l^{(m-1)} \text{ for all other } l. \tag{44}$$

  Note that a suitable pair i, j is guaranteed to exist whenever $\hat{q}^{(m-1)} \neq \hat{p}$.

It is clear that for $m = 1, \ldots, n$, if $\hat{q}^{(m-1)} \neq \hat{p}$, then by construction

$$\left\|\hat{q}^{(m)} - \hat{p}\right\|_1 \quad = \quad \left\|\hat{q}^{(m-1)} - \hat{p}\right\|_1 - \frac{2}{n} \quad = \quad \cdots \quad = \quad \left\|\hat{q}^{(0)} - \hat{p}\right\|_1 - \frac{2m}{n}.$$

Since $\|\hat{q} - \hat{p}\|_1 \leq 2$, it follows that $\hat{q}^{(n)} = \hat{p}$.

Finally, note that for $m = 1, \ldots, n$, if $\hat{q}^{(m-1)} \neq \hat{p}$

$$\begin{aligned}
\frac{f(p, \hat{q}^{(m)})}{f(p, \hat{q}^{(m-1)})} &= \frac{n!}{n_1^{(m)}! \cdots n_k^{(m)}!} \frac{n_1^{(m-1)}! \cdots n_k^{(m-1)}!}{n!} \prod_{l=1}^{k} p_l^{n_l^{(m)} - n_l^{(m-1)}} \\
&= \frac{1}{n_j^{(m-1)} + 1} \frac{n_i^{(m-1)}}{1} \frac{p_j}{p_i} \\
&\geq \frac{n_i + 1}{n_j} \frac{p_j}{p_i} \\
&\geq 1,
\end{aligned}$$

where the first inequality holds by construction, since $n_i^{(m-1)} > n_i$ and $n_j^{(m-1)} < n_j$, and the second inequality holds due to (23). □

Proof of Proposition 2. Let's suppose that there is an index $1 \leq j \leq k$ such that $n_j > 0$ and $p_j = 0$. Replacing in equation (23), it implies that $\forall 1 \leq i \leq k, i \neq j, p_i = 0$ which is impossible since $p_j = 0$. This proves (26). Equation (27) is also a consequence of (23) as the reader can check. Remark that (23)

still hold for indexes $i = j$. Then, summing out, we obtain, for any subset $A \subset \{1, \ldots, k\}$,

$$\sum_{i \in A} \sum_{j=1}^{k} \hat{p}_j p_i \leq \sum_{i \in A} \sum_{j=1}^{k} (\hat{p}_i + \frac{1}{n}) p_j \text{ and} \tag{45}$$

$$\sum_{j \in A} \sum_{i=1}^{k} \hat{p}_j p_i \leq \sum_{j \in A} \sum_{i=1}^{k} (\hat{p}_i + \frac{1}{n}) p_j \tag{46}$$

from which we obtain

$$\forall A \subset \{1, \ldots, k\}, \hat{p}(A) \frac{n}{n+k} \leq p(A) \leq \hat{p}(A) + \frac{\#A}{n}, \tag{47}$$

where $\#A$ is the number of elements in A. Setting $A = \{i\}$ gives (28). Now, from (Cover and Thomas, 1991) page 300,

$$\|p - \hat{p}\|_1 = 2(p(A) - \hat{p}(A)); A = \{1 \leq i \leq k; p_i > \hat{p}_i\}. \tag{48}$$

Using (47), we obtain

$$\|p - \hat{p}\|_1 \leq 2 \frac{\#A}{n} \leq \frac{2(k-1)}{n}. \tag{49}$$

Using equation (23), one can directly check that $\hat{p} \in \mathcal{M}(\hat{p})$. Now, if another type in $\mathcal{P}_n^k$ is also an element of $\mathcal{M}(\hat{p})$, then $\hat{p}$ has a neighbor who is an element of $\mathcal{M}(\hat{p})$, following the argument in the part ($\Leftarrow$) of the proof of Proposition 1. Let's call $\hat{q}$ this neighbor. It is such that $\hat{q}_i = \frac{n_i+1}{n}$ and $\hat{q}_j = \frac{n_j-1}{n}$ for some indexes $1 \leq i, j \leq k$ such that $n_i < n$ and $n_j > 0$. Now, as an element of $\mathcal{M}(\hat{p})$, it satisfies

$$(n_i + 1)\hat{q}_j \geq n_j \hat{q}_i. \tag{50}$$

But this is equivalent to say that $n_i \leq -1$ which is impossible.

Finally,

$$\sup_{p \in \mathcal{M}(\hat{p})} \|p - q\|_1 \leq \frac{2(k-1)}{n} + \|\hat{p} - q\|_1, \tag{51}$$

using the triangular inequality as well as the bound (29). (31) follows from the fact that the type converges to the true distribution in $\|.\|_1$. □

Proof of Proposition 4. Assume, to the contrary, that $p_i^* < p_j^*$ for some $i \neq j$ with $n_i = n_j$ and $q_i \geq q_j$. Define a pmf p^{**} by

$$p_l^{**} = \begin{cases} p_j^* & \text{for } l = i, \\ p_i^* & \text{for } l = j, \\ p_l^* & \text{for } l \neq i \text{ or } j. \end{cases} \tag{52}$$

In other words, construct p^{**} by "switching" the i-th and the j-th entries of p^*. Since $p^* \in \mathcal{M}(\hat{p})$, p^* satisfies (23). But $n_i = n_j$ then implies that, by construction, p^{**} also satisfies (23). Thus $p^{**} \in \mathcal{M}(\hat{p})$. Next, note that

$$\begin{aligned} D(p^*\|q) - D(p^{**}\|q) &= \sum_{l=1}^{k} p_l^* \log \frac{p_l^*}{q_l} - \sum_{l=1}^{k} p_l^{**} \log \frac{p_l^{**}}{q_l} \\ &= p_i^* \log \frac{p_i^*}{q_i} + p_j^* \log \frac{p_j^*}{q_j} - p_j^* \log \frac{p_j^*}{q_i} - p_i^* \log \frac{p_i^*}{q_j} \\ &= p_i^* \log \frac{q_j}{q_i} - p_j^* \log \frac{q_j}{q_i} \\ &= \left(p_i^* - p_j^*\right) \log \frac{q_j}{q_i} \quad \geq 0 \end{aligned}$$

which contradicts Proposition 3, since p^* is the unique minimizer of $D(p\|q)$ in $\mathcal{M}(\hat{p})$. □

References

Bazaraa, M.S., Sherali, H.D. and Shetty, C. (1993). *Nonlinear Programming*. Wiley.

Berger, A.L., Della Pietra, S.A. and Della Pietra, V.J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**, 39–71.

Chen, S.F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 310–318.

Chen, S.F. and Rosenfeld, R. (1999). A gaussian prior for smoothing maximum entropy models. Tech. rep., Carnegie Mellon University.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. John Wiley and Sons, inc.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, **3**, 146–158.

Fletcher, R. (1981). *Practical Methods of Optimization*, vol. 2: constrainted optimization. John Wiley and Sons, Ltd.

Good, I.J. (1953). The populations frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Ha, L.Q., Sicilia, E., Ming, J. and Smith, F.J. (2002). Extension of zipf's law to words and phrases. In *International Conference on Computational Linguistics (COLING'2002)*, 315–320, Taipei, Taiwan.

Hogg, R.V. and Craig, A.T. (1995). *Introduction to Mathematical Statistics*. Prentice-Hall, Inc.

Jaynes, E.T. (1994). Probability theory: The logic of science. http://omega.math.albany.edu:8008/JaynesBook.html.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. The MIT Press.

Kazama, J. and Tsujii, J. (2003). Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 137–144.

Khudanpur, S. (1995). A method of ME estimation with relaxed constraints. In *Proceedings of the Johns Hopkins University Language Modeling Workshop*, 1–17.

Lawrence, C.T., Zhou, J.L. and Tits, A.L. (1997). User's guide for cfsqp version 2.5: A c code for solving (large scale) constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality constraints. Tech. Rep. TR-94-16r1, Institute for Systems Research, University of Maryland.

Li, W. (1999). On-line references on zipf's law. http://linkage.rockefeller.edu/wli/zipf/.

Lidstone, G. (1920). Note on the general case of the Bayes-Laplace formula for inductive or posterior probabilities. *Trans Fac. Actuaries*, 182–192.

McAllester, D. and Schapire, R.E. (2000). On the convergence rate of good-Turing estimators. In *Proc. 13th Annual Conference on Computational Learning Theory*, 1–6, Morgan Kaufmann, San Francisco.

Nemenman, I., Shafee, F. and Bialek, W. (2002). Entropy and inference, revisited. In T.G. Dietterich, S. Becker and Z. Ghahramani, eds., *Advances in Neural Information Processing Systems 14*, MIT Press, Cambridge, MA.

Olkin, I. and Sobel, M. (1979). Admissible and minimax estimation for the multinomial distribution and k independent binomial distributions. *Annals of Mathematical Statistics*, 284–290.

Orlitsky, A., Santhanam, N.P. and Zhang, J. (2003). Always good turing: Asymptotically optimal probability estimation. *Science*, **302**.

Poschel et al, T. (2003). Correction algorithm for finite sample statistics. *Eur. Physics*, 531–541.

Ristad, E.S. (1995). A natural law of succession. Tech. Rep. CS-TR-495-95, Department of Computer Science, Princeton University.

Steinhaus, H. (1957). The problem of estimation. *Annals of Mathematical Statistics*, 633–648.

The Penn Treebank Project (1992). http://www.cis.upenn.edu/~treebank/home.html.