Small Sample p.m.f. Estimation and an Application to Language Modeling

Bruno Jedynak
Dept. of Applied Mathematics and Center for Imaging Science
with **Sanjeev Khudanpur, Ali Yazgan, Damianos Karakos**
Center for Language and Speech Processing
The Johns Hopkins University
bruno.jedynak@jhu.edu
http://cis.jhu.edu/∼bruno

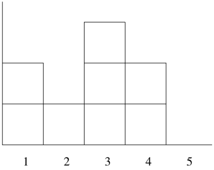
Setting

$x_1, \ldots, x_n$ an iid sample of size n of an unknown distribution [or point mass function (pmf)] p.
Each $x_i \in \{a_1, \ldots, a_k\}$ finite alphabet of numbers, species, words, DNA patterns, ...
we'll use $\{1, \ldots, k\}$ in what follows.
our goal : estimate p when

- small sample situation $n \not>> k$
- domain specific information is available

counts,empirical distribution, type

$x = (x_1, \ldots, x_n)$ an iid sample of size n of a pmf p.
build the *counts* $(n_1, \ldots, n_k)$,
$n_j = \#\{x_i, 1 \leq i \leq n,$ such that $x_i = j\}$
visualize it .. *histogram*



give it names and notations :
$\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ is the *empirical distribution* or *type of* x.

Language modeling

the	56837	of	27155	in	20080
$< \backslash s >$	47108	to	26274	and	19579
N	36068	a	23857	's	11058
... california-backed	... 1	... logs	... 1	... lengthens	... 1

words counts obtained from the UPENN repository database. Three first lines are the largest counts observed. $< \backslash s >$ stands for "end of sentence" and N stands for "numerical expression". This set contains 1,021,203 words. The number of words seen at least once is 37,000.

Likelihood, Kulback-Liebler divergence and Shannon entropy

$x = (x_1, \ldots, x_n)$ a sample of size n of a pmf $p = (p_1, \ldots, p_k)$, $x_i \in \{1, \ldots, k\}$. The type of x is $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$.
Let's compute the log - base 2 - likelihood of the data x

$$
\begin{aligned}
\log p(x) &= \log \prod_{i=1}^{n} p_{x_i} = \log \prod_{j=1}^{k} p_j^{n_j} \\
&= n \sum_{j=1}^{k} \frac{n_j}{n} \log \frac{p_j}{\frac{n_j}{n}} \frac{n_j}{n} \\
&= -n\left(\sum_{j=1}^{k} \frac{n_j}{n} \log \frac{\frac{n_j}{n}}{p_j} - \sum_{j=1}^{k} \frac{n_j}{n} \log \frac{n_j}{n}\right) \\
&= -n(D(\hat{p}, p) + H(\hat{p}))
\end{aligned}
$$

Why (not) using the Maximum Likelihood Estimator

- $\hat{p} \to p$ a.s. fundamental theorem of statistics.
- when k is large, there might be many values for which $p_i << \frac{1}{n}$. In this case, with high probability, $\hat{p}_i = 0$. Leads to a bias in entropy : as soon as $p \neq \delta_j$,

$$E_p(H(\hat{p})) < H(p)$$

- what about prior knowledge ?

Alternatives to the Maximum Likelihood Estimator

- generic
  - Bayesian estimates with Dirichlet prior
  - Minimax estimates

  add-β rules

$$\tilde{p}_i = \frac{n_i + \beta}{n + \beta k} = (1 - \lambda)\frac{n_i}{n} + \lambda\frac{1}{k}, \text{ with } \lambda = \frac{\beta k}{n + \beta k}$$

- specific
  - Good-Turing in langage modeling
  - Bayesian estimates with specific prior

The "Maximum Likelihood Set"

$x = (x_1, \ldots, x_n)$ a sample of size n. The type of x is $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$.
The Maximum Likelihood Set (MLS) is 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*.

$$\mathcal{M}(\hat{p}) = \{p = (p_1, \ldots, p_k), \forall (n'_1, \ldots, n'_k), \sum_{j=1}^{k} n'_j = n,$$
$$Prob_p(n_1, \ldots, n_k) >= Prob_p(n'_1, \ldots, n'_k)\} \Leftrightarrow$$
$$\frac{n!}{n_1! \ldots n_k!} \prod_{l=1}^{k} p_l^{n_l} \geq \frac{n!}{n'_1! \ldots n'_k!} \prod_{l=1}^{k} p_l^{n'_l}\}$$

The "Maximum Likelihood Set" (continued)

$\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$

$$\frac{n!}{n_1! \ldots n_k!} \doteq 2^{nH(\hat{p})} \text{ where } a_n \doteq b_n \iff \frac{1}{n} \log(\frac{a_n}{b_n}) \to 0$$

$$\text{and recall that } \prod_{l=1}^{k} p_l^{n_l} = 2^{-n(D(\hat{p},p)+H(\hat{p}))}$$

$$\text{so that } Prob_p(n_1, \ldots, n_k) \doteq 2^{-nD(\hat{p},p)}$$

$$\text{hence } \mathcal{M}(\hat{p}) \approx \{p; \forall \hat{p}', D(\hat{p}, p) \leq D(\hat{p}', p)\}$$

Characterization of the "Maximum Likelihood Set"

Let $(n_1, \ldots, n_k)$ be the counts. Let's define a neighborhood relationship in the set of types with denominator n. The neighbors of $(n_1, \ldots, n_k)$ are the types obtained by moving a single sample from one value to another one. 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 true which leads to the following

Proposition

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

$$(n_i + 1)p_j \geq n_j p_i \qquad \forall\, 1 \leq i \neq j \leq k,$$

Idea of the proof

Choose $(n_1, \dots, n_i + 1, \dots, n_j - 1, \dots, n_k)$, a neighbor of $(n_1, \dots, n_k)$, then recall that

$$Prob_p(n_1, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} \prod_{j=1}^{k} p_j^{n_j}$$

$$\begin{aligned} Prob_p(n_1, \dots, n_k) &\geq Prob_p(n_1, \dots, n_i + 1, \dots, n_j - 1, \dots, n_k) \\ \Leftrightarrow (n_i + 1) p_j &\geq n_j p_i (1) \end{aligned}$$

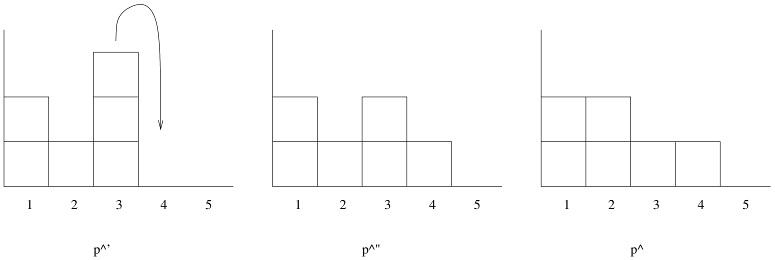
hence, with $\hat{p} = (\frac{n_1}{n}, \dots, \frac{n_k}{n})$,

$$\mathcal{M}(\hat{p}) \subset \{p; (n_i + 1) p_j \geq n_j p_i, i \neq j\}$$

Proof (continue)

conversly, suppose that p verifies $(n_i + 1)p_j \geq n_j p_i, \forall i \neq j$, then for any $\hat{p}'$, choose a neighbor $\hat{p}''$ in the direction of $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ then one can check that

$$Prob_p(\hat{p}'') \geq Prob_p(\hat{p}')$$

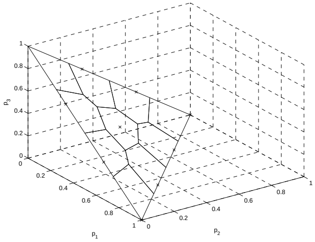


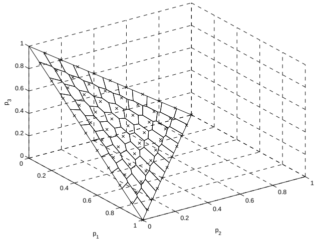
Motivating Examples

For $k = 2$, the MLS is

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

for $k = 3$, **Left :** $n = 3$, **Right :** $n = 10$





More Motivating Examples

$$\mathcal{M}(\frac{n_1}{n}, \ldots, \frac{n_k}{n}) = \{p; (n_i+1)p_j \geq n_j p_i, \qquad \forall 1 \leq \ i \neq j \ \leq k\}$$

if $n_1 = n$ and $n_i = 0, \forall i \neq 1$ then

$$\mathcal{M}(\frac{n_1}{n}, \ldots, \frac{n_k}{n}) = \{p; p_1 \geq np_i, \forall i \neq 1\}$$

- If $n = 1$, then the MLS always contains the Uniform pmf.
- The element of the MLS with Maximum Entropy is

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

If $k > n$, $p_1^* \leq 0.5$ to be contrasted with the estimation $\hat{p}_1 = 1$ given by the type.

Properties of the Maximum Likelihood Set

$$\mathcal{M}(\frac{n_1}{n}, \ldots, \frac{n_k}{n}) = \{p; n_j p_i \ \leq \ (n_i + 1)p_j, \qquad \forall 1 \leq \ i \neq j \ \leq k\}$$

Proposition

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 Maximum Likelihood Set $\mathcal{M}(\hat{p})$ verify

$$\forall 1 \leq j \leq k, n_j > 0 \Rightarrow p_j > 0 \tag{1}$$

$$\forall 1 \leq i, j \leq k, n_i < n_j \Rightarrow p_i \leq p_j \tag{2}$$

$$\forall 1 \leq i \leq k, \hat{p}_i \frac{n}{n+k} \leq p_i \leq \hat{p}_i + \frac{1}{n} \tag{3}$$

More Properties of the Maximum Likelihood Set

$$\mathcal{M}(\frac{n_1}{n}, \ldots, \frac{n_k}{n}) = \{p; n_j p_i \ \le \ (n_i + 1)p_j, \qquad \forall\, 1 \le \ i \ne j \ \le k\}$$

Proposition

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 Maximum Likelihood Set $\mathcal{M}(\hat{p})$ verify

$$\|p - \hat{p}\|_1 = \sum_{i=1}^{k} |p_i - \hat{p}_i| \le \frac{2(k-1)}{n}$$

$\hat{p} \in \mathcal{M}(\hat{p})$, but no other type with denominator n is an element of $\mathcal{M}(\hat{p})$

If $x_i, \ldots, x_n$ are independent samples with common pmf $q \in \mathcal{P}^k$ and type $\hat{p}$, then

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

Selecting an Element from the Maximum Likelihood Set

Proposition

Let $\hat{p} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$ *be a type and* $\mathcal{M}(\hat{p})$ *the MLS associated. 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*

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

Selecting an Element from the Maximum Likelihood Set (more)

Proposition

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

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

has the "monotonicity" property:

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

Hence,

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

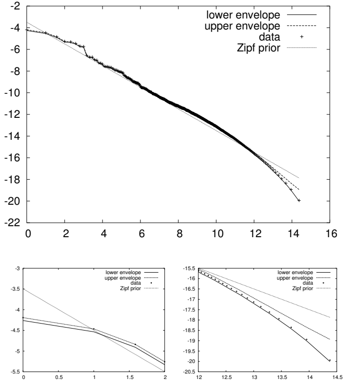
Back to Language Modeling

the	56837	of	27155	in	20080
$< \backslash s >$	47108	to	26274	and	19579
N	36068	a	23857	's	11058
... california-backed	... 1	... logs	... 1	... lengthens	... 1

words counts obtained from the UPENN repository database. Three first lines are the largest counts observed. $< \backslash s >$ stands for "end of sentence" and N stands for "numerical expression". This set contains 1,021,203 words. The number of words seen at least once is 37,000.

Rank ordered data

Zipf Law $(\log i, \log \frac{n_{\sigma(i)}}{n})$ is a straight line with slope -1 provides a reference pmf.



Measuring Performances

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

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

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.

Results

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

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

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. "avg." stands for average and "std. dev." stands for standard deviations.

Numerical aspects

Effective-$k \approx 600$ for $n = 10^6$.

Minimize a convex function over a convex polyhedra in dimension 600.

Tune-up of the convex optimization package CFSQP developped by Lawrence, Zhou and Tits (1997). A C code for solving (large scale) constrained nonlinear (minimax) optimization problems, generating iterates satisfaying all inequality constraints.

Ongoing work

- Estimate $p(w_{(m+1)}|w_m)$ using $p(w)$ as reference pmf. Iterate.
- Estimate pmf from functions of the type.