Estimating Conditional Densities from Sparse Data for Statistical Language Modeling

Damianos Karakos and Sanjeev Khudanpur
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, USA
{damianos,khudanpur}@jhu.edu

***Abstract*— The Maximum Likelihood Set (MLS) was recently introduced in [1] as an effective, parameter-free technique for estimating a probability mass function (pmf) from sparse data. The MLS contains all pmfs that assign merely a higher likelihood to the observed counts than to any other set of counts, for the same sample size. In this paper, the MLS is extended to the case of conditional density estimation. First, it is shown that, when the criterion for selecting a pmf from the MLS is the KL-divergence, the selected conditional pmf naturally has a back-off form, except for a *ceiling* on the probability of high frequency unigrams that are not seen in *particular* contexts. Second, the pmf has a sparse parameterization, leading to efficient algorithms for KL-divergence minimization. Finally, a novel *fattening* of the MLS, called the High Likelihood Set (HLS) is introduced. It contains the MLS, and some neighboring pmfs. Experimental results from bigram and trigram estimation indicate that pmfs selected from the HLS are competitive with state-of-the-art estimates.**

I. The Density Estimation Problem

The problem of probability density estimation may be formulated as follows: a sequence $\mathcal{W} = \{w_1, \ldots, w_N\}$ of independent samples, drawn according to an unknown probability mass function (pmf) P_{True} is observed, and the goal is to estimate P_{True}. It is assumed that the samples w_j belong to a discrete and finite set $\mathcal{V} = \{1, \ldots, V\}$. To facilitate a more concrete exposition, think of $\mathcal{V}$ as the vocabulary of a statistical language model (LM), and $\mathcal{W}$ as the training corpus. This estimation problem is, of course, a recurring problem not only in natural language processing (NLP) but indeed in all of statistics. A popular estimate of P_{True} is the maximum likelihood estimate,

$$\hat{P}_{\text{ML}}(v) \;=\; \frac{c_v}{N} \;=\; \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}(w_t = v), \quad \forall\, v \in \mathcal{V}, \qquad (1)$$

where $\mathbf{1}(A)$ is the indicator function of an event A, and c_v is the observed count of a word v in the corpus $\mathcal{W}$. $\hat{P}_{\text{ML}}$ is usually adequate when $N \gg V$ and $N \gg \max_v \left\{ \frac{1}{P_{\text{True}}(v)} \right\}$. But small samples usually result in $\hat{P}_{\text{ML}}(v)$ being unacceptably small for some $v \in \mathcal{V}$. Zero probabilities, in particular, lead to severe performance degradation when the estimate $\hat{P}_{\text{ML}}$ is subsequently employed, e.g., in automatic speech recognition, parsing, machine translation, and other NLP applications in which statistical models are used.

One standard solution to data sparseness is Bayesian estimation, in which P_{True} itself is viewed as a continuous-valued random variable on the unit simplex $\mathbb{P} \subset \mathbb{R}^V$, with a prior probability density $\pi(P)$. Under a quadratic loss function, the Bayesian estimate $\hat{P}_{\text{Bayes}}$ of P_{True} is the mean of the posterior probability $\pi(P|\mathcal{W})$ given the sample. Formulae such as

$$\hat{P}_{\text{Bayes}} \quad \equiv \quad \hat{P}_{\text{Bayes}}(v) = \frac{c_v + \alpha}{N + \alpha V}, \quad \forall\, v \in \mathcal{V}, \qquad (2)$$

for the Dirichlet prior have been used, where the *hyper-parameter* α is chosen based on some prior knowledge about P_{True}; $\alpha = 1$, $\alpha = \frac{1}{2}$ and $\alpha = \frac{1}{V}$ are often used [2].

Other well-known estimates in language modeling, and elsewhere, include Good-Turing discounting, Witten-Bell discounting, Kneser-Ney discounting, etc [3]. In these methods, the discounted estimate of P_{True} is computed using a back-off formula

$$\hat{P}_{\text{Discount}}(v) \;=\; \begin{cases} \frac{c_v - \delta(c_v)}{N} & \text{if } c_v > 0, \\ \frac{\delta(0)}{N} & \text{if } c_v = 0, \end{cases} \qquad (3)$$

whose discount parameters are estimated via some heuristics, or from a held out portion of $\mathcal{W}$, to provide nonzero probability to unseen words. Readers interested in details of these methods are referred to the survey paper by Chen and Goodman [4]. All these methods make some ad hoc assumptions about the unknown P_{True} that are not substantiated in $\mathcal{W}$, and some further aggravate data sparseness by dividing $\mathcal{W}$ into a training and a held out set. Even in the theoretically pleasing Bayesian case, in which P_{True} itself is viewed as a continuous-valued random variable on the unit simplex $\mathbb{P} \subset \mathbb{R}^V$, the purpose of the prior is often only to ensure that the MAP estimate $\hat{P}_{\text{Bayes}}$ is positive everywhere, and different priors may therefore lead to different estimates.

Maximum entropy estimation is another standard solution to data sparseness. Instead of estimating $\hat{P}(v)$ for each $v \in \mathcal{V}$ according to (1), the maximum entropy method first estimates $\hat{P}(v \in A_j) = \hat{a}_j$ for *select sets* $A_j \subset \mathcal{V}$, $j = 1, \ldots, J$, for which we have sufficient evidence in $\mathcal{W}$. Fixing the probability of some subsets of $\mathcal{V}$ in this manner typically under-specifies the pmf of interest, leading to a set $\mathcal{P}$ of *admissible* pmfs $\mathcal{P} = \{P \in \mathbb{P} \,:\, P(A_j) = \hat{a}_j,\ j = 1, \ldots, J\}$. The admissible pmf with the highest Shannon entropy is then chosen as the estimate of P_{True}: $\hat{P}_{\text{MaxEnt}} = \arg\max_{P \in \mathcal{P}} H(P)$.

In the following section, we briefly describe a technique we have recently developed using the notion of a *maximum likelihood set* [1]. In Section III, we describe how this technique is applied to the estimation of a statistical language model and

how the estimation can be made computationally efficient. In Section IV we explain the connection between the proposed estimate and the notion of back-off in language modeling; we further show that the back-off formula is improved by introducing a *ceiling* on the "backed-off" probabilities of frequent unigrams. We present empirical results for bigram and trigram estimation in Section V, and conclude in Section VI.

II. The Maximum Likelihood Set

Let $\hat{\mathbb{P}}^{(N)} \subset \mathbb{P}$ denote the set of all possible empirical distributions, or *types*, for a sample of size N [5]. A type, which is the same as the maximum likelihood estimate of (1), is fully specified by the counts $(c_1, \ldots, c_V)$, and for N independent samples drawn according to a common pmf $P_{\text{True}} \in \mathbb{P}$, the probability of observing a type $\hat{P}$ is

$$P_{\text{True}}(\hat{P}) \;=\; P_{\text{True}}\,(c_1, \ldots, c_V) \;=\; \frac{N!}{c_1! \ldots c_V!} \prod_v P_{\text{True}}(v)^{c_v} \tag{4}$$

For a given type $\hat{P} \in \hat{\mathbb{P}}^{(N)}$, we define the *maximum likelihood set* (MLS) as

$$\mathcal{M}(\hat{P}) = \left\{ P \in \mathbb{P} \mid P(\hat{P}) \geq P(\hat{P}'), \forall\, \hat{P}' \in \hat{\mathbb{P}}^{(N)} \right\}. \tag{5}$$

In words, $\mathcal{M}(\hat{P})$ is the set of all pmfs under which the observed type $\hat{P}$ is no less likely than any other type in $\hat{\mathbb{P}}^{(N)}$. An equivalent characterization of the MLS is provided in [1]:

$$\mathcal{M}(\hat{P}) = \{P \in \mathbb{P} : (c_u + 1)P(v) \geq c_v P(u),\, \forall\, u, v \in \mathcal{V}\}. \tag{6}$$

The MLS has several desirable properties as described in [1].

1. $\mathcal{M}(\hat{P})$ is a closed, convex, $V \times (V-1)$-sided polyhedron.
2. The observed type $\hat{P}$ is the only type in $\mathcal{M}(\hat{P})$.
3. Diameter: $\|P - \hat{P}\|_1 \leq 2(V-1)N^{-1}\ \forall\, P \in \mathcal{M}(\hat{P})$.
4. Strong consistency: as $N \to \infty$, all pmfs inside the MLS converge to P_{True} almost surely.
5. If a word v has $c_v > 0$, then $P(v) > 0\ \forall\, P \in \mathcal{M}(\hat{P})$.
6. $\mathcal{M}(\hat{P})$ contains pmfs P such that $P(v) > 0\ \forall\, v \in \mathcal{V}$.
7. Faithfulness to evidence: if $c_u < c_v$ then $P(u) \leq P(v)\ \forall\, P \in \mathcal{M}(\hat{P})$.

$\mathcal{M}(\hat{P})$ is proposed as an *admissible* set of pmfs, from which a particular pmf may be chosen as an estimate of the underlying pmf P_{True} using secondary criteria. In particular, if a *reference pmf* Q is available, i.e. an estimate one would find acceptable when $N = 0$, then one way to choose an element of $\mathcal{M}(\hat{P})$ is to minimize the Kullback-Leibler divergence (KL-divergence):

$$\hat{P}_{\text{MLS}} = \underset{P \in \mathcal{M}(\hat{P}_{ML})}{\arg\min}\ D(P\|Q) = \underset{P \in \mathcal{M}(\hat{P}_{ML})}{\arg\min} \sum_{v \in \mathcal{V}} P(v) \log \frac{P(v)}{Q(v)} \tag{7}$$

Attainment of the minimum, and the uniqueness of the minimizer, is guaranteed by $\mathcal{M}(\hat{P})$ being closed and convex and by the convexity of the KL-divergence [6, Theorem 2.1]. Note that if Q is the uniform pmf on $\mathcal{V}$, then the criterion for selecting $\hat{P}_{\text{MLS}}$ through KL-divergence minimization is simply maximum entropy. Moreover, $c_u = c_v$ and $Q(u) = Q(v)$ guarantees $\hat{P}_{\text{MLS}}(u) = \hat{P}_{\text{MLS}}(v)$. This results in great simplifications in the numerical computation of $\hat{P}_{\text{MLS}}$ in our experiments.

III. Statistical Language Modeling: Conditional Estimation and Complexity Issues

Statistical language models are a key component in NLP 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 $P(w_n|w_{n-1}, \ldots, w_1)$, 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 other NLP applications, however, prediction of a word given its *history* plays an important role; language models of order at least $n = 2$ (bigram) and usually $n = 3$ (trigram) improve performance dramatically over the unigram, and are thus preferred. On the other hand, estimation of bigram or trigram models is based on much fewer data samples per conditioning history, thus making the application of smoothing techniques even more necessary.

Conditional density estimation can be performed efficiently using MLS techniques. Each conditioning event (e.g., a single word, or a pair of words encountered in some training data for the bigram and trigram cases, respectively) gives rise to a different MLS: different words are seen following each history. Hence, for each history h, an estimate of the conditional pmf is selected from the corresponding MLS $\mathcal{M}(h)$. To do that, the following issues have to be addressed:

- A reference distribution has to be selected for each conditioning history. In our experiments, we used either a back-off distribution (e.g., unigram for a bigram model, or bigram for a trigram model), or some other conditional pmf estimate of the same order. (Of course, although one is free to pick any arbitrary reference distribution, we feel that the above choices are well-justified as they "encode information" about the particular training corpus at hand.)
- The number of distinct histories can be very high, ranging from a few tens of thousands (single words) to a few hundreds of thousands (pairs of words). Therefore, since a different convex optimization problem (for finding an estimate inside the MLS) needs to be solved for each one of these histories, it is crucial that the optimizations are performed in a computationally efficient way.

To address the second issue above, we identified a number of ways to achieve significant computational savings, and we list them in the following subsection.

A. Computational Complexity Reduction Techniques

Each convex optimization problem is solved using a quadratic program, whose computational complexity is affected by two quantities: (i) The dimensionality of the prob-

lem, which is *at most*[1] equal to the number V of parameters (each parameter corresponds to a word, whose probability needs to be estimated). (ii) The number of constraints, which is equal to $V\times(V-1)$, the number of hyperplanes which bound the MLS. Of course, there is also the usual sum-to-one and non-negativity requirements of probability. These two quantities can be reduced dramatically by considering the following:

- Two distinct words that follow a history, but have the same counts and reference Q probability, may be collapsed when estimating $\hat{P}_{\text{MLS}}$, as mentioned at the end of Section II. For each history, the *effective* size of the alphabet therefore is much smaller than the alphabet size V. Furthermore, as we will see in the next section, the effective alphabet of only words with positive counts needs to be considered; unseen words need not be part of the parameterization in the convex optimization, and their probability is completely determined (collectively) by the probabilities of the seen words.
- Unseen words do not impose upper bounds (6) on the probability of seen (or other unseen) words: if a word w has not been seen following a history h, then for any other word w' we have the inequality

$$c_{(hw)}P(w'|h) \leq (c_{(hw')}+1)P(w|h), \quad P(\cdot|h)\in\mathcal{M}(h),$$

which holds trivially, since $c_{(hw)}$, the number of times that w follows h, is equal to zero. Also, the number of seen words following a history is usually much smaller than the number of unseen words. Hence, if c_{seen} distinct words have been seen following a history, the number of constraints in the definition of the MLS is approximately $c_{\text{seen}}^2+c_{\text{seen}}\times V$, and the remaining $(V-c_{\text{seen}})^2$ constraints imposed by unseen words on other unseen words are redundant. Now, c_{seen} is, on average, about 30 (e.g., for bigram language model estimation) of the Wall Street Journal corpus, and reduces further to just 4 once distinct seen words with the same count and reference probability are collapsed. The number of constraints therefore is about $4^2+4\times 1500=6016$, a tremendous reduction over the 2.5 *billion* constraints suggested by $V\times(V-1)$.

By taking advantage of the above, the minimization of the KL-divergence (more precisely, a *modified* KL-divergence which is defined in Section IV) is done very efficiently, and we are able to estimate bigram and trigram pmfs with modest computing.

IV. Connections with the Back-off Formula

There are many ways of selecting a distribution from the MLS. In the case where the objective is the minimization of the Kullback-Leibler divergence from a reference distribution Q (cf. Section II), we have that

$$D(P\|Q)=\sum_{w:c_{(hw)}>0}P(w|h)\log\frac{P(w|h)}{Q(w)} + \sum_{w:c_{(hw)}=0}P(w|h)\log\frac{P(w|h)}{Q(w)}, \quad (8)$$

[1]We will see in Section IV that it is much lower.

i.e., the calculation is split into the seen $\mathcal{S}$ and the unseen $\mathcal{U}$ portions of the vocabulary. As we mentioned in the previous section, unseen words do not impose upper-bound constraints; hence, for any allocation of probabilities over $\mathcal{S}$, the probabilities over $\mathcal{U}$ have to be such that the following two conditions are satisfied:

1. $P(\mathcal{U})=1-P(\mathcal{S})$ (sum-to-one constraint).
2. $\forall\, w\in\mathcal{U},\ \ P(w|h)\leq\min_{w'\in\mathcal{S}}\frac{P(w'|h)}{c_{(hw')}}\triangleq\gamma(P_{\mathcal{S}})$ (by (6)),

where $P_{\mathcal{S}}$ corresponds to the probability values over only the seen vocabulary. Now, it can be proved that the estimate $P(w|h)$ over $\mathcal{U}$ minimizes the KL-divergence if it is proportional to Q, but with a *ceiling* $\gamma(P_{\mathcal{S}})$ so that condition 2 above is satisfied, as is illustrated in Figure 1. Moreover, the minimization of $D(P\|Q)$ inside the MLS can be done by minimizing a "semi-optimized" KL-divergence variation: i.e., by searching over $\mathcal{S}$ for the set of probability values (whose sum should not exceed unity) that minimizes the functional

$$\tilde{D}(P\|Q)=\sum_{w:c_{(hw)}>0}P(w|h)\log\frac{P(w|h)}{Q(w)} + \sum_{w\in\mathcal{C}}\gamma(P_{\mathcal{S}})\log\frac{\gamma(P_{\mathcal{S}})}{Q(w)}+\sum_{w\in\mathcal{N}}\beta(P_{\mathcal{S}})Q(w)\log(\beta(P_{\mathcal{S}})),$$

where $\beta(P_{\mathcal{S}})$ can be computed in a unique way to satisfy condition 1 above, and $\mathcal{C},\mathcal{N}$ are the "capped" and "non-capped" subsets of $\mathcal{U}$; i.e., sets over which the words have probabilities equal to, or strictly less than, $\gamma(P_{\mathcal{S}})$, respectively.

Hence, the minimizing pmf has the back-off form

$$\hat{P}_{\text{MLS}}(w|h)=\begin{cases}P^*(w|h) & \text{if } c_{(hw)}>0\\ \min\{\gamma(h),\ \ \beta(h)\times Q(w)\} & \text{if } c_{(hw)}=0\end{cases}$$

where $P^*(w|h)$ is the value of the minimizing pmf for $w\in\mathcal{S}$ (which, contrary to the case $w\in\mathcal{U}$, is not given by a closed-form expression), and $\gamma(h)\triangleq\gamma(P^*_{\mathcal{S}})$. The proof of this "capped" back-off formula appears in the full version of this paper; it uses Lagrange multipliers and it is based on the following facts:

- All distributions P inside the MLS give a probability to any unseen word w which cannot exceed $\gamma(P_{\mathcal{S}})$. This is a consequence of the MLS definition; it also implies that no unseen word should be assigned a probability higher than the probability of any of the seen words.
- Any unseen words which are *unconstrained* (i.e., they are assigned a probability strictly less than $\gamma(P_{\mathcal{S}})$, should be proportional to Q with the same constant of proportionality β; this is derived from the fact that the KL-divergence $D(P\|Q)$ is minimized (with respect to P, under a sum constraint) when the two measures P,Q are proportional to each other.

Furthermore, an important observation is in order: the *ceiling* $\gamma(h)$ limits the values of *frequent* unigrams that are nevertheless *unseen* in a context h. This is a very plausible and natural consequence of the MLS: no matter how frequent a word is *in general*, the fact that it is unseen under a *particular* context should be a good indication that it is *infrequent* in that context.

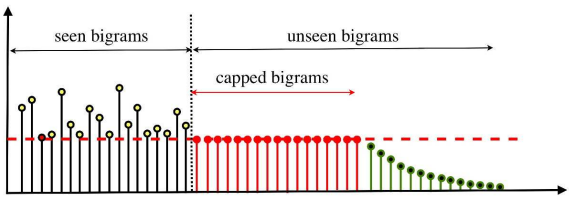

Fig. 1. The MLS estimate places a ceiling on the backed-off probability of frequent unigrams when they are unseen in a history, but otherwise behaves like a back-off estimate.

Hence, its probability should not be just a scaled version of its overall frequency, which is what happens with the back-off formulas of most popular language models. Empirically, most language models are inconsistent in that respect: they give much higher probability to *function* words, even after back-off, than they do to words that actually appear in that context, but are otherwise infrequent. Figure 2 shows the percentage of inconsistent histories of a modified Kneser-Ney model, as a function of their count. Good-Turing and Witten-Bell models give rise to similar plots.

V. Empirical Results for Language Modeling

We have conducted experiments on English text from the Wall Street Journal corpus. A particular subset of this corpus, called the UPenn Treebank corpus, widely used by many researchers in language modeling, has a standard division into Sections, named 00 through 24. We use Sections 00-22, containing about 900K words, as our training corpus, and Sections 23-24, containing 100K words comprise our test corpus.

We make a list of all seen words from Sections 00-22 and augment this vocabulary with a set of unspecified "unseen" words. The decision on how many unseen words to include is presently ad-hoc — based on a leave-one-out estimate, the number of unseen words is set exactly equal to the number of words seen only once in the corpus. To measure the efficacy of a pmf estimate $\hat{P}$, we compute the average codeword length (in bits) that the estimate achieves on the type of the test set, that is,

$$\frac{1}{N_{\text{Test}}} \log(\hat{P}(w_1, \ldots, w_{N_{\text{Test}}})^{-1}), \tag{9}$$

where N_{Test} is the size of the test set $\{w_1, \ldots, w_{N_{\text{Test}}}\}$.

A. Empirical Results for Conditional Density Estimation

We use the counts from Sections 00-22 to compute V. The leave-one-out procedure described above yields V=52,743. Also, the number of distinct bigrams is approximately 350,000, and the number of distinct trigrams is 670,000. For each history in the training set (there are 37,000 distinct unigram histories and 350,000 distinct bigram histories) we collect counts of all words following that history, and we find the pmf inside the MLS (as defined by these counts) that minimizes the KL-divergence from a per-history reference distribution.

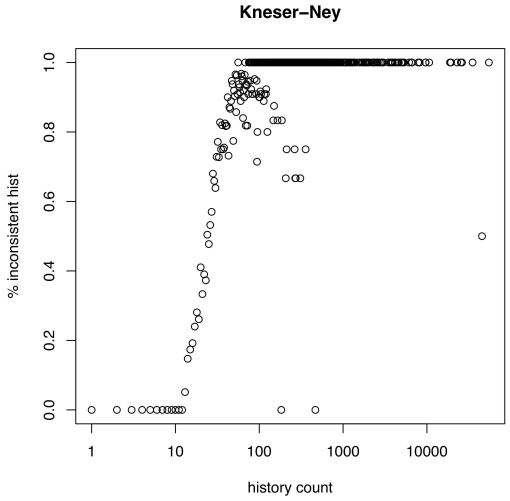

Fig. 2. The fraction of histories (y-axis) whose conditional modified Kneser-Ney model is inconsistent vs. the number of times each of these histories has appeared in the corpus (x-axis).

For each history, the reference distribution used is the corresponding conditional Good-Turing, Witten-Bell or modified Kneser-Ney distribution.

The average lengths of the encoded test words are shown in Table I (for bigram language modeling), and in Table II (for trigram language modeling). These codeword lengths are further broken down into the codeword lengths of test-set words that were seen/not seen, under each conditioning event, in the training data. As can be observed, the bigram MLS-derived estimate has almost identical performance with the Good-Turing and Witten-Bell estimates; and it is slightly worse than the modified Kneser-Ney estimate. In trigram language modeling, the MLS estimate has a slightly higher average codeword length compared to the Witten-Bell and Kneser-Ney models; in all cases, however, the MLS estimate has a lower codeword length on the *unseen* portion of the vocabulary.

Speech recognition WER (word-error-rate) results are shown in Table III for trigram language models. We tested the models on the ARPA93 HUB1 portion of the WSJ; it contains 213 utterances read from the Wall Street Journal, a total of 3446 words.

B. The High Likelihood Set

It may be observed that the admissibility criterion of the MLS, i.e., that the observed counts be more likely than any other set of counts possible for the given sample size, may be too restrictive. In particular, the diameter of the MLS is of the order $O(1/N)$, while the blowing-up lemma of [7] suggests that the empirical distribution lies within $O(1/\sqrt{N})$ of the true distribution with high probability. Hence, enlarging (*fattening*) the MLS by a factor of $\sqrt{N}$, so that it includes the true distribution, seems a plausible approach. This leads to the definition of the High Likelihood Set (HLS), which

Bigram ref. pmf Q		Good-Turing		Witten-Bell		Kneser-Ney	
Avg. length of Q		8.69		8.48		8.36	
seen	unseen	6.30	15.46	6.33	14.55	6.27	14.26
Avg. length of $\hat{P}_{\text{MLS}}$		8.68		8.44		8.39	
seen	unseen	6.30	15.39	6.28	14.55	6.33	14.19
Avg. length of $\hat{P}_{\text{HLS}}$		8.69		8.48		8.36	
seen	unseen	6.29	15.47	6.33	14.56	6.27	14.26

TABLE I

BIGRAM AVERAGE CODEWORD LENGTH RESULTS (IN BITS) FOR VARIOUS REFERENCE DISTRIBUTIONS. THE 2ND AND 3RD ROWS SHOW THE CODEWORD LENGTHS OF THE REFERENCE DISTRIBUTIONS THEMSELVES (GOOD-TURING, WITTEN-BELL OR KNESER-NEY). THE 4TH AND 5TH ROWS CORRESPOND TO THE CODEWORD LENGTHS OF THE MLS ESTIMATES, AND THE 6TH AND 7TH ROWS CORRESPOND TO THE CODEWORD LENGTHS OF THE HLS ESTIMATES (WITH $\alpha(h) = 1/\sqrt{N(h)}$), WHERE THE REFERENCE DISTRIBUTION USED IS INDICATED BY THE COLUMN HEADING.

Trigram ref. pmf Q		Good-Turing		Witten-Bell		Kneser-Ney	
Avg. length of Q		8.47		8.22		8.08	
seen	unseen	4.19	12.02	4.25	11.49	4.12	11.37
Avg. length of $\hat{P}_{\text{MLS}}$		8.46		8.24		8.13	
seen	unseen	4.31	11.88	4.38	11.44	4.36	11.28
Avg. length of $\hat{P}_{\text{HLS}}$		8.41		8.19		8.07	
seen	unseen	4.07	11.98	4.18	11.49	4.11	11.36

TABLE II

TRIGRAM AVERAGE CODEWORD LENGTH RESULTS (IN BITS) FOR THE REFERENCE DISTRIBUTION, AS WELL AS THE MLS AND HLS ESTIMATES.

contains the MLS, as well as many other smooth, neighboring distributions. In general terms, we define the HLS as the set of all pmfs which assign to the observed counts a likelihood which is at least a fraction α of the likelihood they assign to any other counts (for the same sample size), where $0 < \alpha < 1$.

In our experiments, we chose a history-specific $\alpha(h) = 1/\sqrt{N(h)}$, where $N(h)$ is the total number of words following history h; the codeword length results are shown in Tables I and II for bigram and trigram estimation, respectively. The bigram HLS estimate has almost identical performance to the performances of the other models, while the trigram HLS estimate has better performance than the other models. WER results are shown in Table III; it is interesting that the HLS WERs are uniformly slightly lower than the WERs of the other models (although not significantly).

Trigram ref. pmf Q	Good-Turing	Witten-Bell	Kneser-Ney
WER of Q	15.7%	16.0%	15.8%
WER of $\hat{P}_{\text{MLS}}$	15.8%	16.0%	15.8%
WER of $\hat{P}_{\text{HLS}}$	15.7%	15.9%	15.7%

TABLE III

WORD-ERROR-RATE RESULTS FOR VARIOUS TRIGRAM MODELS.

VI. CONCLUDING REMARKS

We have outlined a new way of viewing the problem of conditional pmf estimation, particularly from small samples, commonly encountered in language modeling and in other applications. The view, based on the notion of the maximum likelihood set defined in (5), opens many avenues of investigations not only in language modeling but in other areas of statistical estimation.

We show that, when the optimizing criterion for selecting a pmf from the MLS is the KL-divergence, the minimizing distribution improves upon the standard back-off form by placing a *ceiling* on the backed-off probabilities of very frequent unigrams that are not seen in a history. Moreover, performing the optimization for each history may be carried out efficiently due to the sparse parameterization of the solution: the complexity of the optimization is essentially determined only by the number of distinct *seen* words; this is a result which conforms with intuition. Finally, the experiments presented here demonstrate that pmfs selected using the K-L divergence criterion from the MLS (or HLS) have many desirable properties and better than state-of-the-art performance, even in an application that has been studied for decades. Moreover, preliminary experiments in a speech recognition task show a small reduction in word error rate over the modified Kneser-Ney model when the HLS estimate is used for lattice rescoring.

We also propose to study other conditional estimation problems in NLP that suffer from sparse data, such as rule probabilities in statistical parsers, where the Kneser-Ney estimate is not necessarily the best. We also propose to use the admissible pmfs in the MLS to determine the range of *target expectations* in maximum entropy models — note that in classical formulations of maximum entropy estimation, the target expectation of a feature is either pegged to be exactly its empirical value, or not constrained at all.

VII. ACKNOWLEDGMENTS

We would like to thank Bruno Jedynak for many stimulating discussions. This research was partially supported by the National Science Foundation via Grant No. ITR-0225656 and IIS-9982329.

REFERENCES

[1] B. Jedynak and S. Khudanpur, "Maximum likelihood set for estimating a probability mass function," *Neural Computation*, vol. 17, no. 7, pp. 1508–1530, July 2005.

[2] D. Wolpert and D. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Physical Review E*, vol. 52, pp. 6841–6854, 1995.

[3] Frederick Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, 1998.

[4] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting of the ACL*, 1996, pp. 310–318.

[5] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, NY, 1981.

[6] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, February 1975.

[7] Imre Csiszár and János Körner, *Information Theory: Coding Theorems For Discrete Memoryless Systems*, Academic Press, Budapest, 1981.