Concentration Bounds for Unigrams Language Model

Evgeny Drukh and Yishay Mansour

School of Computer Science, Tel Aviv University, Tel Aviv, Israel.
E-mail: {drukh,mansour}@post.tau.ac.il

Abstract. We show several PAC-style concentration bounds for learning unigrams language model. One interesting quantity is the probability of all words appearing exactly k times in a sample of size m. A standard estimator for this quantity is the Good-Turing estimator. The existing analysis on its error shows a PAC bound of approximately $O\left(\frac{k}{\sqrt{m}}\right)$. We improve its dependency on k to $O\left(\frac{\sqrt[4]{k}}{\sqrt{m}} + \frac{k}{m}\right)$. We also analyze the empirical frequencies estimator, showing that its PAC error bound is approximately $O\left(\frac{1}{k} + \frac{\sqrt{k}}{m}\right)$. We derive a combined estimator, which has an error of approximately $O\left(m^{-\frac{2}{5}}\right)$, for any k.

A standard measure for the quality of a learning algorithm is its expected per-word log-loss. We show that the leave-one-out method can be used for estimating the log-loss of the unigrams model with a PAC error of approximately $O\left(\frac{1}{\sqrt{m}}\right)$, for any distribution.

We also bound the log-loss a priori, as a function of various parameters of the distribution.

1 Introduction and Overview

Natural language processing (NLP) has developed rapidly over the last decades. It has a wide range of applications, including speech recognition, optical character recognition, text categorization and many more. The theoretical analysis has also advanced significantly, though many fundamental questions remain unanswered. One clear challenge, both practical and theoretical, concerns deriving stochastic models for natural languages.

Consider a simple language model, where the distribution of each word in the text is assumed to be independent. Even for such a simplistic model, fundamental questions relating sample size to the learning accuracy are already challenging. This is mainly due to the fact that the sample size is almost always insufficient, regardless of how large it is.

To demonstrate this phenomena, consider the following example. We would like to estimate the distribution of first names in the university. For that, we are given the names list of a graduate seminar: Alice, Bob, Charlie, Dan, Eve, Frank, two Georges, and two Henries. How can we use this sample to estimate the

distribution of students' first names? An empirical frequency estimator would assign Alice the probability of 0.1, since there is one Alice in the list of 10 names, while George, appearing twice, would get estimation of 0.2. Unfortunately, unseen names, such as Michael, will get an estimation of 0. Clearly, in this simple example the empirical frequencies are unlikely to estimate well the desired distribution.

In general, the empirical frequencies estimate well the probabilities of popular names, but are rather inaccurate for rare names. Is there a sample size, which assures us that all the names (or most of them) will appear enough times to allow accurate probabilities estimation? The distribution of first names can be conjectured to follow the Zipf's law. In such distributions, there will be a significant fraction of rare items, as well as a considerable number of non-appearing items, in any sample of reasonable size. The same holds for the language unigrams model, which tries to estimate the distribution of single words. As it has been observed empirically on many occasions ([2], [5]), there are always many rare words and a considerable number of unseen words, regardless of the sample size. Given this observation, a fundamental issue is to estimate the distribution the best way possible.

1.1 Good-Turing Estimators

An important quantity, given a sample, is the probability mass of unseen words (also called "the missing mass"). Several methods exist for smoothing the probability and assigning probability mass to unseen items. The almost standard method for estimating the missing probability mass is the Good-Turing estimator. It estimates the missing mass as the total number of unique items, divided by the sample size. In the names example above, the Good-Turing missing mass estimator is equal 0.6, meaning that the list of the class names does not reflect the true distribution, to put it mildly. The Good-Turing estimator can be extended for higher orders, that is, estimating the probability of all names appearing exactly k times. Such estimators can also be used for estimating the probability of individual words.

The Good-Turing estimators date to World War II, and were published at 1953 ([10], [11]). They have been extensively used in language modeling applications since then ([2], [3], [4], [15]). However, their theoretical convergence rate in various models has been studied only in the recent years ([17], [18], [19], [20], [22]). For estimation of the probability of all words appearing exactly k times in a sample of size m, [19] shows a PAC bound on Good-Turing estimation error of approximately $O\left(\frac{k}{\sqrt{m}}\right)$.

One of our main results improves the dependency on k of this bound to approximately $O\left(\frac{\sqrt[4]{k}}{\sqrt{m}}+\frac{k}{m}\right)$. We also show that the empirical frequencies have an error of approximately $O\left(\frac{1}{k}+\frac{\sqrt{k}}{m}\right)$, for large values of k. Based on the two estimators, we derive a combined estimator with an error of approximately

$O\left(m^{-\frac{2}{5}}\right)$, for any k. We also derive a lower bound of $\Omega\left(\frac{\sqrt[4]{k}}{\sqrt{m}}\right)$ for an error of any estimator based on an independent sample.

Our results give theoretical justification for using the Good-Turing estimator for small values of k, and the empirical frequencies estimator for large values of k. Though in most applications the Good-Turing estimator is used for very small values of k (e.g. $k \leq 5$, as in [15] or [2]), we show that it is fairly accurate in a much wider range.

1.2 Logarithmic Loss

The Good-Turing estimators are used to approximate the probability mass of all the words with a certain frequency. For many applications, estimating this probability mass is not the main optimization criteria. Instead, a certain distance measure between the true and the estimated distributions needs to be minimized.

The most popular distance measure widely used in NLP applications is the *Kullback-Leibler (KL) divergence*. For $P = \{p_x\}$ and $Q = \{q_x\}$, two distributions over some set X, this measure is defined as $\sum_x p_x \ln \frac{p_x}{q_x}$. An equivalent measure, up to the entropy of P, is the *logarithmic loss* (*log-loss*), which equals $\sum_x p_x \ln \frac{1}{q_x}$.

Many NLP applications use the value of *log-loss* to evaluate the quality of the estimated distribution. However, the *log-loss* cannot be directly calculated, since it depends on the underline distribution P, which is unknown. Therefore, estimating *log-loss* using the sample is important, although the sample cannot be independently used for both estimating the distribution and testing it. The *hold-out* estimation splits the sample into two parts: training and testing. The training part is used for learning the distribution, whereas the testing sample is used for evaluating the average per-word log-loss. The main disadvantage of this method is the fact that it uses only part of the available information for learning, whereas in practice one would like to use all the sample.

A widely used general estimation method is called *leave-one-out*. Basically, it means averaging all the possible estimations, where a single item is chosen for testing, and the rest is used for training. This procedure has an advantage of using the entire sample, in addition it is rather simple and usually can be easily implemented. The existing theoretical analysis of the *leave-one-out* method ([14], [16]) shows general PAC-style concentration bounds for the generalization error. However, these techniques are not applicable in our setting.

We show that the *leave-one-out* estimation error for the *log-loss* is approximately $O\left(\frac{1}{\sqrt{m}}\right)$, for any distribution P. In addition, we show a PAC bound for the *log-loss*, as a function of various parameters of the distribution.

1.3 Model and Semantics

We denote the set of all words as V, and $N = |V|$. Let P be a distribution over V, where p_w is the probability of a word $w \in V$. Given a sample S of size m, drawn i.i.d. using P, we denote the number of appearances of a word w in

S as c_w^S, or simply c_w, when a sample S is clear from the context[1]. We define $S_k = \{w \in V : c_w^S = k\}$, and $n_k = |S_k|$.

For a claim Φ regarding a sample S, we write $\forall^\delta S\ \Phi[S]$ for $P(\Phi[S]) \geq 1 - \delta$. For some PAC bound function $f(\cdot)$, we write $\tilde{O}(f(\cdot))$ for $O\left(f(\cdot)\left(\ln \frac{m}{\delta}\right)^c\right)$, where $c > 0$ is some constant, and δ is the PAC error probability.

Due to lack of space, some of the proofs are omitted. A detailed version can be found at [7].

2 Concentration Inequalities

In this section we state several standard Chernoff-style concentration inequalities. We also show some of their corollaries regarding the maximum-likelihood approximation of p_w by $\hat{p}_w = \frac{c_w}{m}$.

Lemma 1. *(Hoeffding's inequality: [13], [18]) Let $Y = Y_1, \ldots, Y_n$ be a set of n independent random variables, such that $Y_i \in [b_i, b_i + d_i]$. Then, for any $\epsilon > 0$,*

$$P\left(\left|\sum_i Y_i - E\left[\sum_i Y_i\right]\right| > \epsilon\right) \leq 2\ \exp\left(-\frac{2\epsilon^2}{\sum_i d_i^2}\right)$$

This inequality has an extension for various functions of $\{Y_1, \ldots, Y_n\}$, which are not necessarily the sum.

Lemma 2. *(Variant of McDiarmid's inequality: [21], [6]) Let $Y = Y_1, \ldots, Y_n$ be a set of n independent random variables, and $f(Y)$ such that any change of Y_i value changes $f(Y)$ by at most d_i. Let $d = \max_i d_i$. Then,*

$$\forall^\delta Y: \qquad |f(Y) - E[f(Y)]| \leq d\sqrt{\frac{n \ln \frac{2}{\delta}}{2}}$$

Lemma 3. *(Angluin-Valiant bound: [1], [18]) Let $Y = Y_1, \ldots, Y_n$ be a set of n independent random variables, where $Y_i \in [0, B]$. Let $\mu = E\left[\sum_i Y_i\right]$. Then, for any $\epsilon > 0$,*

$$P\left(\left|\sum_i Y_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{(2\mu + \epsilon)B}\right)$$

The next lemma shows an explicit upper bound on the binomial distribution probability[2].

[1] Unless mentioned otherwise, all further sample-dependent definitions depend on the sample S.

[2] Its proof is based on Stirling approximation directly, though local limit theorems could be used. This form of bound is needed for the proof of Theorem 4.

Lemma 4. *Let $X \sim Bin(n,p)$ be a binomial random variable, i.e. a sum of n i.i.d. Bernoulli random variables with $p \in (0,1)$. Let $\mu = E[X] = np$. For $x \in (0,n]$, there exist some $T_x = \exp\left(\frac{1}{12x} + O\left(\frac{1}{x^2}\right)\right)$, such that $\forall k \in \{0,\ldots,n\}$, we have $P(X=k) \leq \frac{1}{\sqrt{2\pi\mu(1-p)}} \frac{T_n}{T_\mu T_{n-\mu}}$. For integral values of μ, the equality is achieved at $k=\mu$. (Note that for $x \geq 1$, we have $T_x = \Theta(1)$.)*

The next lemma (by Hoeffding, [12]) deals with the number of successes in independent trials.

Lemma 5. *([12], Theorem 5) Let $Y_1,\ldots,Y_n \in \{0,1\}$ be a set of independent random variables, with $p_i = E[Y_i]$. Let $X = \sum_i Y_i$, and $p = \frac{1}{n}\sum_i p_i$ be the average trial success probability. For any integers b and c such that $0 \leq b \leq np \leq c \leq n$, we have:*

$$\sum_{k=b}^{c} \binom{n}{k} p^k (1-p)^{n-k} \leq P\left(b \leq X \leq c\right) \leq 1$$

Using the above lemma, the next lemma shows a general concentration bound for a sum of arbitrary real-valued functions of a multinomial distribution components. We show that with a small penalty, any Chernoff-style bound pretending the components being independent is valid[3]. We recall that c_w^S, or equivalently c_w, is the number of appearances of the word w in a sample S of size m.

Lemma 6. *Let $\{c'_w \sim Bin(m,p_w) : w \in V\}$ be independent binomial random variables. Let $\{f_w(x) : w \in V\}$ be a set of real valued functions. Let $F = \sum_w f_w(c_w)$ and $F' = \sum_w f_w(c'_w)$. For any $\epsilon > 0$,*

$$P\left(|F - E\left[F\right]| > \epsilon\right) \leq 3\sqrt{m}\, P\left(|F' - E\left[F'\right]| > \epsilon\right)$$

The following lemmas provide concentration bounds for maximum-likelihood estimation of p_w by $\frac{c_w}{m}$.

Lemma 7. *Let $\delta > 0$, and $\lambda \geq 3$. We have $\forall^\delta S$:*

$$\forall w \in V,\ \ s.t.\ \ mp_w \geq 3\ln\tfrac{2m}{\delta}, \quad |mp_w - c_w| \leq \sqrt{3mp_w \ln\frac{2m}{\delta}}$$

$$\forall w \in V,\ \ s.t.\ \ mp_w > \lambda\ln\tfrac{2m}{\delta}, \quad c_w > \left(1 - \sqrt{\frac{3}{\lambda}}\right) mp_w$$

Lemma 8. *Let $\delta \in (0,1)$, and $m > 1$. Then, $\forall^\delta S$: $\forall w \in V$, such that $mp_w \leq 3\ln\frac{m}{\delta}$, we have $c_w \leq 6\ln\frac{m}{\delta}$.*

[3] The *negative association* analysis ([8]) shows that a sum of negatively associated variables must obey Chernoff-style bounds pretending that the variables are independent. The components of a multinomial distribution are negatively associated. Therefore, any Chernoff-style bound is valid for their sum, as well as for the sum of monotone functions of the components. In some sense, our result extends this notion, since it does not require the functions to be monotone.

3 Hitting Mass Estimation

In this section our goal is to estimate the probability of the set of words appearing exactly k times in the sample, which we call "the hitting mass". We analyze the Good-Turing estimator, the empirical frequencies estimator, and the combined estimator.

Definition 1. *We define the hitting mass and its estimators as:* [4]

$$M_k = \sum_{w \in S_k} p_w \qquad \hat{M}_k = \left(\frac{k}{m}\right) n_k \qquad G_k = \left(\frac{k+1}{m-k}\right) n_{k+1}$$

Definition 3 slightly redefines the hitting mass and its estimators. Lemma 9 shows that this redefinition has a negligible influence. Then, we analyze the estimation errors using the concentration inequalities from Section 2.

The expected error of the Good-Turing estimator is bounded, as in [19]. Lemma 14 bounds the deviation of the error, using the negative association analysis. A tighter bound, based on Lemma 6, is achieved at Theorem 1. Theorem 2 analyzes the error of the empirical frequencies estimator. Theorem 3 refers to the combined estimator. Finally, Theorem 4 shows a weak lower bound for the hitting mass estimation.

Definition 2. *For any $w \in V$ and $i \in \{0, \cdots, m\}$, we define $X_{w,i}$ as a random variable equal 1 if $c_w = i$, and 0 otherwise.*

Definition 3. *Let $\alpha > 0$ and $k > 3\alpha^2$. We define $I_{k,\alpha} = \left[\frac{k-\alpha\sqrt{k}}{m}, \frac{k+1+\alpha\sqrt{k+1}}{m}\right]$, and $V_{k,\alpha} = \{w \in V : p_w \in I_{k,\alpha}\}$. We define:*

$$M_{k,\alpha} = \sum_{w \in S_k \cap V_{k,\alpha}} p_w = \sum_{w \in V_{k,\alpha}} p_w X_{w,k}$$

$$G_{k,\alpha} = \frac{k+1}{m-k} |S_{k+1} \cap V_{k,\alpha}| = \frac{k+1}{m-k} \sum_{w \in V_{k,\alpha}} X_{w,k+1}$$

$$\hat{M}_{k,\alpha} = \frac{k}{m} |S_k \cap V_{k,\alpha}| = \frac{k}{m} \sum_{w \in V_{k,\alpha}} X_{w,k}$$

By Lemma 7 and Lemma 8, for large values of k the redefinition coincides with the original definition with high probability:

Lemma 9. *For $\delta > 0$, let $\alpha = \sqrt{6 \ln \frac{4m}{\delta}}$. For $k > 18 \ln \frac{4m}{\delta}$, we have $\forall^\delta S$: $M_k = M_{k,\alpha}$, $G_k = G_{k,\alpha}$, and $\hat{M}_k = \hat{M}_{k,\alpha}$.*

[4] The Good-Turing estimator is usually defined as $(\frac{k+1}{m})n_{k+1}$. The two definitions are almost identical for small values of k. Following [19], we use our definition, which makes the calculations slightly simpler.

Since the minimal probability of a word in $V_{k,\alpha}$ is $\Omega\left(\frac{k}{m}\right)$, we derive:

Lemma 10. *Let $\alpha > 0$ and $k > 3\alpha^2$. Then, $|V_{k,\alpha}| = O\left(\frac{m}{k}\right)$.*

Using Lemma 4, we derive:

Lemma 11. *Let $\alpha > 0$ and $3\alpha^2 < k \leq \frac{m}{2}$. Let $w \in V_{k,\alpha}$. Then, $E[X_{w,k}] = P(c_w = k) = O\left(\frac{1}{\sqrt{k}}\right)$.*

3.1 Good-Turing Estimator

The following lemma, based on the definition of the binomial distribution, was shown in Theorem 1 of [19].

Lemma 12. *For any $k < m$, and $w \in V$, we have:*

$$p_w P(c_w = k) = \frac{k+1}{m-k} P(c_w = k+1)(1 - p_w)$$

The following lemma bounds the expectations of the redefined hitting mass, its Good-Turing estimator, and their difference.

Lemma 13. *Let $\alpha > 0$ and $3\alpha^2 < k < \frac{m}{2}$. We have $E[M_{k,\alpha}] = O\left(\frac{1}{\sqrt{k}}\right)$, $E[G_{k,\alpha}] = O\left(\frac{1}{\sqrt{k}}\right)$, and $|E[G_{k,\alpha}] - E[M_{k,\alpha}]| = O\left(\frac{\sqrt{k}}{m}\right)$.*

Using the *negative association* notion, we can show a preliminary bound for Good-Turing estimation error:

Lemma 14. *For $\delta > 0$ and $18 \ln \frac{8m}{\delta} < k < \frac{m}{2}$, we have $\forall^\delta S$:*

$$|G_k - M_k| = O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right)$$

Lemma 15. *Let $\delta > 0$, $k > 0$. Let $U \subseteq V$. Let $\{b_w : w \in U\}$ be a set of weights, such that $b_w \in [0, B]$. Let $X_k = \sum_{w \in U} b_w X_{w,k}$, and $\mu = E[X_k]$. We have:*

$$\forall^\delta S,\ |X_k - \mu| \leq \max\left\{\sqrt{4B\mu \ln\left(\frac{6\sqrt{m}}{\delta}\right)}, 2B \ln\left(\frac{6\sqrt{m}}{\delta}\right)\right\}$$

Proof. By Lemma 6, combined with Lemma 3, we have:

$$\begin{aligned} P(|X_k - \mu| > \epsilon) &\leq 6\sqrt{m}\ \exp\left(-\frac{\epsilon^2}{B(2\mu + \epsilon)}\right) \\ &\leq \max\left\{6\sqrt{m}\ \exp\left(-\frac{\epsilon^2}{4B\mu}\right), 6\sqrt{m}\ \exp\left(-\frac{\epsilon}{2B}\right)\right\}\ , \quad (1) \end{aligned}$$

where (1) follows by considering $\epsilon \leq 2\mu$ and $\epsilon > 2\mu$ separately. The lemma follows substituting $\epsilon = \max\left\{\sqrt{4B\mu \ln\left(\frac{6\sqrt{m}}{\delta}\right)}, 2B \ln\left(\frac{6\sqrt{m}}{\delta}\right)\right\}$. □

We now derive the concentration bound on the error of the Good-Turing estimator.

Theorem 1. *For $\delta > 0$ and $18 \ln \frac{8m}{\delta} < k < \frac{m}{2}$, we have $\forall^{\delta} S$:*

$$|G_k - M_k| = O\left(\sqrt{\frac{\sqrt{k} \ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m}\right)$$

Proof. Let $\alpha = \sqrt{6 \ln \frac{8m}{\delta}}$. Using Lemma 9, we have $\forall^{\frac{\delta}{2}} S$: $G_k = G_{k,\alpha}$, and $M_k = M_{k,\alpha}$. Recall that $M_{k,\alpha} = \sum_{w \in V_{k,\alpha}} p_w X_{w,k}$ and $G_{k,\alpha} = \sum_{w \in V_{k,\alpha}} \frac{k+1}{m-k} X_{w,k+1}$. Both $M_{k,\alpha}$ and $G_{k,\alpha}$ are linear combinations of $X_{w,k}$ and $X_{w,k+1}$, respectively, where the coefficients' magnitude is $O\left(\frac{k}{m}\right)$, and the expectation, by Lemma 13, is $O\left(\frac{1}{\sqrt{k}}\right)$. By Lemma 15, we have:

$$\forall^{\frac{\delta}{4}} S, |M_{k,\alpha} - E[M_{k,\alpha}]| = O\left(\sqrt{\frac{\sqrt{k} \ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m}\right) \tag{2}$$

$$\forall^{\frac{\delta}{4}} S, \ |G_{k,\alpha} - E[G_{k,\alpha}]| = O\left(\sqrt{\frac{\sqrt{k} \ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m}\right) \tag{3}$$

Combining (2), (3), and Lemma 13, we have $\forall^{\delta} S$:

$$\begin{aligned} |G_k - M_k| &= |G_{k,\alpha} - M_{k,\alpha}| \\ &\leq |G_{k,\alpha} - E[G_{k,\alpha}]| + |M_{k,\alpha} - E[M_{k,\alpha}]| + |E[G_{k,\alpha}] - E[M_{k,\alpha}]| \\ &= O\left(\sqrt{\frac{\sqrt{k} \ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m} + \frac{\sqrt{k}}{m}\right) = O\left(\sqrt{\frac{\sqrt{k} \ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m}\right), \end{aligned}$$

which completes the proof. □

3.2 Empirical Frequencies Estimator

In this section we bound the error of the empirical frequencies estimator $\hat{M}_k$.

Theorem 2. *For $\delta > 0$ and $18 \ln \frac{8m}{\delta} < k < \frac{m}{2}$, we have:*

$$\forall^{\delta} S, \ |M_k - \hat{M}_k| = O\left(\frac{\sqrt{k} \left(\ln \frac{m}{\delta}\right)^{\frac{3}{2}}}{m} + \frac{\sqrt{\ln \frac{m}{\delta}}}{k}\right)$$

Proof. Let $\alpha = \sqrt{6 \ln \frac{8m}{\delta}}$. By Lemma 9, we have $\forall^{\frac{\delta}{2}} S$: $\hat{M}_k = \hat{M}_{k,\alpha}$, and $M_k = M_{k,\alpha}$. Let $V^{-}_{k,\alpha} = \{w \in V_{k,\alpha} : p_w < \frac{k}{m}\}$, and $V^{+}_{k,\alpha} = \{w \in V_{k,\alpha} : p_w > \frac{k}{m}\}$. Let

$$X_- = \sum_{w \in V_{k,\alpha}^-} \left(\frac{k}{m} - p_w\right) X_{w,k}, \qquad X_+ = \sum_{w \in V_{k,\alpha}^+} \left(p_w - \frac{k}{m}\right) X_{w,k},$$

and let $X_?$ specify either X_- or X_+. By the definition, for $w \in V_{k,\alpha}$ we have $\left|\frac{k}{m} - p_w\right| = O\left(\frac{\alpha\sqrt{k}}{m}\right)$. By Lemma 10, $|V_{k,\alpha}| = O\left(\frac{m}{k}\right)$. By Lemma 11, for $w \in V_{k,\alpha}$ we have $E[X_{w,k}] = O\left(\frac{1}{\sqrt{k}}\right)$. Therefore,

$$|E[X_?]| \leq \sum_{w \in V_{k,\alpha}} \left|\frac{k}{m} - p_w\right| E[X_{w,k}] = O\left(\frac{m}{k}\frac{\alpha\sqrt{k}}{m}\frac{1}{\sqrt{k}}\right) = O\left(\frac{\alpha}{k}\right) \tag{4}$$

Both X_- and X_+ are linear combinations of $X_{w,k}$, where the coefficients are $O\left(\frac{\alpha\sqrt{k}}{m}\right)$ and the expectation is $O\left(\frac{\alpha}{k}\right)$. Therefore, by Lemma 15, we have:

$$\forall^{\frac{\delta}{4}} S: \qquad |X_? - E[X_?]| = O\left(\sqrt{\frac{\alpha^4}{m\sqrt{k}}} + \frac{\alpha^3\sqrt{k}}{m}\right) \tag{5}$$

By the definition of X_- and X_+, $M_{k,\alpha} - \hat{M}_{k,\alpha} = X_+ - X_-$. Combining (4) and (5), we have $\forall^\delta S$:

$$\begin{aligned}
|M_k - \hat{M}_k| &= |M_{k,\alpha} - \hat{M}_{k,\alpha}| = |X_+ - X_-| \\
&\leq |X_+ - E[X_+]| + E[X_+] + |X_- - E[X_-]| + E[X_-] \\
&= O\left(\sqrt{\frac{\alpha^4}{m\sqrt{k}}} + \frac{\alpha^3\sqrt{k}}{m} + \frac{\alpha}{k}\right) = O\left(\frac{\sqrt{k}\left(\ln\frac{m}{\delta}\right)^{\frac{3}{2}}}{m} + \frac{\sqrt{\ln\frac{m}{\delta}}}{k}\right),
\end{aligned}$$

since $\sqrt{ab} = O(a+b)$, and we use $a = \frac{\alpha^3\sqrt{k}}{m}$ and $b = \frac{\alpha}{k}$. □

3.3 Combined Estimator

In this section we combine the Good-Turing estimator with the empirical frequencies to derive a combined estimator, which is accurate for all values of k.

Definition 4. *We define $\tilde{M}_k$, a combined estimator for M_k, by:*

$$\tilde{M}_k = \begin{cases} G_k & k \leq m^{\frac{2}{5}} \\ \hat{M}_k & k > m^{\frac{2}{5}} \end{cases}$$

Lemma 16. *(Theorem 3 at [19]) Let $k \in \{0, \ldots, m\}$. For any $\delta > 0$, we have:*

$$\forall^\delta S: \qquad |G_k - M_k| = O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\left(k + \ln\frac{m}{\delta}\right)\right)$$

The next theorem shows that $\tilde{M}_k$ has an error bounded by $\tilde{O}\left(m^{-\frac{2}{5}}\right)$, for any k. For small k, we use Lemma 16. Theorem 1 is used for $18 \ln \frac{8m}{\delta} < k \leq m^{\frac{2}{5}}$. Theorem 2 is used for $m^{\frac{2}{5}} < k < \frac{m}{2}$. The complete proof also handles $k \geq \frac{m}{2}$.

Theorem 3. *Let $\delta > 0$. For any $k \in \{0, \ldots, m\}$, we have:*

$$\forall^{\delta} S, |\tilde{M}_k - M_k| = \tilde{O}\left(m^{-\frac{2}{5}}\right)$$

The next theorem shows a weak lower bound for approximating M_k. It applies to estimating M_k based on an independent sample. This induces the "weak" notation, since G_k, as well as $\hat{M}_k$, base on the same sample as M_k.

Theorem 4. *Suppose that the vocabulary consists of $\frac{m}{k}$ words distributed uniformly (i.e. $p_w = \frac{k}{m}$), where $1 \ll k \ll m$. The variance of M_k is $\Theta\left(\frac{\sqrt{k}}{m}\right)$.*

4 Leave-One-Out Estimation of Log-Loss

Many NLP applications use log-loss as the learning performance criteria. Since the log-loss depends on the underlying probability P, its value cannot be explicitly calculated, and must be approximated. The main result of this section, Theorem 5, shows an upper bound on the leave-one-out estimation of the log-loss, assuming a general family of learning algorithms.

Given a sample $S = \{s_1, \ldots, s_m\}$, the goal of a learning algorithm is to approximate the true probability P by some probability Q. We denote the probability assigned by the learning algorithm to a word w by q_w.

Definition 5. *We assume that any two words with equal sample frequency are assigned equal probabilities in Q, and therefore denote q_w by $q(c_w)$. Let the log-loss of a distribution Q be:*

$$L = \sum_{w \in V} p_w \ln \frac{1}{q_w} = \sum_k M_k \ln \frac{1}{q(k)}$$

Let the leave-one-out estimation, q'_w, be the probability assigned to w, when one of its instances is removed. We assume that any two words with equal sample frequency are assigned equal leave-one-out probability estimation, and therefore denote q'_w by $q'(c_w)$. We define the leave-one-out estimation of the log-loss as:

$$L_{leave-one} = \sum_{w \in V} \frac{c_w}{m} \ln \frac{1}{q'_w} = \sum_{k>0} \frac{k n_k}{m} \ln \frac{1}{q'(k)}$$

Let $L_w = L(c_w) = \ln \frac{1}{q(c_w)}$, and $L'_w = L'(c_w) = \ln \frac{1}{q'(c_w)}$. Let $L_{max} = \max_k \max(L(k), L'(k+1))$.

In this section we discuss a family of learning algorithms, that receive the sample as an input. Assuming an accuracy parameter δ, we require the following properties to hold:

1. Starting from a certain number of appearances, the estimation is close to the sample frequency. Specifically, for some $\alpha, \beta \in [0,1]$,

$$\forall k \geq \ln\left(\frac{4m}{\delta}\right),\ q(k) = \frac{k-\alpha}{m-\beta} \tag{6}$$

2. The algorithm is stable when a word is extracted from the sample:

$$\forall m, \qquad 2 \leq k \leq 10 \ln \tfrac{4m}{\delta}, \qquad |L'(k+1) - L(k)| = O\left(\frac{1}{m}\right) \tag{7}$$

$$\forall m, \forall S \ s.t.\ n_1^S > 0,\ k \in \{0,1\}, \quad |L'(k+1) - L(k)| = O\left(\frac{1}{n_1^S}\right) \tag{8}$$

An example of such an algorithm is the following leave-one-out algorithm (we assume that the vocabulary is large enough so that $n_0 + n_1 > 0$):

$$q_w = \begin{cases} \frac{N-n_0-1}{(n_0+n_1)(m-1)} & c_w \leq 1 \\ \frac{c_w-1}{m-1} & c_w \geq 2 \end{cases}$$

The next lemma shows that the expectation of the leave-one-out method is a good approximation for the per-word expectation of the logarithmic loss.

Lemma 17. *Let $0 \leq \alpha \leq 1$, and $y \geq 1$. Let $B_n \sim Bin(n,p)$ be a binomial random variable. Let $f_y(x) = \ln(\max(x,y))$. Then,*

$$0 \leq E\left[p f_y(B_n - \alpha) - \frac{B_n}{n} f_y(B_n - \alpha - 1)\right] \leq \frac{3p}{n}$$

Sketch of Proof. For a real valued function F (here $F(x) = f_y(x-\alpha)$), we have:

$$\begin{aligned} E\left[\frac{B_n}{n} F(B_n - 1)\right] &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \frac{x}{n} F(x-1) \\ &= p \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} F(x-1) \\ &= pE[F(B_{n-1})]\ , \end{aligned}$$

where we used $\binom{n}{x}\frac{x}{n} = \binom{n-1}{x-1}$. The rest of the proof follows by algebraic manipulations, and the definition of the binomial distribution (see [7] for details). □

Lemma 18. *Let $\delta > 0$. We have $\forall^{\delta} S$: $n_2 = O\left(\left(\sqrt{m \ln \frac{1}{\delta}} + n_1\right) \ln \frac{m}{\delta}\right)$.*

Theorem 5. *For $\delta > 0$, we have:*

$$\forall^{\delta} S, |L - L_{leave-one}| = O\left(L_{max}\sqrt{\frac{(\ln\frac{m}{\delta})^4 \ln\frac{\ln\frac{m}{\delta}}{\delta}}{m}}\right)$$

Proof. Let $y_w = \left(1 - \sqrt{\frac{3}{5}}\right) p_w m - 2$. By Lemma 7, with $\lambda = 5$, we have $\forall^{\frac{\delta}{2}} S$:

$$\forall w \in V : p_w > \frac{3\ln\frac{4m}{\delta}}{m}, \qquad \left|p_w - \frac{c_w}{m}\right| \leq \sqrt{\frac{3p_w \ln\frac{4m}{\delta}}{m}} \tag{9}$$

$$\forall w \in V : p_w > \frac{5\ln\frac{4m}{\delta}}{m}, c_w > y_w + 2 \geq (5 - \sqrt{15})\ln\frac{4m}{\delta} > \ln\frac{4m}{\delta} \tag{10}$$

Let $V_H = \left\{w \in V : p_w > \frac{5\ln\frac{4m}{\delta}}{m}\right\}$ and $V_L = V \setminus V_H$. We have:

$$|L - L_{leave-one}| \leq \left|\sum_{w\in V_H}\left(p_w L_w - \frac{c_w}{m}L'_w\right)\right| + \left|\sum_{w\in V_L}\left(p_w L_w - \frac{c_w}{m}L'_w\right)\right| \tag{11}$$

We start by bounding the first term of (11). By (10), we have $\forall w \in V_H, c_w > y_w + 2 > \ln\frac{4m}{\delta}$. Assumption (6) implies that $q_w = \frac{c_w - \alpha}{m - \beta}$, therefore $L_w = \ln\frac{m-\beta}{c_w - \alpha} = \ln\frac{m-\beta}{\max(c_w - \alpha, y_w)}$, and $L'_w = \ln\frac{m-1-\beta}{c_w - 1 - \alpha} = \ln\frac{m-1-\beta}{\max(c_w - 1 - \alpha, y_w)}$. Let

$$Err_w^H = \frac{c_w}{m}\ln\frac{m-\beta}{\max(c_w - 1 - \alpha, y_w)} - p_w \ln\frac{m-\beta}{\max(c_w - \alpha, y_w)}$$

We have:

$$\begin{aligned}\left|\sum_{w\in V_H}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| &= \left|\sum_{w\in V_H} Err_w^H + \ln\frac{m-1-\beta}{m-\beta}\sum_{w\in V_H}\frac{c_w}{m}\right| \\ &\leq \left|\sum_{w\in V_H} Err_w^H\right| + O\left(\frac{1}{m}\right)\end{aligned} \tag{12}$$

We bound $\left|\sum_{w\in V_H} Err_w^H\right|$ using McDiarmid's inequality. As in Lemma 17, let $f_w(x) = \ln(\max(x, y_w))$. We have:

$$E\left[Err_w^H\right] = \ln(m-\beta)E\left[\frac{c_w}{m} - p_w\right] + E\left[p_w f_w(c_w - \alpha) - \frac{c_w}{m}f_w(c_w - 1 - \alpha)\right]$$

The first expectation equals 0, the second can be bounded using Lemma 17:

$$\left|\sum_{w\in V_H} E\left[Err_w^H\right]\right| \leq \sum_{w\in V_H}\frac{3p_w}{m} = O\left(\frac{1}{m}\right) \tag{13}$$

In order to use McDiarmid's inequality, we bound the change of $\sum_{w\in V_H} Err_w^H$ as a function of a single change in the sample. Suppose that a word u is replaced by a word v. This results in decrease for c_u, and increase for c_v. Recalling that $y_w = \Omega(mp_w)$, the change of Err_u^H, as well as the change of Err_v^H, is bounded by $O\left(\frac{\ln m}{m}\right)$ (see [7] for details).

By (12), (13), and Lemma 2, we have $\forall^{\frac{\delta}{16}} S$:

$$\left|\sum_{w\in V_H}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| = O\left(\sqrt{\frac{(\ln m)^2 \ln\frac{1}{\delta}}{m}}\right) \tag{14}$$

Next, we bound the second term of (11). By Lemma 8, we have $\forall^{\frac{\delta}{4}} S$:

$$\forall w \in V \ s.t. \ \ p_w \le \frac{3\ln\frac{4m}{\delta}}{m}, c_w \le 6\ln\frac{4m}{\delta} \tag{15}$$

Let $b = 5\ln\frac{4m}{\delta}$. By (9) and (15), for any w such that $p_w \le \frac{b}{m}$, we have:

$$\frac{c_w}{m} \le \max\left\{p_w + \sqrt{\frac{3p_w\ln\frac{4m}{\delta}}{m}}, \frac{6\ln\frac{4m}{\delta}}{m}\right\} \le \frac{(5+\sqrt{3*5})\ln\frac{4m}{\delta}}{m} < \frac{2b}{m}$$

Therefore $\forall w \in V_L$, we have $c_w < 2b$. Let $n_k^L = |V_L \cap S_k|$, $G_{k-1}^L = \frac{k}{m-k+1}n_k^L$, and $M_k^L = \sum_{w\in V_L\cap S_k} p_w$. Using algebraic manipulations (see [7] for details), we have:

$$\begin{aligned}\left|\sum_{w\in V_L}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| &= \left|\sum_{k=1}^{2b}\frac{kn_k^L}{m}L'(k) - \sum_{k=0}^{2b-1}M_k^L L(k)\right| \\ &\le \sum_{k=0}^{2b-1}G_k^L|L'(k+1)-L(k)| + \sum_{k=0}^{2b-1}|G_k^L - M_k^L|L(k) + O\left(\frac{bL_{max}}{m}\right)\end{aligned} \tag{16}$$

The first sum of (16) is bounded using (7), (8), and Lemma 18 (with accuracy $\frac{\delta}{16}$). The second sum of (16) is bounded using Lemma 16 separately for every $k < 2b$ with accuracy $\frac{\delta}{16b}$. Since the proof of Lemma 16 also holds for G_k^L and M_k^L (instead of G_k and M_k), we have $\forall^{\frac{\delta}{8}} S$, for every $k < 2b$, $|G_k^L - M_k^L| = O\left(b\sqrt{\frac{\ln\frac{b}{\delta}}{m}}\right)$. Therefore (the details can be found at [7]),

$$\left|\sum_{w\in V_L}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| = O\left(L_{max}\sqrt{\frac{b^4\ln\frac{b}{\delta}}{m}}\right) \tag{17}$$

The proof follows by combining (11), (14), and (17). □

5 Log-Loss A Priori

Section 4 bounds the error of the leave-one-out estimation of the log-loss. In this section we analyze the log-loss itself. We denote the learning error (equivalent to the log-loss) as the KL-divergence between the true and the estimated distribution. We refer to a general family of learning algorithms, and show an upper bound for the learning error.

Let $\alpha \in (0,1)$ and $\tau \geq 1$. We define an (absolute discounting) algorithm $A_{\alpha,\tau}$, which "removes" $\frac{\alpha}{m}$ probability mass from words appearing at most τ times, and uniformly spreads it among the unseen words. We denote by $n_{1\ldots\tau} = \sum_{i=1}^{\tau} n_i$ the number of words with count between 1 and τ. The learned probability Q is defined by :

$$q_w = \begin{cases} \frac{\alpha n_{1\ldots\tau}}{m n_0} & c_w = 0 \\ \frac{c_w - \alpha}{m} & 0 < c_w \leq \tau \\ \frac{c_w}{m} & \tau < c_w \end{cases}$$

Theorem 6. *For any $\delta > 0$ and $\lambda > 3$, such that $\tau < (\lambda - \sqrt{3\lambda}) \ln \frac{8m}{\delta}$, let $x = \frac{\lambda \ln \frac{8m}{\delta}}{m}$ and Let $N_x = |\{w \in V : p_w > x\}|$. Then, the learning error of $A_{\alpha,\tau}$ is bounded $\forall^{\delta} S$ by:*

$$0 \leq \sum_{w \in V} p_w \ln \left(\frac{p_w}{q_w} \right) = \tilde{O} \left(M_0 \ln N + x\sqrt{m} + \frac{N_x}{m} \right)$$

Since N_x includes only words with $p_w > x$, it is bounded by $\frac{1}{x}$. Therefore, $x = m^{-\frac{3}{4}}$ gives a bound of $\tilde{O}\left(M_0 \ln N + m^{-\frac{1}{4}}\right)$. Lower loss can be achieved for specific distributions, such as those with small M_0 and small N_x (for some reasonable x).

Acknowledgements

We are grateful to David McAllester for his important contributions in the early stages of this research.

References

1. D. Angluin and L. G. Valiant. Fast Probabilistic Algorithms for Hamiltonian Circuits and matchings, In *Journal of Computer and System Sciences*, 18:155-193, 1979.
2. S. F. Chen, Building Probabilistic Models for Natural Language, *Ph.D. Thesis*, Harvard University, 1996.
3. S. F. Chen and J. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, *Technical Report* TR-10-98, Harvard University, 1998.

4. K. W. Church and W. A. Gale, A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams, In *Computer Speech and Language*, 5:19-54, 1991.
5. J. R. Curran and M. Osborne, A Very Very Large Corpus Doesn't Always Yield Reliable Estimates, In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 126-131, 2002.
6. L. Devroye, L. Györfi, and G. Lugosi, A Probabilistic Theory of Pattern Recognition, *Springer-Verlag, New York*, 1996.
7. E. Drukh, Concentration Bounds for Unigrams Language Model, *M.Sc. Thesis*, Tel Aviv University, 2004.
8. D. P. Dubhashi and D. Ranjan, Balls and Bins: A Study in Negative Dependence, In *Random Structures and Algorithms*, 13(2):99-124, 1998.
9. W. Gale, Good-Turing Smoothing Without Tears, In *Journal of Quantitative Linguistics*, 2:217–37, 1995.
10. I. J. Good, The Population Frequencies of Species and the Estimation of Population Parameters, In *Biometrika*, 40(16):237-264, 1953.
11. I. J. Good, Turing's Anticipation of Empirical Bayes in Connection with the Cryptanalysis of the Naval Enigma, In *Journal of Statistical Computation and Simulation*, 66(2):101-112, 2000.
12. W. Hoeffding, On the Distribution of the Number of Successes in Independent Trials, In *Annals of Mathematical Statistics*, 27:713-721, 1956.
13. W. Hoeffding, Probability Inequalities for Sums of Bounded Random Variables, In *Journal of the American Statistical Association*, 58:13-30, 1963.
14. S. B. Holden, PAC-like Upper Bounds for the Sample Complexity of Leave-One-Out Cross-Validation, In *Proceesings of the Ninth Annual ACM Workshop on Computational Learning Theory*, pages 41-50, 1996.
15. S. M. Katz, Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400-401, 1987.
16. M. Kearns and D. Ron, Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation, In *Neural Computation*, 11(6):1427-1453, 1999.
17. S. Kutin, Algorithmic Stability and Ensemble-Based Learning, *Ph.D. Thesis*, University of Chicago, 2002.
18. D. McAllester and L. Ortiz, Concentration Inequalities for the Missing Mass and for Histogram Rule Error, In *Journal of Machine Learning Research, Special Issue on Learning Theory*, 4(Oct):895-911, 2003.
19. D. McAllester and R. E. Schapire, On the Convergence Rate of Good-Turing Estimators, In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 1-6, 2000.
20. D. McAllester and R. E. Schapire, Learning Theory and Language Modeling, In *Seventeenth International Joint Conference on Artificial Intelligence*, 2001.
21. C. McDiarmid, On the Method of Bounded Differences, In *Surveys in Combinatorics 1989*, Cambridge University Press, Cambridge, 148-188, 1989.
22. A. Orlitsky, N. P. Santhanam, and J. Zhang, Always Good Turing: Asymptotically Optimal Probability Estimation, In *Science*, 302(Oct):427-431, 2003 (in Reports).