Journal of Machine Learning Research 6 (2005) 1231–1264 Submitted 11/04; Revised 6/05; Published 8/05

Concentration Bounds for Unigram Language Models

Evgeny Drukh DRUKH@POST.TAU.AC.IL
Yishay Mansour MANSOUR@POST.TAU.AC.IL
School of Computer Science
Tel Aviv University
Tel Aviv, 69978, Israel

Editor: John Lafferty

Abstract

We show several high-probability concentration bounds for learning unigram language models. 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 high-probability 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 with high probability its error is bounded by 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. The leave-one-out method can be used for estimating the log-loss of the unigram model. We show that its error has a high-probability bound of approximately $O\left(\frac{1}{\sqrt{m}}\right)$, for any underlying distribution.

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

Keywords: Good-Turing estimators, logarithmic loss, leave-one-out estimation, Chernoff bounds

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 unigram models, which try to estimate the distribution of single words. As it has been observed empirically on many occasions (Chen, 1996; Curran and Osborne, 2002), 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 dates back to World War II, and were published first in 1953 (Good, 1953, 2000). It has been extensively used in language modeling applications since then (Katz, 1987; Church and Gale, 1991; Chen, 1996; Chen and Goodman, 1998). However, their theoretical convergence rate in various models has been studied only in the recent years (McAllester and Schapire, 2000, 2001; Kutin, 2002; McAllester and Ortiz, 2003; Orlitsky et al., 2003). For estimation of the probability of all words appearing exactly k times in a sample of size m, McAllester and Schapire (2000) derive a high probability bound on Good-Turing estimator 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 estimator has 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 weak 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, for example $k \leq 5$, as by Katz (1987) or Chen (1996), 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 used in NLP applications is the *Kullback-Leibler (KL) divergence*. For a true distribution $P = \{p_x\}$, and an estimated distribution $Q = \{q_x\}$, both 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 underlying distribution, 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 performs averaging all the possible estimations, where a single item is chosen for testing, and the rest are used for training. This procedure has an advantage of using the entire sample, and in addition it is rather simple and usually can be easily implemented. The existing theoretical analysis of the *leave-one-out* method (Holden, 1996; Kearns and Ron, 1999) shows general high probability 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 underlying distribution and a general family of learning algorithms. It gives a theoretical justification for effective use of *leave-one-out* estimation for the *log-loss*.

We also analyze the concentration of the *log-loss* itself, not based of an empirical measure. We address the characteristics of the underlying distribution affecting the *log-loss*. We find such a characteristic, defining a tight bound for the *log-loss* value.

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 error bound function $f(\cdot)$, which holds with probability $1-\delta$, 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.

1.4 Paper Organization

Section 2 shows several standard concentration inequalities, together with their technical applications regarding the maximum-likelihood approximation. Section 3 shows the error bounds for the

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

k-hitting mass estimation. Section 4 bounds the error for the leave-one-out estimation of the logarithmic loss. Section 5 shows the bounds for the a priori logarithmic loss. Appendix A includes the technical proofs.

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, 1963) 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 $\varepsilon > 0$,*

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

The next lemma is a variant of an extension of Hoeffding's inequality, by McDiarmid (1989).

Lemma 2 *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, that is*

$$\sup_{\forall j \neq i, Y_j = Y'_j} (|f(Y) - f(Y')|) \leq d_i.$$

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 and Valiant, 1979) 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 $\varepsilon > 0$,*

$$P\left(\sum_i Y_i < \mu - \varepsilon\right) \quad \leq \quad \exp\left(-\frac{\varepsilon^2}{2\mu B}\right),$$
$$P\left(\sum_i Y_i > \mu + \varepsilon\right) \quad \leq \quad \exp\left(-\frac{\varepsilon^2}{(2\mu + \varepsilon)B}\right).$$

Definition 4 *(Dubhashi and Ranjan, 1998) A set of random variables $Y_1, \ldots, Y_n$ is called "negatively associated", if it satisfies for any two disjoint subsets I and J of $\{1, \ldots, n\}$, and any two non-decreasing, or any two non-increasing, functions f from $R^{|I|}$ to R and g from $R^{|J|}$ to R:*

$$E[f(Y_i : i \in I) g(Y_j : j \in J)] \leq E[f(Y_i : i \in I)] E[g(Y_j : j \in J)].$$

The next lemma is based on the *negative association* analysis. It follows directly from Theorem 14 and Proposition 7 of Dubhashi and Ranjan (1998).

Lemma 5 *For any set of N non-decreasing, or N non-increasing functions $\{f_w : w \in V\}$, any Chernoff-style bound on $\sum_{w\in V} f_w(c_w)$, pretending that c_w are independent, is valid. In particular, Lemmas 1 and 2 apply for $\{Y_1,...,Y_n\} = \{f_w(c_w) : w \in V\}$.*

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

Lemma 6 *Let $X \sim Bin(n,p)$ be 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 exists some function $T_x = \exp\left(\frac{1}{12x} + O\left(\frac{1}{x^2}\right)\right)$, 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 deals with the number of successes in independent trials.

Lemma 7 *(Hoeffding, 1956) Let $Y_1,\ldots,Y_n \in \{0,1\}$ be a sequence of independent trials, with $p_i = E[Y_i]$. Let $X = \sum_i Y_i$ be the number of successes, 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*

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

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 8 *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 $\varepsilon > 0$,*

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

The following lemmas provide concentration bounds for maximum-likelihood estimation of p_w by $\hat{p}_w = \frac{c_w}{m}$. The first lemma shows that words with "high" probability have a "high" count in the sample.

Lemma 9 *Let $\delta > 0$, and $\lambda \geq 3$. We have $\forall^\delta S$:*

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

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

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 30.
3. The *negative association* analysis (Lemma 5) shows that a sum of monotone functions of multinomial distribution components must obey Chernoff-style bounds pretending that the components are independent. In some sense, our result extends this notion, since it does not require the functions to be monotone.

The second lemma shows that words with "low" probability have a "low" count in the sample.

Lemma 10 *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 $c_w \leq 6\ln\frac{m}{\delta}$.*

The following lemma derives the bound as a function of the count in the sample (and not as a function of the unknown probability).

Lemma 11 *Let $\delta > 0$. Then, $\forall^{\delta} S$:*

$$\forall w \in V,\ \ s.t. \quad c_w > 18\ln\tfrac{4m}{\delta}, \qquad |mp_w - c_w| \leq \sqrt{6c_w \ln\frac{4m}{\delta}}.$$

The following is a general concentration bound.

Lemma 12 *For any $\delta > 0$, and any word $w \in V$, we have*

$$\forall^{\delta} S, \qquad \left|\frac{c_w}{m} - p_w\right| < \sqrt{\frac{3\ln\frac{2}{\delta}}{m}}.$$

The following lemma bounds the probability of words that do not appear in the sample.

Lemma 13 *Let $\delta > 0$. Then, $\forall^{\delta} S$:*

$$\forall w \notin S, \qquad mp_w < \ln\frac{m}{\delta}.$$

3. K-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 k-hitting mass". We analyze the Good-Turing estimator, the empirical frequencies estimator, and a combined estimator.

Definition 14 *We define the k-hitting mass M_k, its empirical frequencies estimator $\hat{M}_k$, and its Good-Turing estimator G_k as*[4]

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

The outline of this section is as follows. Definition 16 slightly redefines the k-hitting mass and its estimators. Lemma 17 shows that this redefinition has a negligible influence. Then, we analyze the estimation errors using the concentration inequalities from Section 2.

Lemmas 20 and 21 bound the expectation of the Good-Turing estimator error, following McAllester and Schapire (2000). Lemma 23 bounds the deviation of the error, using the negative association analysis. A tighter bound, based on Lemma 8, is achieved at Theorem 25. Theorem 26 analyzes the error of the empirical frequencies estimator. Theorem 29 refers to the combined estimator. Finally, Theorem 30 shows a weak lower bound for the k-hitting mass estimation.

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, as their quotient equals $1 - \frac{k}{m}$. Following McAllester and Schapire (2000), our definition makes the calculations slightly simpler.

Definition 15 *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.*

The following definition concentrates on words whose frequencies are close to their probabilities.

Definition 16 *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:*

$$\begin{aligned} M_{k,\alpha} &= \sum_{w \in S_k \cap V_{k,\alpha}} p_w = \sum_{w \in V_{k,\alpha}} p_w X_{w,k}, \\ G_{k,\alpha} &= \frac{k+1}{m-k} |S_{k+1} \cap V_{k,\alpha}| = \frac{k+1}{m-k} \sum_{w \in V_{k,\alpha}} X_{w,k+1}, \\ \hat{M}_{k,\alpha} &= \frac{k}{m} |S_k \cap V_{k,\alpha}| = \frac{k}{m} \sum_{w \in V_{k,\alpha}} X_{w,k}. \end{aligned}$$

By Lemma 11, for large values of k the redefinition coincides with the original definition with high probability:

Lemma 17 *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 $\hat{M}_k = \hat{M}_{k,\alpha}$.*

Proof By Lemma 11, we have

$$\forall^\delta S, \quad \forall w : c_w > 18 \ln \tfrac{4m}{\delta}, \quad |m p_w - c_w| \leq \sqrt{6 c_w \ln \frac{4m}{\delta}} = \alpha \sqrt{c_w}.$$

This means that any word w with $c_w = k$ has

$$\frac{k - \alpha\sqrt{k}}{m} \leq p_w \leq \frac{k + \alpha\sqrt{k}}{m} < \frac{k + 1 + \alpha\sqrt{k+1}}{m}.$$

Therefore $w \in V_{k,\alpha}$, completing the proof for M_k and $\hat{M}_k$. Since $\alpha < \sqrt{k}$, any word w with $c_w = k+1$ has

$$\frac{k - \alpha\sqrt{k}}{m} < \frac{k + 1 - \alpha\sqrt{k+1}}{m} \leq p_w \leq \frac{k + 1 + \alpha\sqrt{k+1}}{m},$$

which yields $w \in V_{k,\alpha}$, completing the proof for G_k. ■

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

Lemma 18 *Let $\alpha > 0$ and $k > 3\alpha^2$. Then, $|V_{k,\alpha}| = O\left(\frac{m}{k}\right)$.*

Proof We have $\alpha < \frac{\sqrt{k}}{\sqrt{3}}$. Any word $w \in V_{k,\alpha}$ has $p_w \geq \frac{k-\alpha\sqrt{k}}{m} > \frac{k}{m}\left(1-\frac{1}{\sqrt{3}}\right)$. Therefore,

$$|V_{k,\alpha}| < \frac{m}{k}\frac{1}{1-\frac{1}{\sqrt{3}}} = O\left(\frac{m}{k}\right),$$

which completes the proof. ■

Using Lemma 6, we derive:

Lemma 19 *Let* $\alpha > 0$ *and* $3\alpha^2 < k \leq \frac{m}{2}$. *Let* $w \in V_{k,\alpha}$. *Then,* $E[X_{w,k}] = P(c_w = k) = O\left(\frac{1}{\sqrt{k}}\right)$.

Proof Since $c_w \sim Bin(m, p_w)$ is a binomial random variable, we use Lemma 6:

$$E[X_{w,k}] = P(c_w = k) \leq \frac{1}{\sqrt{2\pi m p_w(1-p_w)}}\frac{T_m}{T_{mp_w}T_{m(1-p_w)}}.$$

For $w \in V_{k,\alpha}$, we have $mp_w = \Omega(k)$, which implies $\frac{T_m}{T_{mp_w}T_{m(1-p_w)}} = O(1)$. Since $p_w \in I_{k,\alpha}$ and $3\alpha^2 < k \leq \frac{m}{2}$, we have

$$\begin{aligned}
\frac{1}{\sqrt{2\pi m p_w(1-p_w)}} &\leq \frac{1}{\sqrt{2\pi\left(k-\alpha\sqrt{k}\right)\left(1-\left(\frac{k+1+\alpha\sqrt{k+1}}{m}\right)\right)}} \\
&< \frac{1}{\sqrt{2\pi k\left(1-\frac{1}{\sqrt{3}}\right)\left(1-\frac{k+1}{m}\left(1+\frac{1}{\sqrt{3}}\right)\right)}} \\
&< \frac{1}{\sqrt{2\pi k\left(1-\frac{1}{\sqrt{3}}\right)\left(1-\left(\frac{1}{2}+\frac{1}{m}\right)\left(1+\frac{1}{\sqrt{3}}\right)\right)}} \\
&= O\left(\frac{1}{\sqrt{k}}\right),
\end{aligned}$$

which completes the proof. ■

3.1 Good-Turing Estimator

The following lemma, directly based on the definition of the binomial distribution, was shown in Theorem 1 of McAllester and Schapire (2000).

Lemma 20 *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 k-hitting mass, its Good-Turing estimator, and their difference.

Lemma 21 *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* $|E[G_{k,\alpha}] - E[M_{k,\alpha}]| = O\left(\frac{\sqrt{k}}{m}\right)$.

Lemma 22 *Let* $\delta > 0$, $k \in \{1,\ldots,m\}$. *Let* $U \subseteq V$, *such that* $|U| = O\left(\frac{m}{k}\right)$. *Let* $\{b_w : w \in U\}$, *such that* $\forall w \in U, b_w \geq 0$ *and* $max_{w\in U} b_w = O\left(\frac{k}{m}\right)$. *Let* $X_k = \sum_{w\in U} b_w X_{w,k}$. *We have* $\forall^{\delta} S$:

$$|X_k - E[X_k]| = O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right).$$

Proof We define $Y_{w,k} = \sum_{i\leq k} X_{w,i}$ be random variable indicating $c_w \leq k$ and $Z_{w,k} = \sum_{i<k} X_{w,i} = Y_{w,k} - X_{w,k}$ be random variable indicating $c_w < k$. Let $Y_k = \sum_{w\in U} b_w Y_{w,k}$ and $Z_k = \sum_{w\in U} b_w Z_{w,k}$. We have

$$X_k \quad = \quad \sum_{w\in U} b_w X_{w,k} = \sum_{w\in U} b_w [Y_{w,k} - Z_{w,k}] = Y_k - Z_k.$$

Both Y_k and Z_k, can be bounded using the Hoeffding inequality. Since $\{b_w Y_{w,k}\}$ and $\{b_w Z_{w,k}\}$ are monotone with respect to $\{c_w\}$, Lemma 5 applies for them. This means that the concentration of their sum is at least as tight as if they were independent. Recalling that $|U| = O\left(\frac{m}{k}\right)$ and $max_{w\in U} b_w = O\left(\frac{k}{m}\right)$, and using Lemma 2 for Y_k and Z_k, we have

$$\forall^{\frac{\delta}{2}} S, \quad |Y_k - E[Y_k]| = O\left(\frac{k}{m}\sqrt{\frac{m}{k}\ln\frac{1}{\delta}}\right),$$
$$\forall^{\frac{\delta}{2}} S, \quad |Z_k - E[Z_k]| = O\left(\frac{k}{m}\sqrt{\frac{m}{k}\ln\frac{1}{\delta}}\right).$$

Therefore,

$$\begin{aligned} |X_k - E[X_k]| &= |Y_k - Z_k - E[Y_k - Z_k]| \\ &\leq |Y_k - E[Y_k]| + |Z_k - E[Z_k]| = O\left(\sqrt{\frac{k\ln\frac{1}{\delta}}{m}}\right), \end{aligned}$$

which completes the proof. ■

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

Lemma 23 *For* $\delta > 0$ *and* $18\ln\frac{8m}{\delta} < k < \frac{m}{2}$, *we have* $\forall^{\delta} S$:

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

Proof Let $\alpha = \sqrt{6 \ln \frac{8m}{\delta}}$. By Lemma 17, we have

$$\forall^{\frac{\delta}{2}} S, \quad G_k = G_{k,\alpha} \wedge M_k = M_{k,\alpha}. \tag{1}$$

By Lemma 21,

$$|E[G_k - M_k]| = |E[G_{k,\alpha} - M_{k,\alpha}]| = O\left(\frac{\sqrt{k}}{m}\right). \tag{2}$$

By Definition 16, $M_{k,\alpha} = \sum_{w \in V_{k,\alpha}} p_w X_{w,k}$ and $G_{k,\alpha} = \sum_{w \in V_{k,\alpha}} \left(\frac{k+1}{m-k}\right) X_{w,k+1}$. By Lemma 18, we have $|V_{k,\alpha}| = O\left(\frac{m}{k}\right)$. Therefore, using Lemma 22 with k for $M_{k,\alpha}$, and with $k+1$ for $G_{k,\alpha}$, we have

$$\forall^{\frac{\delta}{4}} S, \quad |M_{k,\alpha} - E[M_{k,\alpha}]| = O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right), \tag{3}$$

$$\forall^{\frac{\delta}{4}} S, \quad |G_{k,\alpha} - E[G_{k,\alpha}]| = O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right). \tag{4}$$

Combining Equations (1), (2), (3), and (4), we have $\forall^{\delta} S$:

$$\begin{aligned} |G_k - M_k| &= |G_{k,\alpha} - M_{k,\alpha}| \\ &\leq |G_{k,\alpha} - E[G_{k,\alpha}]| + |M_{k,\alpha} - E[M_{k,\alpha}]| + |E[G_{k,\alpha}] - E[M_{k,\alpha}]| \\ &= O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right) + O\left(\frac{\sqrt{k}}{m}\right) = O\left(\sqrt{\frac{k \ln \frac{1}{\delta}}{m}}\right), \end{aligned}$$

which completes the proof. ■

Lemma 24 *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*

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

Proof By Lemma 8, combined with Lemma 3, we have

$$\begin{aligned} P(|X_k - \mu| > \varepsilon) &\leq 6\sqrt{m} \exp\left(-\frac{\varepsilon^2}{B(2\mu + \varepsilon)}\right) \\ &\leq \max\left\{6\sqrt{m} \exp\left(-\frac{\varepsilon^2}{4B\mu}\right), 6\sqrt{m} \exp\left(-\frac{\varepsilon}{2B}\right)\right\}, \end{aligned} \tag{5}$$

where Equation (5) follows by considering $\varepsilon \leq 2\mu$ and $\varepsilon > 2\mu$ separately. The lemma follows substituting $\varepsilon = \max\left\{\sqrt{4B\mu\ln\left(\frac{6\sqrt{m}}{\delta}\right)}, 2B\ln\left(\frac{6\sqrt{m}}{\delta}\right)\right\}$. ■

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

Theorem 25 *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 17, 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 21, is $O\left(\frac{1}{\sqrt{k}}\right)$. By Lemma 24, we have

$$\forall^{\frac{\delta}{4}} S, \quad |M_{k,\alpha} - E[M_{k,\alpha}]| = O\left(\sqrt{\frac{\sqrt{k}\ln\frac{m}{\delta}}{m}} + \frac{k\ln\frac{m}{\delta}}{m}\right), \tag{6}$$

$$\forall^{\frac{\delta}{4}} S, \quad |G_{k,\alpha} - E[G_{k,\alpha}]| = O\left(\sqrt{\frac{\sqrt{k}\ln\frac{m}{\delta}}{m}} + \frac{k\ln\frac{m}{\delta}}{m}\right). \tag{7}$$

Combining Equations (6), (7), and Lemma 21, 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 26 *For $\delta > 0$ and $18\ln\frac{8m}{\delta} < k < \frac{m}{2}$, we have*

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

Proof Let $\alpha = \sqrt{6 \ln \frac{8m}{\delta}}$. By Lemma 17, 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 18, $|V_{k,\alpha}| = O\left(\frac{m}{k} \right)$. By Lemma 19, for $w \in V_{k,\alpha}$ we have $E[X_{w,k}] = O\left(\frac{1}{\sqrt{k}} \right)$. Therefore,

$$|E[X_?]| \quad \leq \quad \sum_{w \in V_{k,\alpha}} \left| \frac{k}{m} - p_w \right| E[X_{w,k}] = O\left(\frac{m}{k} \frac{\alpha\sqrt{k}}{m} \frac{1}{\sqrt{k}} \right) = O\left(\frac{\alpha}{k} \right). \tag{8}$$

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 24, we have

$$\forall^{\frac{\delta}{4}} S: \qquad |X_? - E[X_?]| \quad = \quad O\left(\sqrt{\frac{\alpha^4}{m\sqrt{k}}} + \frac{\alpha^3\sqrt{k}}{m} \right). \tag{9}$$

By the definition of X_- and X_+, $M_{k,\alpha} - \hat{M}_{k,\alpha} = X_+ - X_-$. Combining Equations (8) and (9), 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 uniformly accurate for all values of k.

Definition 27 *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 28 *(McAllester and Schapire, 2000) 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 following 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 28. Theorem 25 is used for $18\ln\frac{8m}{\delta} < k \leq m^{\frac{2}{5}}$. Theorem 26 is used for $m^{\frac{2}{5}} < k < \frac{m}{2}$. The complete proof also handles $k \geq \frac{m}{2}$. The theorem refers to $\tilde{M}_k$ as a probability estimator, and does not show that it is a probability distribution by itself.

Theorem 29 *Let $\delta > 0$. For any $k \in \{0,\ldots,m\}$, we have*

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

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

Theorem 30 *Suppose that the vocabulary consists of $\frac{m}{k}$ words distributed uniformly (that is $p_w = \frac{k}{m}$), 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 32, is 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 31 *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*

$$L \quad = \quad \sum_{w \in V} p_w \ln\frac{1}{q_w} = \sum_{k \geq 0} M_k \ln\frac{1}{q(k)}.$$

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 averaging the loss of each sample word, when it is extracted from the sample and pretended to be the test sample:

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

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 the maximal loss be*

$$L_{max} = \max_k \max \left\{ L(k), L'(k+1) \right\}.$$

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 $\alpha, \beta \in [0,1]$,

$$\forall k \geq \ln \left(\frac{4m}{\delta} \right), \qquad q(k) = \frac{k - \alpha}{m - \beta}. \tag{10}$$

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

$$\forall m, \qquad 2 \leq k \leq 10 \ln \tfrac{4m}{\delta}, \qquad \left| L'(k+1) - L(k) \right| = O\left(\frac{1}{m} \right), \tag{11}$$

$$\forall m, \quad \forall S \; s.t. \; n_1^S > 0, \; k \in \{0,1\}, \qquad \left| L'(k+1) - L(k) \right| = O\left(\frac{1}{n_1^S} \right). \tag{12}$$

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

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

Equation (10) is satisfied by $\alpha = \beta = 1$. Equation (11) is satisfied for $k \geq 2$ by $L(k) - L'(k+1) = \ln \left(\frac{m-1}{m-2} \right) = O\left(\frac{1}{m} \right)$. Equation (12) is satisfied for $k \leq 1$:

$$|L'(1) - L(0)| = \left| \ln \left(\frac{N - n_0 - 1}{N - n_0 - 2} \frac{m-2}{m-1} \right) \right| = O\left(\frac{1}{N - n_0} + \frac{1}{m} \right) = O\left(\frac{1}{n_1} \right),$$
$$|L'(2) - L(1)| = \left| \ln \left(\frac{n_0 + n_1 + 1}{n_0 + n_1} \frac{m-2}{m-1} \right) \right| = O\left(\frac{1}{n_0 + n_1} + \frac{1}{m} \right) = O\left(\frac{1}{n_1} \right).$$

The following is the main theorem of this section. It bounds the deviation between the between the true loss and the *leave one out* estimate. This bound shows that for a general family of learning algorithms, leave-one-out technique can be effectively used to estimate the logarithmic loss, given the sample only. The estimation error bound decreases roughly in proportion to the square root of the sample size, regardless of the underlying distribution.

Theorem 32 *For a learning algorithm satisfying Equations (10), (11), and (12), and* $\delta > 0$*, we have:*

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

The proof of Theorem 32 bounds the estimation error separately for the high-probability and low-probability words. We use Lemma 20 (McAllester and Schapire, 2000) to bound the estimation error for low-probability words. The expected estimation error for the high-probability words is bounded elementarily using the definition of the binomial distribution (Lemma 33). Finally, we use McDiarmid's inequality (Lemma 2) to bound its deviation.

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 33 *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,*

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

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 $\binom{n}{x}\frac{x}{n} = \binom{n-1}{x-1}$. Since $B_n \sim B_{n-1} + B_1$, we have:

$$\begin{aligned} E\left[p f_y(B_n - \alpha) - \frac{B_n}{n} f_y(B_n - \alpha - 1)\right] &= p(E[f_y(B_{n-1} + B_1 - \alpha)] - E[f_y(B_{n-1} - \alpha)]) \\ &= pE\left[\ln\frac{\max(B_{n-1} + B_1 - \alpha, y)}{\max(B_{n-1} - \alpha, y)}\right] \\ &\leq pE\left[\ln\frac{\max(B_{n-1} - \alpha + B_1, y + B_1)}{\max(B_{n-1} - \alpha, y)}\right] \\ &= pE\left[\ln(1 + \frac{B_1}{\max(B_{n-1} - \alpha, y)})\right] \\ &\leq pE\left[\frac{B_1}{\max(B_{n-1} - \alpha, y)}\right]. \end{aligned}$$

Since B_1 and B_{n-1} are independent, we get

$$\begin{aligned}
pE\left[\frac{B_1}{\max(B_{n-1}-\alpha,y)}\right] &= pE[B_1]E\left[\frac{1}{\max(B_{n-1}-\alpha,y)}\right] \\
&= p^2E\left[\frac{1}{\max(B_{n-1}-\alpha,y)}\right] \\
&= p^2\sum_{x=0}^{n-1}\binom{n-1}{x}p^x(1-p)^{n-1-x}\frac{1}{\max(x-\alpha,y)} \\
&= p^2\sum_{x=0}^{n-1}\binom{n-1}{x}p^x(1-p)^{n-1-x}\frac{1}{x+1}\frac{x+1}{\max(x-\alpha,y)} \\
&\leq \frac{p}{n}\max_x\left(\frac{x+1}{\max(x-\alpha,y)}\right)\sum_{x=0}^{n-1}\binom{n}{x+1}p^{x+1}(1-p)^{n-(x+1)} \\
&\leq \frac{3p}{n}(1-(1-p)^n) < \frac{3p}{n}. \qquad (13)
\end{aligned}$$

Equation (13) follows by the following observation: $x+1 \leq 3(x-\alpha)$ for $x \geq 2$, and $x+1 \leq 2y$ for $x \leq 1$. Finally, $pE\left[\ln\frac{\max(B_{n-1}-\alpha+B_1,y)}{\max(B_{n-1}-\alpha,y)}\right] \geq 0$, which implies the lower bound of the lemma. ■

The following lemma bounds n_2 as a function of n_1.

Lemma 34 *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 32 **Proof** Let $y_w = \left(1-\sqrt{\frac{3}{5}}\right)p_w m - 2$. By Lemma 9, with $\lambda = 5$, we have $\forall^{\frac{\delta}{2}} S$:

$$\forall w \in V : p_w > \frac{3\ln\frac{4m}{\delta}}{m}, \qquad \left|p_w - \tfrac{c_w}{m}\right| \leq \sqrt{\tfrac{3p_w\ln\frac{4m}{\delta}}{m}} \qquad (14)$$

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

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

$$|L - L_{leave-one}| \leq \left|\sum_{w\in V_H}\left(p_w L_w - \frac{c_w}{m}L'_w\right)\right| + \left|\sum_{w\in V_L}\left(p_w L_w - \frac{c_w}{m}L'_w\right)\right|. \qquad (16)$$

We start by bounding the first term of Equation (16). By Equation (15), we have $\forall w \in V_H, c_w > y_w + 2 > \ln\frac{4m}{\delta}$. Equation (10) 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

$$\left|\sum_{w\in V_H}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| = \left|\sum_{w\in V_H} Err_w^H + \ln\frac{m-1-\beta}{m-\beta}\sum_{w\in V_H}\frac{c_w}{m}\right| \leq \left|\sum_{w\in V_H} Err_w^H\right| + O\left(\frac{1}{m}\right). \quad (17)$$

We bound $\left|\sum_{w\in V_H} Err_w^H\right|$ using McDiarmid's inequality. As in Lemma 33, 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 33:

$$\left|\sum_{w\in V_H} E\left[Err_w^H\right]\right| \leq \sum_{w\in V_H}\left|E\left[p_w f_w(c_w-\alpha) - \frac{c_w}{m}f_w(c_w-1-\alpha)\right]\right| \leq \sum_{w\in V_H}\frac{3p_w}{m} = O\left(\frac{1}{m}\right). \quad (18)$$

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 $O\left(\frac{\ln m}{m}\right)$, as follows:

The change of $p_u \ln\frac{m-\beta}{\max(c_u-\alpha, y_u)}$ would be 0 if $c_u - \alpha \leq y_u$. Otherwise,

$$\left|p_u\ln\frac{m-\beta}{\max(c_u-1-\alpha,y_u)} - p_u\ln\frac{m-\beta}{\max(c_u-\alpha,y_u)}\right| \leq p_u[\ln(c_u-\alpha) - \ln(c_u-1-\alpha)] = p_u\ln\left(1+\frac{1}{c_u-1-\alpha}\right) = O\left(\frac{p_u}{c_u}\right).$$

Since $c_u \geq y_u = \Omega(mp_u)$, the change is bounded by $O(\frac{p_u}{c_u}) = O(\frac{1}{m})$. The change of $\frac{c_u}{m}\ln\frac{m-\beta}{\max(c_u-1-\alpha,y_u)}$ would be $O(\frac{\ln m}{m})$ if $c_u - 1 - \alpha \leq y_u$. Otherwise,

$$\left|\frac{c_u-1}{m}\ln\frac{m-\beta}{\max(c_u-2-\alpha,y_u)} - \frac{c_u}{m}\ln\frac{m-\beta}{\max(c_u-1-\alpha,y_u)}\right| \leq \frac{c_u-1}{m}\left|\ln\frac{m-\beta}{\max(c_u-2-\alpha,y_u)} - \ln\frac{m-\beta}{\max(c_u-1-\alpha,y_u)}\right| + \frac{1}{m}\ln\frac{m-\beta}{\max(c_u-1-\alpha,y_u)} \leq \frac{c_u-1}{m}\ln\left(1+\frac{1}{c_u-2-\alpha}\right) + \frac{\ln m}{m} = O\left(\frac{\ln m}{m}\right).$$

The change of Err_v^H is bounded in a similar way.

By Equations (17) and (18), and Lemma 2, we have $\forall^{\frac{\delta}{16}} S$:

$$\begin{aligned}
&\left|\sum_{w\in V_H}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| \\
&\leq \left|\sum_{w\in V_H} Err_w^H - E\left[\sum_{w\in V_H} Err_w^H\right]\right| + \left|E\left[\sum_{w\in V_H} Err_w^H\right]\right| + O\left(\frac{1}{m}\right) \\
&\leq O\left(\frac{\ln m}{m}\sqrt{m\ln\frac{1}{\delta}} + \frac{1}{m} + \frac{1}{m}\right) = O\left(\sqrt{\frac{(\ln m)^2 \ln\frac{1}{\delta}}{m}}\right). \qquad (19)
\end{aligned}$$

Next, we bound the second term of Equation (16). By Lemma 10, we have $\forall^{\frac{\delta}{4}} S$:

$$\forall w \in V \; s.t. \;\; p_w \leq \frac{3\ln\frac{4m}{\delta}}{m}, \quad c_w \leq 6\ln\frac{4m}{\delta}. \qquad (20)$$

Let $b = 5\ln\frac{4m}{\delta}$. By Equations (14) and (20), for any w such that $p_w \leq \frac{b}{m}$, we have

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

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 $M_k^L = \sum_{w\in V_L\cap S_k} p_w$. We have

$$\begin{aligned}
&\left|\sum_{w\in V_L}\left(\frac{c_w}{m}L'_w - p_w L_w\right)\right| \\
&= \left|\sum_{k=1}^{2b}\frac{kn_k^L}{m}L'(k) - \sum_{k=0}^{2b-1} M_k^L L(k)\right| \\
&\leq \left|\sum_{k=1}^{2b}\frac{kn_k^L}{m-k+1}L'(k) - \sum_{k=0}^{2b-1} M_k^L L(k)\right| + \sum_{k=1}^{2b} kn_k^L L'(k)\left(\frac{1}{m-k+1} - \frac{1}{m}\right) \\
&= \left|\sum_{k=1}^{2b} G_{k-1}^L L'(k) - \sum_{k=0}^{2b-1} M_k^L L(k)\right| + O\left(\frac{bL_{max}}{m}\right) \\
&= \left|\sum_{k=0}^{2b-1} G_k^L L'(k+1) - \sum_{k=0}^{2b-1} M_k^L L(k)\right| + O\left(\frac{bL_{max}}{m}\right) \\
&\leq \sum_{k=0}^{2b-1} G_k^L |L'(k+1) - L(k)| + \sum_{k=0}^{2b-1} |G_k^L - M_k^L| L(k) + O\left(\frac{bL_{max}}{m}\right). \qquad (21)
\end{aligned}$$

The first sum of Equation (21) is bounded using Equations (11) and (12), and Lemma 34:

$$\sum_{k=0}^{2b-1} G_k^L |L'(k+1) - L(k)|$$
$$= \sum_{k=2}^{2b-1} G_k^L |L'(k+1) - L(k)| + G_0 |L'(1) - L(0)| + G_1 |L'(2) - L(1)|. \quad (22)$$

The first term of Equation (22) is bounded by Equation (11):

$$\sum_{k=2}^{2b-1} G_k^L |L'(k+1) - L(k)| \leq \sum_{k=2}^{2b-1} G_k^L \cdot O\left(\frac{1}{m}\right) = O\left(\frac{1}{m}\right). \quad (23)$$

The other two terms are bounded using Lemma 34. For $n_1 > 0$, we have $\forall^{\frac{\delta}{16}} S$, $n_2 = O\left(b\left(\sqrt{m \ln \frac{1}{\delta}} + n_1\right)\right)$. By Equation (12), we have

$$G_0 |L'(1) - L(0)| + G_1 |L'(2) - L(1)|$$
$$\leq \frac{n_1}{m} \cdot O\left(\frac{1}{n_1}\right) + \frac{2n_2}{m-1} \cdot O\left(\frac{1}{n_1}\right) = O\left(b\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right). \quad (24)$$

For $n_1 = 0$, Lemma 34 results in $n_2 = O\left(b\sqrt{m \ln \frac{1}{\delta}}\right)$, and Equation (24) transforms into

$$G_1 |L'(2) - L(1)| \leq \frac{2n_2 L_{max}}{m-1} = O\left(bL_{max}\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right). \quad (25)$$

Equations (22), (23), (24), and (25) sum up to

$$\sum_{k=0}^{2b-1} G_k^L |L'(k+1) - L(k)| = O\left(bL_{max}\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right). \quad (26)$$

The second sum of Equation (21) is bounded using Lemma 28 separately for every $k < 2b$ with accuracy $\frac{\delta}{16b}$. Since the proof of Lemma 28 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, together with Equations (21) and (26), we have

$$\left|\sum_{w \in V_L} \left(\frac{c_w}{m} L'_w - p_w L_w\right)\right| \leq O\left(bL_{max}\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right) + \sum_{k=0}^{2b-1} L(k) O\left(b\sqrt{\frac{\ln \frac{b}{\delta}}{m}}\right) + O\left(\frac{bL_{max}}{m}\right)$$
$$= O\left(L_{max}\sqrt{\frac{b^4 \ln \frac{b}{\delta}}{m}}\right). \quad (27)$$

The proof follows by combining Equations (16), (19), and (27). ■

5. Log-Loss A Priori

Section 4 bounds the error of the leave-one-out estimation of the log-loss. It shows that the log-loss can be effectively estimated, for a general family of learning algorithms.

Another question to be considered is the log-loss distribution itself, without the empirical estimation. That is, how large (or low) is it expected to be, and which parameters of the distribution affect it.

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 lower and upper bounds for the learning error.

The upper bound (Theorem 39) can be divided to three parts. The first part is the missing mass. The other two build a trade-off between a threshold (lower thresholds leads to a lower bound), and the number of words with probability exceeding this threshold (fewer words lead to a lower bound). It seems that this number of words is a necessary lower bound, as we show at Theorem 35.

Theorem 35 *Let the distribution be uniform: $\forall w \in V : p_w = \frac{1}{N}$, with $N \ll m$. Also, suppose that the learning algorithm just uses maximum-likelihood approximation, meaning $q_w = \frac{c_w}{m}$. Then, a typical learning error would be $\Omega(\frac{N}{m})$.*

The proof of Theorem 35 bases on the Pinsker inequality (Lemma 36). It first shows a lower bound for L_1 norm between the true and the expected distributions, and then transforms it to the form of the learning error.

Lemma 36 *(Pinsker Inequality) Given any two distributions P and Q, we have*

$$KL(P||Q) \geq \frac{1}{2}(L_1(P,Q))^2.$$

Theorem 35 Proof We first show that $L_1(P,Q)$ concentrates near $\Omega\left(\sqrt{\frac{N}{m}}\right)$. Then, we use Pinsker inequality to show lower bound[5] of $KL(P||Q)$.

First we find a lower bound for $E[|p_w - q_w|]$. Since c_w is a binomial random variable, $\sigma^2[c_w] = mp_w(1-p_w) = \Omega\left(\frac{m}{N}\right)$, and with some constant probability, $|c_w - mp_w| > \sigma[c_w]$. Therefore, we have

$$\begin{aligned} E[|q_w - p_w|] &= \frac{1}{m}E[|c_w - mp_w|] \\ &\geq \frac{1}{m}\sigma[c_w]P(|c_w - mp_w| > \sigma[c_w]) = \Omega\left(\frac{1}{m}\sqrt{\frac{m}{N}}\right) = \Omega\left(\frac{1}{\sqrt{mN}}\right) \end{aligned}$$

$$E\left[\sum_{w\in V}|p_w - q_w|\right] = \Omega\left(N\frac{1}{\sqrt{mN}}\right) = \Omega\left(\sqrt{\frac{N}{m}}\right).$$

5. This proof does not optimize the constants. Asymptotic analysis of logarithmic transform of binomial variables by Flajolet (1999) can be used to achieve explicit values for $KL(P||Q)$.

A single change in the sample changes $L_1(P,Q)$ by at most $\frac{2}{m}$. Using McDiarmid inequality (Lemma 2) on $L_1(P,Q)$ as a function of sample words, we have $\forall^{\frac{1}{2}} S$:

$$\begin{aligned} L_1(P,Q) &\geq E[L_1(P,Q)] - |L_1(P,Q) - E[L_1(P,Q)]| \\ &= \Omega\left(\sqrt{\frac{N}{m}}\right) - O\left(\frac{\sqrt{m}}{m}\right) = \Omega\left(\sqrt{\frac{N}{m}}\right). \end{aligned}$$

Using Pinsker inequality (Lemma 36), we have

$$\forall^{\frac{1}{2}} S, \qquad \sum_{w\in V} p_w \ln \frac{p_w}{q_w} \geq \frac{1}{2}\left(\sum_{w\in V} |p_w - q_w|\right)^2 = \Omega\left(\frac{N}{m}\right),$$

which completes the proof. ■

Definition 37 *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 :*

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

The α parameter can be set to some constant, or to make the missing mass match the Good-Turing missing mass estimator, that is $\frac{\alpha n_{1\ldots\tau}}{m} = \frac{n_1}{m}$.

Definition 38 *Given a distribution P, and $x \in [0,1]$, let $F_x = \sum_{w\in V: p_w \leq x} p_w$, and $N_x = |\{w \in V : p_w > x\}|$. Clearly, for any distribution P, F_x is a monotone function of x, varying from 0 to 1, and N_x is a monotone function of x, varying from N to 0. Note that N_x is bounded by $\frac{1}{x}$.*

The next theorem shows an upper bound for the learning error.

Theorem 39 *For any $\delta > 0$ and $\lambda > 3$, such that $\tau < (\lambda - \sqrt{3\lambda}) \ln \frac{8m}{\delta}$, the learning error of $A_{\alpha,\tau}$ is bounded $\forall^{\delta} S$ by*

$$\begin{aligned} 0 \leq \sum_{w\in V} p_w \ln\left(\frac{p_w}{q_w}\right) &\leq M_0 \ln\left(\frac{n_0 \ln \frac{4m}{\delta}}{\alpha n_{1\ldots\tau}}\right) + \frac{\lambda \ln \frac{8m}{\delta}}{1-\alpha}\left(\sqrt{\frac{3\ln\frac{8}{\delta}}{m}} + M_0\right) \\ &\quad + \frac{\alpha}{1-\alpha} F_{\frac{\lambda \ln \frac{8m}{\delta}}{m}} + \sqrt{\frac{3\ln\frac{8}{\delta}}{m}} + \frac{3\lambda \ln \frac{8m}{\delta}}{2(\sqrt{\lambda}-\sqrt{3})^2 m} N_{\frac{\lambda \ln \frac{8m}{\delta}}{m}}. \end{aligned}$$

The proof of Theorem 39 bases directly on Lemmas 40, 41, and 43. We can rewrite this bound roughly as

$$\sum_{w\in V} p_w \ln\left(\frac{p_w}{q_w}\right) \leq \tilde{O}\left(M_0 + \frac{\lambda}{\sqrt{m}} + \frac{N_{\frac{\lambda}{m}}}{m}\right).$$

This bound implies the characteristics of the distribution influencing the log-loss. It shows that a "good" distribution can involve many low-probability words, given that the missing mass is low. However, the learning error would increase if the dictionary included many mid-range-probability words. For example, if a typical word's probability were $m^{-\frac{3}{4}}$, the bound would become $\tilde{O}\left(M_0 + m^{-\frac{1}{4}}\right)$.

Lemma 40 *For any $\delta > 0$, the learning error for non-appearing words can be bounded with high probability by*

$$\forall^{\delta} S, \qquad \sum_{w\notin S} p_w \ln\left(\frac{p_w}{q_w}\right) \leq M_0 \ln\left(\frac{n_0 \ln\frac{m}{\delta}}{\alpha n_{1\ldots\tau}}\right).$$

Proof By Lemma 13, we have $\forall^{\delta} S$, the real probability of any non-appearing word does not exceed $\frac{\ln\frac{m}{\delta}}{m}$. Therefore,

$$\begin{aligned}\sum_{w\notin S} p_w \ln\left(\frac{p_w}{q_w}\right) &= \sum_{w\notin S} p_w \ln\left(p_w \frac{m}{\alpha}\frac{n_0}{n_{1\ldots\tau}}\right) \\ &\leq \sum_{w\notin S} p_w \ln\left(\frac{\ln\frac{m}{\delta}}{m}\frac{m}{\alpha}\frac{n_0}{n_{1\ldots\tau}}\right) = M_0 \ln\left(\frac{n_0 \ln\frac{m}{\delta}}{\alpha n_{1\ldots\tau}}\right),\end{aligned}$$

which completes the proof. ■

Lemma 41 *Let $\delta > 0$, $\lambda > 0$. Let $V_L = \left\{w \in V : p_w \leq \frac{\lambda \ln\frac{2m}{\delta}}{m}\right\}$, and $V'_L = V_L \cap S$. The learning error for V'_L can be bounded with high probability by*

$$\forall^{\delta} S, \qquad \sum_{w\in V'_L} p_w \ln\left(\frac{p_w}{q_w}\right) \leq \frac{\lambda \ln\frac{2m}{\delta}}{1-\alpha}\left(\sqrt{\frac{3\ln\frac{2}{\delta}}{m}} + M_0\right) + \frac{\alpha}{1-\alpha} F_{\frac{\lambda \ln\frac{2m}{\delta}}{m}}.$$

Proof We use $\ln(1+x) \leq x$.

$$\sum_{w\in V'_L} p_w \ln\frac{p_w}{q_w} \leq \sum_{w\in V'_L} p_w \frac{p_w - q_w}{q_w}.$$

For any appearing word w, $q_w \geq \frac{1-\alpha}{m}$. Therefore,

$$\begin{aligned}
\sum_{w\in V'_L} p_w \frac{p_w - q_w}{q_w} &\leq \frac{m}{1-\alpha} \sum_{w\in V'_L} p_w(p_w - q_w) \\
&= \frac{m}{1-\alpha}\left[\sum_{w\in V'_L} p_w\left(p_w - \frac{c_w}{m}\right) + \sum_{w\in V'_L} p_w\left(\frac{c_w}{m} - q_w\right)\right] \\
&\leq \frac{m}{1-\alpha}\left|\sum_{w\in V'_L} p_w\left(p_w - \frac{c_w}{m}\right)\right| + \frac{m}{1-\alpha}\sum_{w\in V'_L} p_w\frac{\alpha}{m} \\
&\leq \frac{m}{1-\alpha}\frac{\lambda\ln\frac{2m}{\delta}}{m}\left|\sum_{w\in V'_L}\left(p_w - \frac{c_w}{m}\right)\right| + \frac{\alpha}{1-\alpha}\sum_{w\in V'_L} p_w \\
&\leq \frac{\lambda\ln\frac{2m}{\delta}}{1-\alpha}\left|\sum_{w\in V'_L}\left(p_w - \frac{c_w}{m}\right)\right| + \frac{\alpha}{1-\alpha}F_{\frac{\lambda\ln\frac{2m}{\delta}}{m}}. \qquad (28)
\end{aligned}$$

We apply Lemma 12 on v_L, the union of words in V_L. Let $p_{v_L} = \sum_{w\in V_L} p_w$ and $c_{v_L} = \sum_{w\in V_L} c_w$. We have $\forall^\delta S$:

$$\begin{aligned}
\left|\sum_{w\in V'_L}\left(p_w - \frac{c_w}{m}\right)\right| &= \left|\sum_{w\in V_L}\left(p_w - \frac{c_w}{m}\right) - \sum_{w\in V_L\setminus S}\left(p_w - \frac{c_w}{m}\right)\right| \\
&\leq \left|\sum_{w\in V_L}\left(p_w - \frac{c_w}{m}\right)\right| + \sum_{w\in V_L\setminus S} p_w \\
&\leq \left|p_{v_L} - \frac{c_{v_L}}{m}\right| + M_0 \\
&\leq \sqrt{\frac{3\ln\frac{2}{\delta}}{m}} + M_0. \qquad (29)
\end{aligned}$$

The proof follows combining Equations (28) and (29). ■

Lemma 42 *Let $0 < \Delta < 1$. For any $x \in [-\Delta, \Delta]$, we have $\ln(1+x) \geq x - \frac{x^2}{2(1-\Delta)^2}$.*

Lemma 43 *Let $\delta > 0$, $\lambda > 3$, such that $\tau < (\lambda - \sqrt{3\lambda})\ln\frac{4m}{\delta}$. Let the high-probability words set be $V_H = \left\{w \in V : p_w > \frac{\lambda\ln\frac{4m}{\delta}}{m}\right\}$, and $V'_H = V_H \cap S$. The learning error for V'_H can be bounded with high probability by*

$$\forall^\delta S, \qquad \sum_{w\in V'_H} p_w \ln\left(\frac{p_w}{q_w}\right) \leq \sqrt{\frac{3\ln\frac{4}{\delta}}{m}} + \frac{3\lambda\ln\frac{4m}{\delta}}{2(\sqrt{\lambda}-\sqrt{3})^2 m} N_{\frac{\lambda\ln\frac{4m}{\delta}}{m}}.$$

Proof

$$\begin{aligned}\sum_{w\in V'_H} p_w \ln\left(\frac{p_w}{q_w}\right) &= \sum_{w\in V'_H} p_w \ln\left(\frac{p_w}{\frac{c_w}{m}}\right) + \sum_{w\in V'_H} p_w \ln\left(\frac{\frac{c_w}{m}}{q_w}\right) \\ &= \sum_{w\in V'_H} p_w \ln\left(\frac{mp_w}{c_w}\right) + \sum_{w\in V'_H, c_w\le\tau} p_w \ln\left(\frac{c_w}{c_w-\alpha}\right). \end{aligned} \tag{30}$$

Using Lemma 9 with λ, we have $\forall^{\frac{\delta}{2}} S$:

$$\begin{aligned} &\forall w\in V_H, && \left|p_w - \frac{c_w}{m}\right| \le \sqrt{\frac{3p_w \ln\frac{4m}{\delta}}{m}}, \\ &\forall w\in V_H, && c_w \ge (\lambda - \sqrt{3\lambda})\ln\frac{4m}{\delta}. \end{aligned} \tag{31}$$

This means that for a reasonable choice of τ (meaning $\tau < (\lambda-\sqrt{3\lambda})\ln\frac{4m}{\delta}$), the second term of Equation (30) is 0, and $V'_H = V_H$. Also,

$$\left|\frac{\frac{c_w}{m}-p_w}{p_w}\right| \le \frac{1}{p_w}\sqrt{\frac{3p_w\ln\frac{4m}{\delta}}{m}} \le \sqrt{\frac{m}{\lambda\ln\frac{2m}{\delta}}\frac{3\ln\frac{2m}{\delta}}{m}} = \sqrt{\frac{3}{\lambda}}.$$

Therefore, we can use Lemma 42 with $\Delta = \sqrt{\frac{3}{\lambda}}$:

$$\begin{aligned}\sum_{w\in V'_H} p_w \ln\left(\frac{mp_w}{c_w}\right) &= -\sum_{w\in V_H} p_w \ln\left(1+\frac{\frac{c_w}{m}-p_w}{p_w}\right) \\ &\le -\sum_{w\in V_H} p_w \left[\frac{\frac{c_w}{m}-p_w}{p_w} - \frac{1}{2\left(1-\sqrt{\frac{3}{\lambda}}\right)^2}\left(\frac{\frac{c_w}{m}-p_w}{p_w}\right)^2\right] \\ &= \sum_{w\in V_H}\left(p_w - \frac{c_w}{m}\right) + \frac{\lambda}{2\left(\sqrt{\lambda}-\sqrt{3}\right)^2}\sum_{w\in V_H}\frac{\left(\frac{c_w}{m}-p_w\right)^2}{p_w}. \end{aligned} \tag{32}$$

We apply Lemma 12 on the v_H, the union of all words in V_H. Let $p_{v_H} = \sum_{w\in V_H} p_w$ and $c_{v_H} = \sum_{w\in V_H} c_w$. The bound on the first term of Equation (32) is:

$$\forall^{\frac{\delta}{2}} S, \qquad \left|\sum_{w\in V_H}\left(p_w - \frac{c_w}{m}\right)\right| = \left|p_{v_H} - \frac{c_{v_H}}{m}\right| \le \sqrt{\frac{3\ln\frac{4}{\delta}}{m}}. \tag{33}$$

Assuming that Equation (31) holds, the second term of Equation (32) can also be bounded:

$$\sum_{w\in V_H}\frac{\left(\frac{c_w}{m}-p_w\right)^2}{p_w}\leq\sum_{w\in V_H}\frac{1}{p_w}\frac{3p_w\ln\frac{4m}{\delta}}{m}=\frac{3\ln\frac{4m}{\delta}}{m}N_{\lambda\ln\frac{4m}{\delta}}\,. \tag{34}$$

The proof follows by combining Equations (30), (32), (33) and (34). ■

Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, by a grant from the Israel Science Foundation and an IBM faculty award. This publication only reflects the authors' views.

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

Appendix A. Technical Proofs

A.1 Concentration Inequalities

Lemma 6 **Proof** We use Stirling approximation $\Gamma(x+1)=\sqrt{2\pi x}\left(\frac{x}{e}\right)^x T_x$, where

$$T_x=\exp\left(\frac{1}{12x}+O\left(\frac{1}{x^2}\right)\right).$$

$$\begin{aligned}
P(X=k) &= \binom{n}{k}p^k(1-p)^{n-k}\\
&\leq \frac{\Gamma(n+1)}{\Gamma(\mu+1)\Gamma(n-\mu+1)}\left(\frac{\mu}{n}\right)^{\mu}\left(\frac{n-\mu}{n}\right)^{n-\mu}\\
&= \frac{\sqrt{2\pi n}}{\sqrt{2\pi\mu}\sqrt{2\pi(n-\mu)}}\frac{n^n}{\mu^\mu(n-\mu)^{n-\mu}}\frac{\mu^\mu}{n^\mu}\frac{(n-\mu)^{n-\mu}}{n^{n-\mu}}\frac{T_n}{T_\mu T_{n-\mu}}\\
&= \frac{1}{\sqrt{2\pi\mu}}\sqrt{\frac{n}{n-\mu}}\frac{T_n}{T_\mu T_{n-\mu}}\\
&= \frac{1}{\sqrt{2\pi\mu(1-p)}}\frac{T_n}{T_\mu T_{n-\mu}}.
\end{aligned}$$

Clearly, for integral values of μ, the equality is achieved at $k=\mu$. ■

Lemma 8 **Proof** Let $m'=\sum_{w\in V}c'_w$. Using Lemma 7 for m' with $b=c=E[m']=m$, the probability $P(m'=m)$ achieves its minimum when $\forall w\in V, p_w=\frac{1}{N}$. Under this assumption, we have $m'\sim Bin(mN,\frac{1}{N})$. Using Lemma 6, we have

$$P\left(m' = m\right) = \frac{1}{\sqrt{2\pi m N \frac{1}{N}\left(1-\frac{1}{N}\right)}} \frac{T_{mN}}{T_m T_{mN-m}} \geq \frac{1}{3\sqrt{m}}.$$

Therefore, for any distribution $\{p_w : w \in V\}$, we have

$$P(m' = m) \geq \frac{1}{3\sqrt{m}}.$$

Obviously, $E[F'] = \sum_w E[f_w(c'_w)] = E[F]$. Also, the distribution of $\{c'_w\}$ given that $m' = m$ is identical to the distribution of $\{c_w\}$, therefore the distribution of F' given that $m' = m$ is identical to the distribution of F. We have

$$\begin{aligned} P(|F' - E[F']| > \varepsilon) &= \sum_i P(m' = i) P(|F' - E[F']| > \varepsilon | m' = i) \\ &\geq P(m' = m) P(|F' - E[F']| > \varepsilon | m' = m) \\ &= P(m' = m) P(|F - E[F]| > \varepsilon) \\ &\geq \frac{1}{3\sqrt{m}} P(|F - E[F]| > \varepsilon), \end{aligned}$$

which completes the proof. ■

Lemma 44 *For any $\delta > 0$, and a word $w \in V$, such that $p_w \geq \frac{3\ln\frac{2}{\delta}}{m}$, we have*

$$P\left(\left|p_w - \frac{c_w}{m}\right| > \sqrt{\frac{3p_w \ln\frac{2}{\delta}}{m}}\right) \leq \delta.$$

Proof The proof follows by applying Lemma 3, substituting $\varepsilon = \sqrt{3mp_w \ln\frac{2}{\delta}}$. Note that for $3\ln\frac{2}{\delta} \leq mp_w$ we have $\varepsilon \leq mp_w$:

$$\begin{aligned} P\left(\left|p_w - \frac{c_w}{m}\right| \geq \sqrt{\frac{3p_w \ln\frac{2}{\delta}}{m}}\right) &= P\left(|mp_w - c_w| \geq \varepsilon\right) \\ &\leq 2\exp\left(-\frac{\varepsilon^2}{2E[c_w] + \varepsilon}\right) \\ &\leq 2\exp\left(-\frac{3mp_w \ln\frac{2}{\delta}}{3mp_w}\right) = \delta, \end{aligned}$$

which completes the proof. ■

Lemma 9 **Proof** There are at most m words with probability $p_w \geq \frac{3\ln\frac{2m}{\delta}}{m}$. The first claim follows using Lemma 44 together with union bound over all these words (with accuracy $\frac{\delta}{m}$ for each word).

Using the first claim, we derive the second. We show a lower bound for $\frac{c_w}{m}$, using $\frac{\ln\frac{2m}{\delta}}{m} < \frac{1}{\lambda}p_w$:

$$\frac{c_w}{m} \geq p_w - \sqrt{\frac{3p_w \ln\frac{2m}{\delta}}{m}} > p_w - p_w\sqrt{\frac{3}{\lambda}} = \left(1-\sqrt{\frac{3}{\lambda}}\right)p_w.$$

The final inequality follows from simple algebra. ■

Lemma 10 **Proof** Let $b = 3\ln(\frac{m}{\delta})$. Note that $\delta \in [0,1]$ and $m > 1$ yield $b > 2$. First, suppose that there are up to m words with $p_w \leq \frac{b}{m}$. For each such word, we apply Lemma 3 on c_w, with $\varepsilon = b$. We have:

$$P\left(c_w > 6\ln\frac{m}{\delta}\right) \leq P\left(c_w > mp_w + \varepsilon\right) \leq \exp\left(-\frac{b^2}{2mp_w + b}\right) \leq \frac{\delta}{m}.$$

Since we assume that there are up to m such words, the total mistake probability is δ.

Now we assume the general case, that is, without any assumption on the number of words. Our goal is to reduce the problem to the former conditions, that is, to create a set of size m of words with probability smaller than $\frac{b}{m}$.

We first create m empty sets $v_1, \ldots, v_m$. Let the probability of each set v_i, p_{v_i}, be the sum of the probabilities of all the words it includes. Let the actual count of v_i, c_{v_i}, be the sum of the sample counts of all the words w it includes.

We divide all the words w between these sets in a bin-packing-approximation manner. We sort the words w in decreasing probability order. Then, we do the following loop: insert the next word w to the set v_i with the currently smallest p_{v_i}.

We claim that $p_{v_i} \leq \frac{b}{m}$ for each v_i at the end of the loop. If this inequality does not hold, then some word w made this "overflow" first. Obviously, p_w must be smaller than $\frac{b}{2m}$, otherwise it would be one of the first $\frac{2m}{b} < m$ words ordered, and would enter an empty set. If $p_w < \frac{b}{2m}$ and it made an "overflow", then the probability of each set at the moment w was entered must exceed $\frac{b}{2m}$, since w must have entered the lightest set available. This means that the total probability of all words entered by that moment was greater than $m\frac{b}{2m} > 1$.

Applying the case of m words to the sets $v_1, \ldots, v_m$, we have $\forall^{\delta} S$: for every v_i, $c_{v_i} \leq 2b$. Also, if the count of each set v_i does not exceed $2b$, so does the count of each word $w \in v_i$. That is,

$$P\left(\exists w : p_w \leq \frac{b}{m}, c_w > 2b\right) \leq P\left(\exists v_i : p_{v_i} \leq \frac{b}{m}, c_{v_i} > 2b\right) \leq \delta,$$

which completes the proof. ■

Lemma 11 **Proof** By Lemma 9 with some $\lambda > 3$ (which will be set later), we have $\forall^{\frac{\delta}{2}} S$:

$$\forall w : p_w \geq \frac{3\ln\frac{4m}{\delta}}{m}, \qquad |mp_w - c_w| \leq \sqrt{3mp_w \ln\frac{4m}{\delta}}, \tag{35}$$

$$\forall w : p_w > \frac{\lambda\ln\frac{4m}{\delta}}{m}, \qquad c_w > \left(1 - \sqrt{\frac{3}{\lambda}}\right) mp_w. \tag{36}$$

By Equation (35), for any word w such that $\frac{3\ln\frac{4m}{\delta}}{m} \leq p_w \leq \frac{\lambda\ln\frac{4m}{\delta}}{m}$, we have $c_w \leq mp_w + \sqrt{3mp_w \ln\frac{4m}{\delta}} \leq \left(\lambda + \sqrt{3\lambda}\right)\ln\frac{4m}{\delta}$. By Lemma 10, we have

$$\forall^{\frac{\delta}{2}} S, \quad \forall w \; s.t. \; p_w \leq \frac{3\ln\frac{4m}{\delta}}{m}, \quad c_w \leq 6\ln\frac{4m}{\delta}.$$

It means that for any $w : mp_w \leq \lambda\ln\frac{4m}{\delta}$, we have $c_w \leq \left(\lambda + \sqrt{3\lambda}\right)\ln\frac{4m}{\delta}$. This means that for any w such that $c_w > \left(\lambda + \sqrt{3\lambda}\right)\ln\frac{4m}{\delta}$, we have $mp_w > \lambda\ln\frac{4m}{\delta}$. By Equation (36), this means $mp_w \leq \frac{1}{1-\sqrt{\frac{3}{\lambda}}} c_w$, and by Equation (35):

$$|mp_w - c_w| \leq \sqrt{3mp_w \ln\frac{4m}{\delta}} \leq \sqrt{\frac{3c_w \ln\frac{4m}{\delta}}{1-\sqrt{\frac{3}{\lambda}}}} = \sqrt{\frac{3c_w\sqrt{\lambda}\ln\frac{4m}{\delta}}{\sqrt{\lambda}-\sqrt{3}}}.$$

Substituting $\lambda = 12$ results in

$$\forall^{\delta} S : \quad \forall w \; s.t. \; c_w > 18\ln\frac{4m}{\delta}, \quad |mp_w - c_w| \leq \sqrt{6c_w \ln\frac{4m}{\delta}},$$

which completes the proof. ■

Lemma 12 **Proof** If $p_w \geq \frac{3\ln\frac{2}{\delta}}{m}$, we can apply Lemma 44. We have

$$\forall^{\delta} S, \qquad \left|\frac{c_w}{m} - p_w\right| \leq \sqrt{\frac{3p_w \ln\frac{2}{\delta}}{m}} \leq \sqrt{\frac{3\ln\frac{2}{\delta}}{m}}.$$

Otherwise, we can apply Lemma 10. We have:

$$\forall^{\delta} S, \qquad \left|\frac{c_w}{m} - p_w\right| \leq \max\left\{p_w, \frac{c_w}{m}\right\} \leq \frac{6\ln\frac{m}{\delta}}{m} \leq \sqrt{\frac{3\ln\frac{2}{\delta}}{m}},$$

which completes the proof. ■

Lemma 13 **Proof** Let $b = \ln \frac{m}{\delta}$. We note that there are at most $\frac{m}{b}$ words with probability $p_w \geq \frac{b}{m}$.

$$\begin{aligned} P\left(\exists w : c_w = 0, p_w \geq \frac{b}{m}\right) &\leq \sum_{w: p_w \geq \frac{b}{m}} P(c_w = 0) \\ &= \sum_{w: p_w \geq \frac{b}{m}} (1-p_w)^m \leq \frac{m}{b}\left(1-\frac{b}{m}\right)^m < me^{-b} = \delta, \end{aligned}$$

which completes the proof. ■

A.2 K-Hitting Mass Estimation

Lemma 21 **Proof** We have $\sum_{w \in V_{k,\alpha}} p_w \leq 1$. Using Lemma 19, we bound $P(c_w = k)$ and $P(c_w = k+1)$:

$$\begin{aligned} E[M_{k,\alpha}] &= \sum_{w \in V_{k,\alpha}} p_w P(c_w = k) = O\left(\frac{1}{\sqrt{k}}\right) \\ |E[G_{k,\alpha}] - E[M_{k,\alpha}]| &= \left| \sum_{w \in V_{k,\alpha}} \left[\frac{k+1}{m-k} P(c_w = k+1) - p_w P(c_w = k) \right] \right| \\ &= \sum_{w \in V_{k,\alpha}} p_w \frac{k+1}{m-k} P(c_w = k+1) = O\left(\frac{\sqrt{k}}{m}\right). \end{aligned} \tag{37}$$

Equation (37) follows by Lemma 20. By Lemma 18, we have $|V_{k,\alpha}| = O\left(\frac{m}{k}\right)$:

$$E[G_{k,\alpha}] = \frac{k+1}{m-k} \sum_{w \in V_{k,\alpha}} P(c_w = k+1) = O\left(\frac{k}{m}\frac{m}{k}\frac{1}{\sqrt{k}}\right) = O\left(\frac{1}{\sqrt{k}}\right),$$

which completes the proof. ■

Theorem 29 **Proof** The proof is done by examining four cases of k. For $k \leq 18 \ln \frac{8m}{\delta}$, we can use Lemma 28. We have

$$\forall^\delta S, \quad |\tilde{M}_k - M_k| = |G_k - M_k| = O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\left(k + \ln \frac{m}{\delta}\right)\right) = \tilde{O}\left(\frac{1}{\sqrt{m}}\right).$$

For $18 \ln \frac{8m}{\delta} < k \leq m^{\frac{2}{5}}$, we can use Theorem 25. We have

$$\forall^\delta S, \quad |\tilde{M}_k - M_k| = |G_k - M_k| = O\left(\sqrt{\frac{\sqrt{k}\ln \frac{m}{\delta}}{m}} + \frac{k \ln \frac{m}{\delta}}{m}\right) = \tilde{O}\left(m^{-\frac{2}{5}}\right).$$

For $m^{\frac{2}{5}} < k < \frac{m}{2}$, we can use Theorem 26. We have

$$\forall^{\delta} S, \quad |\tilde{M}_k - M_k| = |\hat{M}_k - M_k| = O\left(\frac{\sqrt{k}\left(\ln\frac{m}{\delta}\right)^{\frac{3}{2}}}{m} + \frac{\sqrt{\ln\frac{m}{\delta}}}{k}\right) = \tilde{O}\left(m^{-\frac{2}{5}}\right).$$

For $k \geq \frac{m}{2}$, let $\alpha = \sqrt{6\ln\frac{8m}{\delta}}$. By Lemma 17, we have $\forall^{\frac{\delta}{2}} S$, $M_k = M_{k,\alpha} \wedge \hat{M}_k = \hat{M}_{k,\alpha}$. By Lemma 18, $|V_{k,\alpha}| = O\left(\frac{m}{k}\right) = O(1)$. Let c be the bound on $|V_{k,\alpha}|$. Using Lemma 12 for each $w \in V_{k,\alpha}$ with accuracy $\frac{\delta}{2c}$, we have

$$\forall^{\frac{\delta}{2}} S, \quad \forall w \in V_{k,\alpha}, \quad \left|\frac{c_w}{m} - p_w\right| = O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right).$$

Therefore, we have $\forall^{\delta} S$:

$$|\tilde{M}_k - M_k| = |\hat{M}_{k,\alpha} - M_{k,\alpha}| \leq \sum_{w \in V_{k,\alpha}} \left|\frac{k}{m} - p_w\right| X_{w,k} = O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) = \tilde{O}\left(\frac{1}{\sqrt{m}}\right),$$

which completes the proof. ■

Theorem 30 **Proof** First, we show that for any two words u and v, $Cov(X_{u,k}, X_{v,k}) = \Theta\left(\frac{k}{m^2}\right)$. Note that $\{c_v | c_u = k\} \sim Bin\left(m-k, \frac{k}{m-k}\right)$. By Lemma 6, we have:

$$\begin{aligned} P(c_u = k) = P(c_v = k) &= \frac{1}{\sqrt{2\pi k\left(1-\frac{k}{m}\right)}} \frac{T_m}{T_k T_{m-k}}, \\ P(c_v = k | c_u = k) &= \frac{1}{\sqrt{2\pi k\left(1-\frac{k}{m-k}\right)}} \frac{T_{m-k}}{T_k T_{m-2k}}. \end{aligned} \tag{38}$$

Using $T_x = \Theta(1)$ for $x \geq k$, we have

$$\begin{aligned} &Cov(X_{u,k}, X_{v,k}) \\ &= E[X_{u,k}X_{v,k}] - E[X_{u,k}]E[X_{v,k}] \\ &= P(c_u = k)[P(c_v = k | c_u = k) - P(c_v = k)] \\ &= \frac{1}{2\pi k\sqrt{\left(1-\frac{k}{m}\right)}} \frac{T_m}{T_k T_{m-k}} \left[\frac{1}{\sqrt{\left(1-\frac{k}{m-k}\right)}} \frac{T_{m-k}}{T_k T_{m-2k}} - \frac{1}{\sqrt{\left(1-\frac{k}{m}\right)}} \frac{T_m}{T_k T_{m-k}}\right] \\ &= \Theta\left(\frac{1}{k}\right) \left[\frac{1}{\sqrt{\left(1-\frac{k}{m-k}\right)}} \frac{T_{m-k}}{T_{m-2k}} - \frac{1}{\sqrt{\left(1-\frac{k}{m}\right)}} \frac{T_m}{T_{m-k}}\right] \end{aligned}$$

$$= \Theta\left(\frac{1}{k}\right)\left[\frac{T_{m-k}}{T_{m-2k}}\left(\frac{1}{\sqrt{\left(1-\frac{k}{m-k}\right)}}-\frac{1}{\sqrt{\left(1-\frac{k}{m}\right)}}\right)\right.$$
$$\left.+\frac{1}{\sqrt{\left(1-\frac{k}{m}\right)}}\left(\frac{T_{m-k}}{T_{m-2k}}-\frac{T_m}{T_{m-k}}\right)\right]. \quad (39)$$

We can bound the first term of Equation (39):

$$\frac{1}{\sqrt{\left(1-\frac{k}{m-k}\right)}}-\frac{1}{\sqrt{\left(1-\frac{k}{m}\right)}} = \left(\frac{\sqrt{1-\frac{k}{m}}-\sqrt{1-\frac{k}{m-k}}}{\sqrt{\left(1-\frac{k}{m-k}\right)\left(1-\frac{k}{m}\right)}}\right)\left(\frac{\sqrt{1-\frac{k}{m}}+\sqrt{1-\frac{k}{m-k}}}{\sqrt{1-\frac{k}{m}}+\sqrt{1-\frac{k}{m-k}}}\right)$$
$$= \Theta\left(1-\frac{k}{m}-1+\frac{k}{m-k}\right)=\Theta\left(\frac{k^2}{m^2}\right). \quad (40)$$

Since $T_x = \exp\left(\frac{1}{12x}+O\left(\frac{1}{x^2}\right)\right) = 1+\frac{1}{12x}+O\left(\frac{1}{x^2}\right)$ for $x \geq m-2k$ (note that $k \ll m$), we have

$$\frac{T_{m-k}}{T_{m-2k}}-\frac{T_m}{T_{m-k}} = \frac{T_{m-k}^2-T_mT_{m-2k}}{T_{m-2k}T_{m-k}}$$
$$= \frac{1}{T_{m-2k}T_{m-k}}\left[\frac{1}{6(m-k)}-\frac{1}{12m}-\frac{1}{12(m-2k)}+O\left(\frac{1}{m^2}\right)\right]$$
$$= -\Theta\left(\frac{k^2}{m^3}\right)+O\left(\frac{1}{m^2}\right). \quad (41)$$

Combining Equations (39), (40), and (41), we have

$$Cov(X_{u,k},X_{v,k}) = \Theta\left(\frac{1}{k}\right)\left[\Theta\left(\frac{k^2}{m^2}\right)-\Theta\left(\frac{k^2}{m^3}\right)+O\left(\frac{1}{m^2}\right)\right]=\Theta\left(\frac{k}{m^2}\right).$$

Now we show that $\sigma^2[X_{w,k}] = \Theta\left(\frac{1}{\sqrt{k}}\right)$. By Equation (38), we have

$$\sigma^2[X_{w,k}] = P(c_w=k)(1-P(c_w=k)) = \Theta\left(\frac{1}{\sqrt{k}}\right)\left(1-\Theta\left(\frac{1}{\sqrt{k}}\right)\right)=\Theta\left(\frac{1}{\sqrt{k}}\right).$$

Now we find a bound for $\sigma^2[M_k]$.

$$\sigma^2[M_k] = \sigma^2\left[\sum_w p_w X_{w,k}\right]$$
$$= \sum_w p_w^2\sigma^2[X_{w,k}]+\sum_{u\neq v}p_up_vCov(X_{u,k},X_{v,k})$$
$$= \frac{m}{k}\left(\frac{k}{m}\right)^2\Theta\left(\frac{1}{\sqrt{k}}\right)+\frac{m}{k}\left(\frac{m}{k}-1\right)\left(\frac{k}{m}\right)^2\Theta\left(\frac{k}{m^2}\right)$$
$$= \Theta\left(\frac{\sqrt{k}}{m}\right),$$

which completes the proof. ■

A.3 Leave-One-Out Estimation of Log-Loss

Lemma 34 **Proof** Using Lemma 9, we have $\forall^{\frac{\delta}{2}}$: $n_2 = |U \cap S_2|$ and $n_1 = |U \cap S_1|$, where $U = \{w \in V : mp_w \leq c \ln \frac{m}{\delta}\}$, for some $c > 0$. Let $n'_2 = |U \cap S_2|$ and $n'_1 = |U \cap S_1|$. Let $b = \ln \frac{m}{\delta}$.

First, we show that $E[n'_2] = O(bE[n'_1])$.

$$\begin{aligned} E[n'_2] &= \sum_{w \in U} \binom{m}{2} p_w^2 (1-p_w)^{m-2} \\ &= \sum_{w \in U} mp_w (1-p_w)^{m-1} \left[\frac{m-1}{2} \frac{p_w}{1-p_w} \right] \\ &= \sum_{w \in U} mp_w (1-p_w)^{m-1} O(b) = O(bE[n'_1]). \end{aligned}$$

Next, we bound the deviation of n'_1 and n'_2. A single change in the sample changes n'_1, as well as n'_2, by at most 1. Therefore, using Lemma 2 for n'_1 and n'_2, we have

$$\begin{aligned} \forall^{\frac{\delta}{4}} S: &\qquad n'_1 \geq E[n'_1] - O\left(\sqrt{m \ln \frac{1}{\delta}}\right), \\ \forall^{\frac{\delta}{4}} S: &\qquad n'_2 \leq E[n'_2] + O\left(\sqrt{m \ln \frac{1}{\delta}}\right). \end{aligned}$$

Therefore,

$$n'_2 \leq E[n'_2] + O\left(\sqrt{m \ln \frac{1}{\delta}}\right) = O\left(bE[n'_1] + \sqrt{m \ln \frac{1}{\delta}}\right) = O\left(b\left(n'_1 + \sqrt{m \ln \frac{1}{\delta}}\right)\right),$$

which completes the proof. ■

A.4 Log-Loss A Priori

Theorem 39 **Proof** The KL-divergence is of course non-negative. By Lemma 40, we have

$$\forall^{\frac{\delta}{4}} S, \qquad \sum_{w \notin S} p_w \ln \left(\frac{p_w}{q_w}\right) \leq M_0 \ln \left(\frac{n_0 \ln \frac{4m}{\delta}}{\alpha n_{1\ldots\tau}}\right). \tag{42}$$

By Lemma 41 with λ, we have $\forall^{\frac{\delta}{4}} S$:

$$\sum_{w \in S: p_w \leq \frac{\lambda \ln \frac{8m}{\delta}}{m}} p_w \ln \left(\frac{p_w}{q_w} \right) \leq \frac{\lambda \ln \frac{8m}{\delta}}{1-\alpha} \left(\sqrt{\frac{3 \ln \frac{8}{\delta}}{m}} + M_0 \right) + \frac{\alpha}{1-\alpha} F_{\frac{\lambda \ln \frac{8m}{\delta}}{m}}. \tag{43}$$

By Lemma 43 with λ, we have $\forall^{\frac{\delta}{2}} S$:

$$\sum_{w \in S: p_w > \frac{\lambda \ln \frac{8m}{\delta}}{m}} p_w \ln \left(\frac{p_w}{q_w} \right) \leq \sqrt{\frac{3 \ln \frac{8}{\delta}}{m}} + \frac{3\lambda \ln \frac{8m}{\delta}}{2(\sqrt{\lambda} - \sqrt{3})^2 m} N_{\frac{\lambda \ln \frac{8m}{\delta}}{m}}. \tag{44}$$

The proof follows by combining Equations (42), (43), and (44). ■

Lemma 42 **Proof** Let $f(x) = \frac{x^2}{2(1-\Delta)^2} - x + \ln(1+x)$. Then,

$$\begin{aligned} f'(x) &= \frac{x}{(1-\Delta)^2} - 1 + \frac{1}{1+x}, \\ f''(x) &= \frac{1}{(1-\Delta)^2} - \frac{1}{.}(1+x)^2 \end{aligned}$$

Clearly, $f(0) = f'(0) = 0$. Also, $f''(x) \geq 0$ for any $x \in [-\Delta, \Delta]$. Therefore, $f(x)$ is non-negative in the range above, and the lemma follows. ■

References

D. Angluin and L. G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18:155–193, 1979.

S. F. Chen. *Building Probabilistic Models for Natural Language*. PhD thesis, Harvard University, 1996.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, 1998.

K. W. Church and W. A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5: 19–54, 1991.

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.

D. P. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

P. Flajolet. Singularity analysis and asymptotics of Bernoulli sums. *Theoretical Computer Science*, 215:371–381, 1999.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(16):237–264, 1953.

I. J. Good. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. *Journal of Statistical Computation and Simulation*, 66(2):101–112, 2000.

W. Hoeffding. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics*, 27:713–721, 1956.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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.

S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

S. Kutin. *Algorithmic Stability and Ensemble-Based Learning*. PhD thesis, University of Chicago, 2002.

D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research, Special Issue on Learning Theory*, 4(Oct): 895–911, 2003.

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.

D. McAllester and R. E. Schapire. Learning theory and language modeling. In *Seventeenth International Joint Conference on Artificial Intelligence*, 2001.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. London Math. Soc. Lectures Notes 141, Cambridge University Press, 1989.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(Oct):427–431, 2003.