On the Convergence Rate of Good-Turing Estimators

David McAllester **Robert E. Schapire**
AT&T Labs − Research
Shannon Laboratory
180 Park Avenue
Florham Park, NJ 07932
{dmac, schapire}@research.att.com

Abstract

Good-Turing adjustments of word frequencies are an important tool in natural language modeling. In particular, for any sample of words, there is a set of words not occuring in that sample. The total probability mass of the words not in the sample is the so-called missing mass. Good showed that the fraction of the sample consisting of words that occur only once in the sample is a nearly unbiased estimate of the missing mass. Here, we give a PAC-style high-probability confidence interval for the actual missing mass. More generally, for $k \geq 0$, we give a confidence interval for the true probability mass of the set of words occuring k times in the sample.

1 INTRODUCTION

Since the publication of the Good-Turing estimators in 1953 [4], these estimators have been used extensively in language modeling applications [2, 3, 6]. In spite of the extensive use of Good-Turing estimators, little theoretical work has been done on these estimators since the original theorems showing that they have negligible bias. In this paper, we briefly review the Good-Turing estimators and then prove new convergence rates, i.e., we give PAC-style high-probability confidence intervals for the true values of the estimated quantities.

Perhaps the most significant Good-Turing estimator is the estimate of the missing mass. We assume that there is some unknown underlying distribution on some unknown set of objects, e.g., an unknown frequency for each word in English. We assume that a sample is constructed by drawing objects independently according to this unknown distribution. If the number of objects with nonzero probability is infinite then for any finite sample there will be objects of nonzero probability that do not occur in the sample. It is well known that in any sample of English text there will be English words not occurring in the sample. The missing mass of a sample is the total probability mass of the objects not occurring in the sample. The Good-Turing estimate of the missing mass is the fraction of the sample consisting of objects that occur exactly once in the sample. The fundamental Good-Turing result is that this estimate has negligible bias. However, to our knowledge, the convergence rate of this estimator has never been formally analyzed.

According to Good [5], the Good-Turing estimators were developed by Alan Turing during World War II while breaking Enigma codes. The Enigma was an encryption device used by the German navy. The Enigma used, as part of its encryption key, a three letter sequence. These three letter sequences were selected from a book containing all such sequences in a random order. However, a person opening the book and selecting an entry was likely to select a previously used entry, say the entry on the top of a page where the binding of the book was creased. Given a sample of previously used entries, Turing wanted to estimate the likelihood that the current unknown entry was one that had been previously used, and further, to estimate the probability distribution over the previously used entries. This lead to the development of the estimators of the missing mass and estimates of the true probability mass of the set of items occuring k times in the sample. Good worked with Turing during the war and, with Turing's permission, published the analysis of the bias of these estimators in 1953. As mentioned above, these estimators have now become standard in a variety of natural language processing applications.

In this paper, we analyze the convergence rate of the Good-Turing estimators. Let G_0 be the fraction of the sample consisting of words that occur only once in the sample and let M_0 be the actual missing mass, i.e., the total probability mass of the items not occurring in the sample. We prove that with probability at least $1 - \delta$ over the choice of the sample, we have that M_0 is no larger than

$$G_0 + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right)$$

where m is the size of the sample. This is true independent of the underlying distribution. We also give a somewhat weaker PAC lower bound on M_0 and PAC bounds on the true total probability mass of the set of words occurring k times in the sample.

2 THE GOOD-TURING ESTIMATORS

We assume an unknown probability distribution P on a countable *vocabulary* V and we denote the probability of word w by P_w. In practice, this is often taken to be the words of some natural language, such as English, although of course are results are applicable when the vocabulary is any countable universe of objects. We consider a sample S of m words drawn independently from V according to distribution P. Throughout the paper, we will write $\forall^\delta S\ \Phi[S]$ to mean that with probability at least $1-\delta$ over the choice of the sample we have that $\Phi[S]$ holds.

For a sample S of m words and for any word $w \in V$ we define $c(w)$ to be the number of times word w occurs in the sample S. For any integer $k \geq 0$, we define S_k to be the set of words $w \in V$ such that $c(w) = k$. Note that S_0 is the set of words in V not occuring in S. We define M_k to be probability of drawing a word in the set S_k:

$$M_k \equiv \sum_{w \in S_k} P_w.$$

Note that M_k depends on the sample, i.e., it is a random variable.

The Good-Turing estimators estimate the quantities M_k. These quantities are conceptually useful in constructing language models. The quantity M_0 is the so-called *missing mass*, i.e., the total probability mass of words not occuring in the sample. Intuitively, a language model should reserve some probability mass for words not in the sample since it is unlikely (or even impossible if the vocabulary is larger than the sample) that all the words in a large vocabulary will be seen in the sample. Similarly, for $k \geq 1$ the quantitiy M_k is useful in estimating the true probability of a word that occurs k times in the sample. Specifically, for $w \in S_k$, if we know M_k, then a good estimate of P_w would be $M_k/|S_k|$. For k small, we usually have that M_k is significantly smaller than its "natural" estimate $k|S_k|/m$. For example, if all words in a large sample occur only once, then S_1 is the entire sample but M_1 is almost certainly near zero.

The Good-Turing estimate of M_k, which we denote G_k, can be defined as follows:[1]

$$G_k \equiv \frac{k+1}{m-k}|S_{k+1}|.$$

Good [4] showed that for k small and m large this estimate has small bias, that is, the expectation of G_k is very close to the expectation of M_k. We prove a variant of Good's theorem here:

Theorem 1 *For $k < m$ we have*

$$\mathrm{E}\left[M_k\right] = \mathrm{E}\left[G_k\right] - \frac{k+1}{m-k}\mathrm{E}\left[M_{k+1}\right].$$

Proof: Note that $\mathrm{E}\left[M_k\right]$ can be written as follows:

$$\begin{aligned}
\mathrm{E}\left[M_k\right] &= \sum_{w \in V} P_w \Pr\left[w \in S_k\right] \\
&= \sum_{w \in V} \binom{m}{k} P_w^{k+1}(1-P_w)^{m-k} \\
&= \sum_{w \in V} \Pr\left[w \in S_{k+1}\right] \frac{\binom{m}{k}}{\binom{m}{k+1}}(1-P_w) \\
&= \sum_{w \in V} \frac{k+1}{m-k} \Pr\left[w \in S_{k+1}\right](1-P_w) \\
&= \frac{k+1}{m-k}\sum_{w \in V} \Pr\left[w \in S_{k+1}\right] \\
&\quad -\frac{k+1}{m-k}\sum_{w \in V} \Pr\left[w \in S_{k+1}\right] P_w \\
&= \frac{k+1}{m-k}\mathrm{E}\left[|S_{k+1}|\right] - \frac{k+1}{m-k}\mathrm{E}\left[M_{k+1}\right] \\
&= \mathrm{E}\left[G_k\right] - \frac{k+1}{m-k}\mathrm{E}\left[M_{k+1}\right].
\end{aligned}$$

■

Theorem 1 immediately implies that for k much smaller than m we have that G_k is a nearly unbiased estimate of M_k. More specifically, since $M_{k+1} \in [0,1]$ we have the following corollary of Theorem 1.

Corollary 2 *For $k < m$ we have*

$$|\mathrm{E}\left[M_k\right] - \mathrm{E}\left[G_k\right]| \leq \frac{k+1}{m-k}.$$

Note in particular that $|\mathrm{E}\left[G_0\right] - \mathrm{E}\left[M_0\right]| \leq 1/m$.

It is interesting to note that it is possible to "unwind" the equation in Theorem 1. For example, we can use $G_0 - G_1/m$ as an improved estimate of M_0. By observing that $M_2 \leq 1$ we get that the bias of this improved estimate is at most $2/(m(m-1))$. More generally, the bias of an estimator based on using the equation in Theorem 1 d times will be $O(1/m^d)$. However, it seems that the variance of these estimators is large compared to $1/m$, so reducing the bias below $O(1/m)$ is not a significant improvement.

3 CONVERGENCE OF THE GOOD-TURING ESTIMATORS

The first main result of this paper bounds the rate at which the Good-Turing estimators converge. More specifically, we have the following:

Theorem 3

$$\begin{aligned}
&\forall \delta > 0\ \ \forall^\delta S\quad |G_k - M_k| \leq \\
&\quad \frac{k+2}{m-k} + \sqrt{\frac{2\ln(\frac{3}{\delta})}{m}} \times \\
&\quad\quad \left[\frac{k+1}{1-k/m} + k + \sqrt{2k\ln\left(\frac{3m}{\delta}\right)} + 2\ln\left(\frac{3m}{\delta}\right)\right].
\end{aligned}$$

Note that for fixed k and δ, we have that the bound on $|G_k - M_k|$ converges to zero as m increases at the rate $O((\ln m)/\sqrt{m})$ independent of the size or distribution of the underlying vocabulary. Furthermore, the width of the confidence interval has only logarithmic dependence on the confidence parameter δ. For k small compared to $\ln(3m/\delta)$, the bound is approximately

$$2\ln\left(\frac{3m}{\delta}\right)\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{m}}.$$

[1] The Good-turing estimate is often defined to be $\frac{k+1}{m}|S_{k+1}|$. For k much smaller than m this is essentially the same as the definition used here. However, the estimate $\frac{k+1}{m-k}|S_{k+1}|$ has slighly smaller bias and is theoretically easier to work with.

For k large compared to $\ln(3m/\delta)$, but still small compared to m, the bound is approximately

$$2k\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{m}}.$$

The bound is vacuous for $k \geq \sqrt{m}$.

The basic idea behind the proof is to introduce a threshold Θ such that, with high confidence, all words w with $P_w > \Theta$ occur more than k times and hence do not influence M_k. Given an upper bound on P_w for words influencing M_k we have that a single (plausible) change in the sample can change M_k by at most 2Θ. Given a bound on the influence of a single sample element on M_k (and also G_k), we can apply McDiarmid's theorem which gives a convergence rate for any function of the sample where single changes in the sample have limited influence.

To establish an appropriate value for Θ we use the following lemma:

Lemma 4 *If a biased coin has probability p of being heads, and $\hat{p}$ is the fraction of times the coin comes up heads in a sample S of m independent tosses, then we can bound p in terms of $\hat{p}$ as follows.*

$$\forall \delta > 0 \ \ \forall^\delta S \ \ p \leq \hat{p} + \sqrt{\frac{2\hat{p}\ln(1/\delta)}{m}} + \frac{2\ln(1/\delta)}{m}.$$

Proof: The relative Chernoff bound [1] states the following for $\gamma > 0$:

$$\Pr\left[\hat{p} < (1-\gamma)p\right] \leq e^{-pm\gamma^2/2}.$$

Setting this probability equal to δ and solving for γ we can rephrase this bound as follows:

$$\forall^\delta S \quad p - \hat{p} \leq \sqrt{\frac{2p\ln(\frac{1}{\delta})}{m}}. \tag{1}$$

We use "high confidence implication" which states that if $\forall^\delta S \ \Phi[S]$ and $\Phi[S]$ implies $\Psi[S]$, then $\forall^\delta S \ \Psi[S]$. In particular, consider any sample satisfying the body of Eq. (1). The body of Eq. (1) implies that

$$m(p-\hat{p})^2 \leq 2p\ln(1/\delta),$$

that is,

$$mp^2 - (2m\hat{p} + 2\ln(1/\delta))p + m\hat{p}^2 \leq 0,$$

which implies that p is at most

$$\begin{aligned} & \frac{(2m\hat{p} + 2\ln(1/\delta)) + \sqrt{(2m\hat{p}+2\ln(1/\delta))^2 - 4m^2\hat{p}^2}}{2m} \\ = \quad & \hat{p} + \frac{\ln(1/\delta)}{m} + \sqrt{\frac{8m\hat{p}\ln(1/\delta) + 4\ln^2(1/\delta)}{4m^2}} \\ = \quad & \hat{p} + \frac{\ln(1/\delta)}{m} + \sqrt{\frac{2\hat{p}\ln(1/\delta)}{m} + \frac{\ln^2(1/\delta)}{m^2}} \\ \leq \quad & \hat{p} + \frac{2\ln(1/\delta)}{m} + \sqrt{\frac{2\hat{p}\ln(1/\delta)}{m}} \end{aligned}$$

completing the proof. ■

We now define $\Theta(\hat{p}, \delta)$ to be the bound in Lemma 4:

$$\Theta(\hat{p}, \delta) \equiv \hat{p} + \sqrt{\frac{2\hat{p}\ln(1/\delta)}{m}} + \frac{2\ln(1/\delta)}{m}.$$

We also define M_k^δ as follows:

$$M_k^\delta \equiv \sum_{w \in S_k:\ P_w \leq \Theta(k/m, \delta/m)} P_w.$$

Note that M_k^δ consists of that fragment of M_k due to "low frequency" words. The frequency threshold $\Theta(k/m,\ \delta/m)$ is selected so that M_k^δ is essentially the same as M_k; with high confidence, $M_k^\delta = M_k$ and their expectations differ by at most $1/m$.

Lemma 5 *For $m > 1$ we have that*

$$\forall \delta > 0 \ \ \forall^\delta S \quad M_k^\delta = M_k.$$

Proof: First we use "union bound quantification" which states that if W is a finite set such that

$$\forall x \in W \ \ \forall \delta > 0 \ \ \forall^\delta S \ \ \Phi[x,\ S,\ \delta]$$

then

$$\forall \delta > 0 \ \ \forall^\delta S \ \ \forall x \in W \ \ \Phi[x,\ S,\ \delta/|W|].$$

This is simply a formulation of the union bound. Applying union bound quantification to Lemma 4 with W being the set of words w such that $P_w > \frac{1}{m}$, we get that

$$\forall \delta > 0 \ \ \forall^\delta S \ \ \forall w : P_w > \frac{1}{m}, \quad P_w \leq \Theta\left(\frac{c(w)}{m},\ \frac{\delta}{m}\right). \tag{2}$$

By high confidence implication, it now suffices to show that the body of (2) implies $M_k^\delta = M_k$. Assume the body of (2). To show $M_k^\delta = M_k$ we must show that for any word w with $P_w > \Theta(k/m,\ \delta/m)$ we have $c(w) > k$. Let w be any such word. One can check that for $m > 1$ we have $\Theta(k/m,\ \delta/m) > 1/m$. Hence $P_w > 1/m$ and so by the body of (2) we have $P_w \leq \Theta(c(w)/m,\ \delta/m)$. But this implies $\Theta(k/m,\ \delta/m) < P_w \leq \Theta(c(w)/m,\ \delta/m)$ which implies $c(w) > k$. ■

Lemma 6

$$\forall \delta \in [0,1], \quad \left|\mathrm{E}\left[M_k\right] - \mathrm{E}\left[M_k^\delta\right]\right| \leq \frac{1}{m}$$

Proof: First note the following:

$$\mathrm{E}\left[M_k\right] - \mathrm{E}\left[M_k^\delta\right] = \sum_{w:\ P_w > \Theta(k/m,\ \delta/m)} P_w \Pr\left[w \in S_k\right].$$

It now suffices to show that for $P_w > \Theta(k/m,\ \delta/m)$ we have $\Pr\left[w \in S_k\right] \leq 1/m$. Lemma 4 can be rephrased as

$$\Pr\left[\Theta\left(\frac{c(w)}{m},\ \frac{\delta}{m}\right) < P_w\right] \leq \frac{\delta}{m}.$$

For $P_w > \Theta(k/m,\ \delta/m)$ this implies

$$\Pr\left[\Theta\left(\frac{c(w)}{m},\ \frac{\delta}{m}\right) \leq \Theta\left(\frac{k}{m},\ \frac{\delta}{m}\right)\right] \leq \frac{\delta}{m},$$

and therefore

$$\Pr\left[c(w) \le k\right] \le \frac{\delta}{m}.$$

So we have $\Pr[w \in S_k] \le \Pr\left[c(w) \le k\right] \le \delta/m \le 1/m$. ■

Now that we have established that M_k^δ behaves much like M_k, we use the fact that a single change in the sample can not have much influence on the value of M_k^δ. The following theorem of McDiarmid [7] states that any function of the sample for which a single change in the sample has limited effect must converge to its expectation as the sample gets large.

Theorem 7 (McDiarmid) *Let $X_1, \ldots, X_m$ be independent random variables taking values in a set V and let $f : V^m \to \mathbb{R}$ be such that*

$$\sup |f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)|$$

is at most c_i where the supremum is taken over all $x_1, \ldots, x_m$, $x_i' \in V$. Then with probability at least $1 - \delta$

$$f(X_1, \ldots, X_m) \le E[f(X_1, \ldots, X_m)] + \sqrt{\frac{\ln(\frac{1}{\delta}) \sum_{i=1}^m c_i^2}{2}},$$

and with probability at least $1 - \delta$

$$f(X_1, \ldots, X_m) \ge E[f(X_1, \ldots, X_m)] - \sqrt{\frac{\ln(\frac{1}{\delta}) \sum_{i=1}^m c_i^2}{2}}.$$

A natural special case is $x_i \in [0,1]$ and $f(x_1, \ldots, x_n) = \frac{1}{m}\sum_{i=1}^m x_i$. In this case, $c_i = 1/m$ and McDiarmid's theorem reduces to the Heoffding inequalities.

The "union bound conjunction principle" states that, for any positive numbers j and k, if

$$\forall \delta > 0 \ \ \forall^\delta S \ \ \Phi\left[S, \ \frac{\delta}{j}\right]$$

and

$$\forall \delta > 0 \ \ \forall^\delta S \ \ \Psi\left[S, \ \frac{\delta}{k}\right]$$

then

$$\forall \delta > 0 \ \ \forall^\delta S \ \ \left(\Phi\left[S, \ \frac{\delta}{k+j}\right] \wedge \Psi\left[S, \ \frac{\delta}{k+j}\right]\right).$$

This can be rephrased equivalently to say that if

$$\forall \delta > 0 \ \ \forall^{j\delta} S \ \ \Phi[S, \ \delta]$$

and

$$\forall \delta > 0 \ \ \forall^{k\delta} S \ \ \Psi[S, \ \delta]$$

then

$$\forall \delta > 0 \ \ \forall^{(j+k)\delta} S \ \ \left(\Phi[S, \ \delta] \wedge \Psi[S, \ \delta]\right)$$

which clearly follows from the union bound.

Applying union bound conjunction to the two conclusions in McDiarmid's theorem gives that, with probability at least $1 - \delta$,

$$|f(X_1, \ldots, X_n) - \mathrm{E}\left[f(X_1, \ldots, X_n)\right]| \le \sqrt{\frac{\ln(\frac{2}{\delta}) \sum_{i=1}^m c_i^2}{2}}. \qquad (3)$$

Using Eq. (3) we can prove the following:

Lemma 8

$$\forall \delta > 0 \ \ \forall^\delta S \quad |(G_k - M_k^\gamma) - \mathrm{E}\left[G_k - M_k^\gamma\right]| \le$$

$$\left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\gamma}{m}\right)\right)\sqrt{2m \ln\left(\frac{2}{\delta}\right)}$$

Proof: We apply Eq. (3) with V being the vocabulary of possible words and X_i being the ith word in the sample. We take $f(X_1, \ldots, X_n)$ to be $G_k - M_k^\gamma$. Note that when a word is replaced in the sample, one word increases its count while another word decreases its count. This implies that a single replacement can change $|S_k|$ by at most 2. So a single replacement can change G_k by at most $2(k+1)/(m-k)$. A single replacement can change M_k^γ by at most $2\Theta(k/m, \ \gamma/m)$. So a single change in the sample can change $G_k - M_k^\gamma$ by at most

$$2\left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\gamma}{m}\right)\right).$$

Eq. (3) then implies the lemma. ■

Proof of Theorem 3: We apply union bound conjunction to lemmas 5 and 8 with $\delta/3$ inserted for γ in Lemma 8. When then get that the following holds with probability at least $1 - \delta$:

$$\begin{aligned}
&|G_k - M_k| \\
&\quad = \left|G_k - M_k^{\delta/3}\right| \\
&\quad \le \left|\mathrm{E}\left[G_k\right] - \mathrm{E}\left[M_k^{\delta/3}\right]\right| \\
&\qquad + \left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\delta}{3m}\right)\right)\sqrt{2m \ln\left(\frac{3}{\delta}\right)} \\
&\quad \le \left|\mathrm{E}\left[G_k\right] - \mathrm{E}\left[M_k\right]\right| + \left|\mathrm{E}\left[M_k^{\delta/3}\right] - \mathrm{E}\left[M_k\right]\right| \\
&\qquad + \left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\delta}{3m}\right)\right)\sqrt{2m \ln\left(\frac{3}{\delta}\right)} \\
&\quad \le \frac{k+1}{m-k} + \frac{1}{m} \\
&\qquad + \left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\delta}{3m}\right)\right)\sqrt{2m \ln\left(\frac{3}{\delta}\right)} \\
&\quad \le \frac{k+2}{m-k} + \left(\frac{k+1}{m-k} + \Theta\left(\frac{k}{m}, \ \frac{\delta}{3m}\right)\right)\sqrt{2m \ln\left(\frac{3}{\delta}\right)} \\
&\quad = \frac{k+2}{m-k} + \sqrt{\frac{2\ln(\frac{3}{\delta})}{m}} \ \times \\
&\qquad \left[\frac{k+1}{1-k/m} + k + \sqrt{2k \ln\left(\frac{3m}{\delta}\right)} + 2\ln\left(\frac{3m}{\delta}\right)\right]
\end{aligned}$$

This inequality is trivially true when $m = 1$ and Theorem 3 follows. ■

4 A TIGHTER UPPER BOUND ON THE MISSING MASS

In the case of the missing mass M_0, it is possible to give a significantly tighter upper bound than that given in Theorem 3, namely, the following:

Theorem 9 $\forall\delta > 0\ \ \forall^{\delta} S \quad M_0 \leq G_0 + (2\sqrt{2}+\sqrt{3})\sqrt{\frac{\ln(\frac{3}{\delta})}{m}}.$

Note that this bound only applies to one of the tails. It remains open whether a similar bound holds on the other tail as well.

To prove this theorem, we divide M_0 into a high frequency component M_0^+ and a low frequency component M_0^- as follows:

$$\begin{aligned} M_0^+ &\equiv \sum_{w: P_w > 1/m,\ c(w)=0} P_w. \\ M_0^- &\equiv \sum_{w: P_w \leq 1/m,\ c(w)=0} P_w. \end{aligned}$$

We prove the following two lemmas seperately:

Lemma 10 $\forall\delta > 0\ \ \forall^{\delta} S \quad M_0^+ \leq \mathrm{E}\left[M_0^+\right] + \sqrt{\frac{3\ln(\frac{1}{\delta})}{m}}.$

Lemma 11 $\forall\delta > 0\ \ \forall^{\delta} S \quad M_0^- \leq \mathrm{E}\left[M_0^-\right] + \sqrt{\frac{2\ln(\frac{1}{\delta})}{m}}.$

Lemma 11 follows from an application of McDiarmid's theorem and the observation that a single change in the sample can change M_0^- by at most $2/m$. Lemma 10 is more involved and is proved at the end of this section. Note that $M_0 = M_0^- + M_0^+$ and hence, by union bound conjunction, Lemmas 10 and 11 together imply that

$$\forall\delta > 0\ \ \forall^{\delta} S\ \ M_0 \leq \mathrm{E}\left[M_0\right] + (\sqrt{2}+\sqrt{3})\sqrt{\frac{\ln(\frac{2}{\delta})}{m}}. \quad (4)$$

We also need the following two lemmas where the first follows from Theorem 1 and the second follows from an application of McDiarmid's theorem to G_0:

Lemma 12 $\mathrm{E}\left[M_0\right] \leq \mathrm{E}\left[G_0\right]$.

Lemma 13 $\forall\delta > 0\ \ \forall^{\delta} S \quad \mathrm{E}\left[G_0\right] \leq G_0 + \sqrt{\frac{2\ln(\frac{1}{\delta})}{m}}.$

Theorem 9 now follows by applying union bound conjunction to Eq. (4) and Lemma 13 so that the bodies of Eq. (4), Lemma 12 and Lemma 13 all hold simultaneously.

It now remains only to prove Lemma 10. The proof is based on Chernoff's method. The first step is to prove the following:

Lemma 14 *For $\lambda > 0$ and $\epsilon > 0$ we have*

$$\Pr\left[M_0^+ \geq \mathrm{E}\left[M_0^+\right] + \epsilon\right] \leq e^{F(\lambda) - \lambda\epsilon}$$

where

$$F(\lambda) \equiv \sum_{w:\ P_w > 1/m} \left(\ln(Q_w e^{\lambda P_w} + (1-Q_w)) - \lambda P_w Q_w\right)$$

and $Q_w = (1-P_w)^m$ is the probability that word w does not occur in the sample.

Proof: In Chernoff's method, we bound the tail probability using Markov's inequality:

$$\begin{aligned} &\Pr\left[M_0^+ \geq \mathrm{E}\left[M_0^+\right] + \epsilon\right] \\ &\quad= \Pr\left[\exp\left(\lambda(M_0^+ - \mathrm{E}\left[M_0^+\right] - \epsilon)\right) \geq 1\right] \\ &\quad\leq \mathrm{E}\left[\exp\left(\lambda(M_0^+ - \mathrm{E}\left[M_0^+\right] - \epsilon)\right)\right] \\ &\quad= e^{-\lambda(\mathrm{E}\left[M_0^+\right] + \epsilon)}\ \mathrm{E}\left[e^{\lambda M_0^+}\right]. \end{aligned} \quad (5)$$

Let $B = \{w \in V : P_w > 1/m\}$. For each word $w \in B$, we introduce a random variable X_w which is 1 if w does *not* occur in the sample and 0 otherwise. We can then write M_0^+ as

$$M_0^+ = \sum_{w \in B} X_w P_w.$$

Clearly, $\mathrm{E}\left[X_w\right] = Q_w$ so

$$\mathrm{E}\left[M_0^+\right] = \sum_{w \in B} Q_w P_w. \quad (6)$$

Now

$$\begin{aligned} e^{\lambda M_0^+} &= \exp\left(\lambda \sum_{w \in B} P_w X_w\right) \\ &= \prod_{w \in B} e^{\lambda P_w X_w} \\ &= \prod_{w \in B} \left(1 + \left(e^{\lambda P_w} - 1\right) X_w\right) \end{aligned} \quad (7)$$

where the last equality uses the fact that $X_w \in \{0,1\}$. Multiplying out the product, we can write Eq. (7) as a polynomial:

$$\prod_{w \in B} \left(1 + \left(e^{\lambda P_w} - 1\right) X_w\right) = \sum_{A \subseteq B} c_A \prod_{w \in A} X_w \quad (8)$$

for some coefficients c_A. Furthermore, because $\lambda P_w \geq 0$, all of the coefficients c_A are nonnegative.

Note that $\prod_{w \in A} X_w$ is 1 if none of the words w in A occur in the sample S and is 0 otherwise. Thus,

$$\begin{aligned} \mathrm{E}\left[\prod_{w \in A} X_w\right] &= \left(1 - \sum_{w \in A} P_w\right)^m \\ &\leq \left(\prod_{w \in A} (1 - P_w)\right)^m \\ &= \prod_{w \in A} Q_w. \end{aligned} \quad (9)$$

The inequality here can be proved by induction on $|A|$ using the fact that $1 - p - q \leq (1-p)(1-q)$ for $p, q \geq 0$. Thus, combining Eqs. (7), (8) and (9) gives

$$\begin{aligned} \mathrm{E}\left[e^{\lambda M_0^+}\right] &= \sum_{A \subseteq B} c_A \mathrm{E}\left[\left(\prod_{w \in A} X_w\right)\right] \\ &\leq \sum_{A \subseteq B} c_A \prod_{w \in A} Q_w \\ &= \prod_{w \in B} \left(1 + \left(e^{\lambda P_w} - 1\right) Q_w\right). \end{aligned}$$

Combined with Eqs. (5) and (6) this gives

$$\begin{aligned}
&\Pr\left[M_0^+ \geq \mathrm{E}\left[M_0^+\right] + \epsilon\right] \\
&\quad\leq \exp\left(-\lambda\epsilon - \lambda \sum_{w\in B} P_w Q_w\right) \cdot \\
&\qquad \prod_{w\in B} \left(1 + \left(e^{\lambda P_w} - 1\right) Q_w\right) \\
&\quad= e^{F(\lambda)-\lambda\epsilon}.
\end{aligned}$$

■

Next we prove the following bound on the function $F(\lambda)$:

Lemma 15 *For* $\lambda \leq m/2$

$$F(\lambda) \leq \frac{\lambda^2}{(e-1)m}.$$

Proof: First, note that $F(0) = 0$. Now let $F'(\lambda)$ denote the first derivative of F, i.e., $dF/d\lambda$ evaluated at λ. Then

$$F'(\lambda) = \sum_{w:\, P_w > 1/m} \frac{Q_w P_w}{(1-Q_w)e^{-\lambda P_w} + Q_w} - Q_w P_w$$

Note that $F'(0) = 0$. Now letting $F''(\lambda)$ denote the second derivative of F we get that

$$\begin{aligned}
F''(\lambda) &= \sum_{w:\, P_w > 1/m} \frac{Q_w P_w^2 (1-Q_w) e^{-\lambda P_w}}{\left[(1-Q_w)e^{-\lambda P_w} + Q_w\right]^2} \\
&\leq \sum_{w:\, P_w > 1/m} \frac{Q_w P_w^2 (1-Q_w) e^{-\lambda P_w}}{\left[(1-Q_w)e^{-\lambda P_w}\right]^2} \\
&= \sum_{w:\, P_w > 1/m} \frac{Q_w P_w^2}{(1-Q_w)e^{-\lambda P_w}} \\
&= \sum_{w:\, P_w > 1/m} P_w \frac{Q_w P_w e^{\lambda P_w}}{(1-Q_w)} \\
&\leq \sum_{w:\, P_w > 1/m} P_w \frac{P_w e^{(\lambda - m)P_w}}{(1-Q_w)} \\
&\leq \sum_{w:\, P_w > 1/m} P_w \frac{P_w e^{(\lambda - m)P_w}}{(1-1/e)}
\end{aligned}$$

where the last two inequalities use the inequality $Q_w = (1 - P_w)^m \leq e^{-mP_w}$ which is at most $1/e$ for $P_w \geq 1/m$. For $\alpha > 0$ and $x \geq 0$ one can show, by maximizing over x, that

$$xe^{-\alpha x} \leq \frac{1}{\alpha e}.$$

For $\lambda < m$, we can use this inequality with $\alpha = (m - \lambda)$ and get that

$$\begin{aligned}
F''(\lambda) &\leq \sum_{w:\, P_w > 1/m} P_w \frac{1}{(e-1)(m-\lambda)} \\
&\leq \frac{1}{(e-1)(m-\lambda)}.
\end{aligned}$$

Since $\lambda \leq m/2$ we then have that

$$F''(\lambda) \leq \frac{2}{(e-1)m}.$$

The lemma now follows from $F(0) = 0$, $F'(0) = 0$ and $F''(\lambda) \leq 2/(e-1)m$. ■

Proof of Lemma 10: Let $\lambda = m\epsilon/2$. Lemmas 14 and 15 together imply that

$$\begin{aligned}
\Pr\left[M_0^+ \geq \mathrm{E}\left[M_0^+\right] + \epsilon\right] &\leq \exp\left(\frac{\lambda^2}{(e-1)m} - \lambda\epsilon\right) \\
&= \exp\left(\frac{m\epsilon^2}{4(e-1)} - \frac{m\epsilon^2}{2}\right) \\
&\leq e^{-m\epsilon^2/3}.
\end{aligned}$$

Lemma 10 now follows by setting this probability equal to δ and solving for ϵ. This completes the proof of Theorem 9. ■

References

[1] Dana Angluin and Leslie G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, April 1979.

[2] Stanley F. Chen. *Building probabilistic models for natural language*. PhD thesis, Harvard University, May 1996.

[3] Kenneth W. Church and William 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.

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

[5] I. J. Good. Turing's anticipation of emprical Bayes in connection with the cryptanalysis of the Naval Enigma. *Journal of Statistical Computation and Simulation*, in press.

[6] Slava 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*, ASSP-35(3):400–401, March 1987.

[7] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.