On the Convergence Rate of Good-Turing Estimators

by David McAllester and Robert E. Schapire

Brief Overview: Ramesh Sridharan and Matthew Johnson

1 Intuitive Problem Statement

Consider an urn containing colored balls, where the number of different colors is unknown. Suppose we draw a sample of balls from the urn (with replacement) and find some number of red and blue balls. Given our sample, we wish to estimate the probability of drawing from the urn a red ball, a blue ball, or a ball of an unseen color (the "missing mass").

The Good-Turing class of estimators provides estimates for these and similar quantities. This paper discusses bounds on the error convergence of the Good-Turing estimators as functions of a confidence parameter, and in particular provides a relatively tight upper bound on the true "missing mass."

2 Formal Setup

V	Vocabulary (countable set of words)
P	Probability distribution over V
P_w	Probability of drawing word $w \in V$ according to P
S	Sample of words from V
m	$\|S\|$: Sample size
$c(w)$	Number of times w occurs in sample S
S_k	$\{w \| c(w) = k\}$: Set of words that occur k times in sample
M_k	$\sum_{w \in S_k} P_w$: Total mass of words in S_k, M_0 is the "missing mass"
G_k	Good-Turing estimate of M_k
M_0^+	$\sum_{w:P_w > 1/m, c(w)=0} P_w$
M_0^-	$\sum_{w:P_w \leq 1/m, c(w)=0} P_w$
$\forall^\delta$	"with probability at least $1-\delta$ over the sample choice"

Table 1: Notation and quantities of interest.

We consider a countable vocabulary V and a probability distribution P over V such that word w has probability P_w. We draw a sample set S of size m i.i.d. from P. Let $c(w)$ be the number of times word w occurs in S, and S_k be the set of words such that $c(w) = k$; that is, it is the set of words that occur k times in our sample.

We additionally define M_k to be the probability of drawing a word in S_k from V. In particular, M_0 is called the "missing mass": it is the cumulative probability mass of the words in V that don't appear in our sample.

The Good-Turing estimators provide estimates of M_k (which can then provide estimates of P_w for sampled words). The Good-Turing estimate of M_k will be denoted by G_k, and it is given by

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

In particular, the estimator G_0 of the missing mass is given by $G_0 = \frac{1}{m} |S_1|$, which is simply the fraction of words that occur only once in the sample.

Intuitively, a more "natural" estimate of M_k would be something like $k|S_k|/m$. However, that natural estimate would estimate the missing mass at zero, and may therefore also greatly overestimate the mass of the words that do appear in the sample. Consider a very large vocabulary with a uniform distribution and a

relatively small sample; it is likely that all words in the sample only occur once, in which case the "natural" estimate of M_1 would be 1, while its true value is near zero.

Good's Theorem, given below, is an important bound on the bias of the Good-Turing estimators as a function of m and k. It is also the result that the paper seeks to extend via notions of confidence.

Theorem (Good's Theorem)**.** *Theorem 1 in the paper states the following:*

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

Since $M_{k+1} \in [0,1]$, we can use this to provide a bound on the estimator bias (Corollary 2 in the paper):

$$|\mathbb{E}[M_k] - \mathbb{E}[G_k]| \leq \frac{k+1}{m-k}$$

In particular we have $|\mathbb{E}[G_0] - E[M_0]| \leq 1/m$.

3 Main Results

(1) The error in the Good-Turing estimator converges with a confidence of $1-\delta$ according to

$$\forall \delta > 0 \quad \forall^{\delta} S \quad |G_k - M_k| \leq \begin{cases} 2\ln\left(\frac{3m}{\delta}\right)\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{m}} & k \text{ small compared to } \ln(3m/\delta) \\ 2k\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{m}} & k \text{ large compared to } \ln(3m/\delta) \end{cases}$$

which, for fixed δ, converges to zero at a rate of $O((\ln m)/\sqrt{m})$. This result is Theorem 3 in the paper. Note that k must always be small compared to m.

(2) As a tighter upper bound on the special case of G_0 we have

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

or, in other words, with probability at least $1-\delta$ over the choice of the sample,

$$M_0 \leq G_0 + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right)$$

independent of the underlying distribution P. This result is Theorem 9 in the paper.

We will focus on Main Result (2), since estimating M_0 is of primary interest when discussing Good-Turing estimators. The hierarchy of the proof is given in Figure 1.

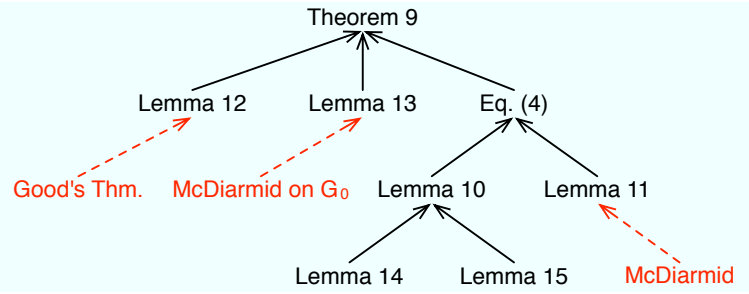

Figure 1: Proof Hierarchy for Main Result (2), titled Theorem 9 in the paper. The black is the main dependency on the lemmas given in the paper, and the red gives an idea as to how some lemmas are proven.