Distribution-Dependent Performance of the Good-Turing Estimator for the Missing Mass

Mesrob I. Ohannessian and Munther A. Dahleh

***Abstract*— The Good-Turing estimator for the missing mass has certain bias and concentration properties which define its performance. In this paper we give distribution-dependent conditions under which this performance can or cannot be matched by a trivial estimator, that is one which does not depend on observation. We introduce the notion of accrual function for a distribution, and derive our conditions from the fact that the latter governs the decay rate of the mean of the missing mass. These results shed light on the inner workings of the Good-Turing estimator, and explain why it applies particularly well for heavy-tailed distributions such as those that arise when modeling natural language.**

I. INTRODUCTION

Let $(\mathcal{X}, \mathbb{P})$ be a discrete probability space, possibly countable. We write the outcomes as $\mathcal{X} = \{1, 2, \cdots\}$ and their probability distribution as $\mathbb{P} = (p_1, p_2, \cdots)$. We associate with this space a sampling source: a process $X_1, X_2, \cdots$ of independent and identically $\mathbb{P}$-distributed random variables. For each positive integer n, we refer to $\{X_1, \cdots, X_n\}$ as an observation sample.

Definition. For each n, we define the missing mass as:

$$M_n = \mathbb{P}\left\{i \in \mathcal{X} : i \notin \{X_1, \cdots, X_n\}\right\}.$$

The name derives from the fact that M_n is the probability mass which is not represented in the observation sample. For any given n, it is a random variable, and therefore it defines a random process. The problem of interest in this paper is the estimation of M_n from the observation $\{X_1, \cdots, X_n\}$, without or with partial information about $\mathbb{P}$.

A. Estimator Performance

Let $\hat{M}_n(x_1, \cdots, x_n)$ be some estimator for M_n. We will focus in particular on two performance valuations: bias and concentration.

Definition. If there exists a function $f(n)$ such that $|\mathbf{E}[\hat{M}_n] - \mathbf{E}[M_n]| = O\left(f(n)\right)$, we say that $\hat{M}_n$ has asymptotic bias of order $f(n)$.

Particularly, when $|\mathbf{E}[\hat{M}_n] - \mathbf{E}[M_n]| \to 0$ as $n \to \infty$, we say that $\hat{M}_n$ is asymptotically unbiased and, intuitively, the bias vanishes no slower than $f(n)$.

Research partially supported by NSF grant EFRI-0735956.

The authors are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA mesrob@mit.edu, dahleh@mit.edu

Definition. If there exist constants a and b, such that for all $\epsilon > 0$ and n we have

$$\mathbb{P}\{\hat{M}_n < M_n - \epsilon\} < a \exp(-b\epsilon^2 n),$$

we say that $\hat{M}_n$ concentrates *above* M_n. If this holds for $\hat{M}_n > M_n + \epsilon$ instead, then we say $\hat{M}_n$ concentrates *below* M_n. If both hold, then $\hat{M}_n$ concentrates around M_n.

Concentration results can have various forms of the exponent. The choice of $\epsilon^2 n$ reflects the results in the literature, such as [1] and [2], based on Chernoff-like bounds.

B. The Good-Turing Estimator

We now consider the popular estimator for the missing mass, proposed by Good [3] (with due credit to Turing).

Definition. The Good-Turing estimator for the missing mass is defined as:

$$G_n = \frac{1}{n} \sum_j \mathbf{1}_{X_j \notin \{X_k \,:\, k \neq j\}}.$$

In particular, note that G_n is simply the fraction of symbols occurring exactly once in the observation.

It was known to Good that G_n is asymptotically unbiased and that for all $\mathbb{P}$, G_n has asymptotic bias of order $1/n$. More recently, McAllester and Schapire [1] showed that G_n concentrates above M_n, uniformly for all $\mathbb{P}$. There are no parallel results for concentration from below.

C. Trivial Estimators

Let's call an estimator $\hat{M}_n$ trivial if it does not depend on the observation sample. In other words, a trivial estimator is a function that depends on n alone. We will use such estimators as comparative benchmark against the performance of the Good-Turing estimator. In particular, we would like to meet the order of asymptotic bias and assert concentration from above.

D. Overview of Results

The aforementioned performance bounds for the Good-Turing estimator are distribution independent, and thus assume no information about $\mathbb{P}$. In this paper we investigate whether, for certain distributions, G_n fails to be informative about M_n. We make this notion concrete by comparing these performance bounds with those of trivial estimators.

When the same performance can be met by a trivial estimator, there is no discernible benefit in using the Good-Turing estimator. Conversely, when no trivial estimator can meet these guarantees, the value of the estimator is reinforced, and its use is promoted.

As an example, consider the simplest trivial estimator: one which evaluates to the constant 0 for all observations and for all n. Indeed, the (random) event $\{i \in \mathcal{X} : i \notin \{X_1, \cdots, X_n\}\}$, by construction, is not represented in the empirical measure. Therefore the 0-estimator is the empirical estimate of M_n. Furthermore, it is easy to show that $\mathbf{E}[M_n]$ converges to 0, and thus the 0-estimator is also asymptotically unbiased. One consequence of the present work is that if $\mathbb{P}$ falls in a certain category, then the 0-estimator has asymptotic bias decaying at the same order as that of G_n. In another category, however, G_n distinctly outperforms this trivial estimator. Such categories turn out to depend on how probability decays within $\mathbb{P}$.

The rest of the paper is organized as follows. In section II we introduce the notions of accrual function and accrual rates of a distribution. These characterize the aforementioned decay of probability, and do so intrinsically without reference to an arbitrary index. Then, in section III-A, we use accrual rates to determine the asymptotic behavior of the expected value of the missing mass. We apply these in section III-B to describe the distribution-dependent performance of the Good-Turing estimator. Lastly we conclude in section IV with descriptive and predictive consequences of our results.

II. ACCRUAL FUNCTION AND RATES

We now introduce the notion of accrual function, and use it to characterize probability distributions. Our goal is to capture the notion of probability decay and heaviness of tail without an arbitrary indexing or ordering of the symbols, such as by descending mass. We therefore give an intrinsic definition, as follows.

Definition. We define the accrual function of a distribution $\mathbb{P}$ as:

$$F(x) = \sum_{p_i \leq x} p_i.$$

Note that $F(x)$ is not a cumulative distribution. Rather, it describes how the probability mass accrues from the ground up, whence the name. It is intrinsic, because it parametrizes by probability, rather than by index. More importantly, probability decay can be described by considering its behavior near $x = 0$, using the concept of accrual rates.

Definition. Let the distribution $\mathbb{P}$ have an accrual function $F(x)$. We define the lower and upper accrual rates of $\mathbb{P}$ respectively as:

$$\underline{\rho} = \liminf_{x \to 0} \frac{\log F(x)}{\log x}, \text{ and } \overline{\rho} = \limsup_{x \to 0} \frac{\log F(x)}{\log x}.$$

If the limit exists, we simply say that $\mathbb{P}$ has accrual rate ρ.

We illustrate this with three examples. First, note that when the support of $\mathbb{P}$ is finite the accrual function is 0 near $x = 0$, therefore its accrual rate is infinite. This is indeed the steepest possible form of decay. Next, consider the case of the geometric distribution with parameter q, i.e. $p_i = (1-q)^{i-1}q$. In this case, one can compute the accrual function to be $F(x) = x/q$, at every $x = p_i$, and a piecewise step in between. Consequently, the accrual rate is 1. Lastly, consider the power law $p_i = 6/(\pi i)^2$. In this case, the accrual function is $F(x) = \sum 6/(\pi i)^2$, for $i \geq \pi^{-1}\sqrt{6/x}$. Near $x = 0$, we can approximate the sum with an integral, and find $F(x) \approx C \cdot \sqrt{x}$. Therefore, the accrual rate is $1/2$.

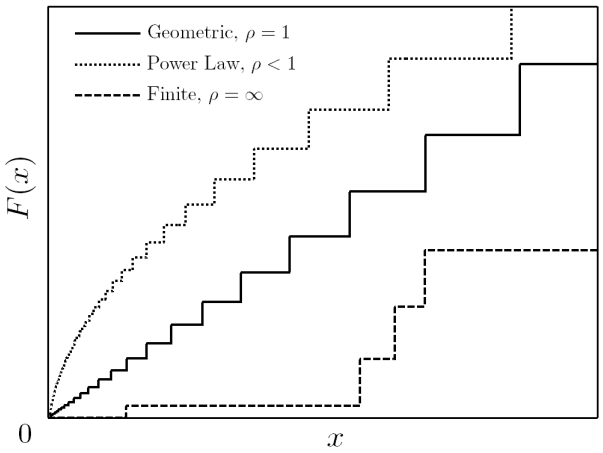

Fig. 1. Accrual functions for geometric, power law, and finite distributions.

These examples suggest a taxonomy that relates accrual rate to conventional notions of decay. Exponentially decaying tails have an accrual rate of one. Above one, the decay is super-exponential; below one, it is sub-exponential. It is interesting that one arrives at these notions, without an explicit indexing in the definition of the accrual function.

III. DISTRIBUTION DEPENDENCE

A. Behavior of $\mathbf{E}[M_n]$

We start by giving basic bounds for the expected missing mass using the accrual function. We then specialize these bounds to obtain asymptotic behavior, based on the accrual rates of the distribution.

Lemma. *Let the distribution $\mathbb{P}$ have an accrual function F. Then for any $x, y \in [0, 1]$, the expected missing mass $\mathbf{E}[M_n]$ can be bounded as follows:*

$$(1-x)^n F(x) \leq \mathbf{E}[M_n] \leq (1-y)^n + F(y). \qquad (1)$$

Proof. Let $\{Y_i\}$, $i = 1, 2, \cdots$, be indicator random variables which are 1 when symbol i does not appear in the sample and 0 otherwise. Then it follows that $M_n = \sum_i p_i Y_i$ and consequently that:

$$\mathbf{E}[M_n] = \sum_i p_i \mathbf{E}[Y_i] = \sum_i p_i (1-p_i)^n.$$

Both bounds are then obtained by splitting this sum around values of p_i below and above a given value. For the lower bound:

$$\begin{aligned} \mathbf{E}[M_n] &= \textstyle\sum_{p_i \leq x} p_i (1-p_i)^n + \sum_{p_i > x} p_i (1-p_i)^n \\ &\geq \textstyle\sum_{p_i \leq x} p_i \cdot (1-x)^n + \sum_{p_i > x} p_i \cdot 0. \end{aligned}$$

And for the upper bound:

$$\mathbf{E}[M_n] \leq \textstyle\sum_{p_i \leq y} p_i \cdot 1 + (1-y)^n.$$

The lemma then follows from the definition of F. □

Theorem 1. *Let $\mathbb{P}$ have lower and upper accrual rates $0 < \underline{\rho} \le \overline{\rho} < \infty$. Then for every $\delta > 0$ there exists n_0 such that for all $n > n_0$ we have:*

$$n^{-(\overline{\rho}+\delta)} \le \mathbf{E}[M_n] \le n^{-(\underline{\rho}-\delta)}$$

or, equivalently, for every $\delta > 0$ we have that $\mathbf{E}[M_n]$ is both $\Omega\big(n^{-(\overline{\rho}+\delta)}\big)$ and $O\big(n^{-(\underline{\rho}-\delta)}\big)$.

Proof. We start with the lower bound. Choose any α in $(\overline{\rho}, \overline{\rho}+\delta)$. Then there exists x_1 such that for all $x < x_1$ we have $\log F(x)/\log x \le \alpha$, or alternatively $F(x) \ge x^\alpha$.

Consider the expression $(1-x)^n x^\alpha$, and note that its maximal value is achieved at $\overline{x} = \alpha/(n+\alpha)$. For n large enough, $\overline{x} < x_0$, and we can use the bound for $F(x)$ together with the left-hand inequality in (1) to write:

$$\begin{aligned}\mathbf{E}[M_n] &\ge (1-\overline{x})^n F(\overline{x}) \\ &\ge (1-\overline{x})^n \overline{x}^\alpha = \left(1-\tfrac{\alpha}{n+\alpha}\right)^n \tfrac{\alpha^\alpha}{(n+\alpha)^\alpha} \\ &\ge e^{-\alpha}\alpha^\alpha \tfrac{1}{(n+\alpha)^\alpha},\end{aligned}$$

therefore there exists n_1, such that for all $n > n_1$ we have $\mathbf{E}[M_n] \ge n^{-(\overline{\rho}+\delta)}$.

For the lower bound, choose any β in $(\underline{\rho}-\delta, \underline{\rho})$. Then there exists x_2 such that for all $x < x_2$ we have $\log F(x)/\log x \ge \beta$, or alternatively $F(x) \le x^\beta$.

Now consider the expression $(1-x)^n + x^\beta$, which we evaluate at the test point $\underline{x} = 1-(\beta/n)^{\beta/n}$. Note that using the fact that $e^z \ge 1+z$ with $z = \frac{\beta}{n}\log\frac{\beta}{n}$ we can show that $\underline{x} \le \frac{\beta}{n}\log\frac{n}{\beta}$, and thus for n large enough we will have $\underline{x} < x_2$. Using the bound for $F(x)$ with the right-hand inequality in (1), we can write:

$$\begin{aligned}\mathbf{E}[M_n] &\le (1-\underline{x})^n + F(\underline{x}) \\ &\le (1-\underline{x})^n + \underline{x}^\beta = \left(\tfrac{\beta}{n}\right)^\beta + \left(1-\left(\tfrac{\beta}{n}\right)^{\frac{\beta}{n}}\right)^\beta \\ &\le \left(\tfrac{\beta}{n}\right)^\beta + \left(\tfrac{\beta}{n}\log\tfrac{n}{\beta}\right)^\beta,\end{aligned}$$

therefore there exists n_2, such that for all $n > n_2$ we have $\mathbf{E}[M_n] \le n^{-(\underline{\rho}-\delta)}$. Finally, set $n_0 = n_1 \vee n_2$. □

B. Performance of G_n

We apply these results to categorize the performance of the Good-Turing estimator, based on the accrual rates of the distribution. We start with a general statement pertaining to trivial estimators, and then derive corollaries about when such estimators have the same performance guarantees as the Good-Turing estimator.

Theorem 2. *Let $0 < r < \infty$ be given. Then there exists a trivial estimator, namely $\hat{M}_n = 1/n^r$, that achieves asymptotic bias of order $1/n^r$ for all distributions with lower accrual rate $\underline{\rho} > r$.*

Conversely, given any trivial estimator, there exists a distribution with upper accrual $\overline{\rho} < r$ for which the trivial estimator does not have asymptotic bias of order $1/n^r$.

Another way to state the converse statement is that, given a trivial estimator, there is always a distribution with $\overline{\rho} < r$ such that for all n_0, there exists $n > n_0$ where the bias is larger than $1/n^r$.

Proof. Consider the forward statement. Set $\hat{M}_n = 1/n^r$, and let $\mathbb{P}$ be any distribution such that $\underline{\rho} > r$. From theorem 1, we know that for large enough n we have $\mathbf{E}[M_n] \le 1/n^r$. It immediately follows that $0 < \hat{M}_n - \mathbf{E}[M_n] < 1/n^r$, demonstrating the bias claim.

For the converse, assume to the contrary that there exists an estimator $\hat{M}^n$, such that for all $\mathbb{P}$ with $\overline{\rho} < r$ the bias is of order $1/n^r$, that is: there exists n_0 such that for all $n > n_0$ we have $|\mathbf{E}[M_n] - \hat{M}_n| < 1/n^r$.

Now, pick two distributions. First, let $\mathbb{P}$ be such that $0 < \underline{\rho} \le \overline{\rho} < r$. Then, let $\mathbb{P}'$ be such that $\overline{\rho}' < \underline{\rho}$. Also pick any t and s such that $\overline{\rho}' < t < s < \underline{\rho}$. We will show that if $\hat{M}^n$ has proper bias with $\mathbb{P}'$, it has to decay slowly, and therefore will fail to have proper bias with $\mathbb{P}$.

Focusing on $\mathbb{P}'$, let n_0 be large enough such that for every $n > n_0$, all of the following hold:

- $\mathbf{E}[M_n] - \hat{M}_n < 1/n^r$ for $\mathbb{P}'$ by our assumption,
- $\mathbf{E}[M_n] > 1/n^t$ for $\mathbb{P}'$ by theorem 1,
- $n^{s-t} > 2$, and $n^{r-s} > 1$.

In particular, we get $\hat{M}_n > \mathbf{E}[M_n] - 1/n^r > 1/n^t - 1/n^r$.

Now, moving on to $\mathbb{P}$, choose $n_1 > n_0$ such that $\mathbf{E}[M_n] < 1/n^s$, by theorem 1. Then, for all $n > n_1$ we have:

$$\begin{aligned}\hat{M}_n - \mathbf{E}[M_n] &> 1/n^t - 1/n^s - 1/n^r \\ &= \left(n^{r-s}(n^{s-t}-1)-1\right)/n^r > 1/n^r.\end{aligned}$$

But this contradicts our assumption, therefore it's false. □

This immediately results in the following corollary.

Corollary. *If $\mathbb{P}$ has accrual rates greater than 1, then the trivial estimator $\hat{M}_n = 1/n$ matches the performance of the Good-Turing estimator asymptotically.*

Proof. The bias claim follows from the forward part of theorem 2, by setting $r = 1$. For the concentration result, first note that from theorem 1 we know that for large enough n, we have $\mathbf{E}[M_n] \le 1/n = \hat{M}_n$, then invoke the fact that M_n concentrates around its own mean. Namely, there exist constants a and b such that for every $\epsilon > 0$ and n we have:

$$\mathbb{P}\{\mathbf{E}[M_n] < M_n - \epsilon\} < a\exp(-b\epsilon^2 n), \tag{2}$$

This was shown first by McAllester and Schapire [1], and later with tighter constants (in addition to concentration from below) by McAllester and Ortiz [2]. From here, one needs only to observe that for large enough n, an event of the form $\{\hat{M}_n < Z\}$ is a subset of event $\{\mathbf{E}[M_n] < Z\}$, and particularly

$$\mathbb{P}\{\hat{M}_n < M_n - \epsilon\} \le \mathbb{P}\{\mathbf{E}[M_n] < M_n - \epsilon\},$$

and asymptotic concentration follows from (2). □

It is worthwhile to remark that, in this case, even the 0-estimator achieves bias of order 1/n, since $\mathbf{E}[M_n] \le 1/n$ for large enough n, but it does not match concentration performance.

As we illustrated in section II, accrual rates above one are characteristic of super-exponential tails. Distributions in many applications fall under this category, and it is instructive to see that there is no (asymptotic) advantage in using the Good-Turing estimator in such situations.

One would like to assert a converse, to the effect that when the accrual rate of a specific $\mathbb{P}$ is below 1, then no trivial estimators can match the performance of the Good-Turing estimator. However, this naive converse is not true: if one has knowledge of the precise expression of $\mathbf{E}[M_n]$, and uses it as a trivial estimator, then bias would be zero, and concentration would follow from (2). Of course, this would constitute much more than a partial knowledge about accrual rate. In fact, the converse part of theorem 2 formalizes how lack of such precise knowledge dooms trivial estimators to failure in this case. We restate this as a corollary.

Corollary. *Given any trivial estimator $\hat{M}_n$, there is always some distribution $\mathbb{P}$ with accrual rates less than 1, for which $\hat{M}_n$ fails to match the performance of the Good-Turing estimator.*

Accrual rates above one are characteristic of sub-exponential, heavy, tails. We have thus shown that here, in contrast to the super-exponential case, one cannot construct a single trivial estimator that works as well as the Good-Turing estimator, without further knowledge about the distribution. Therefore, the Good-Turing estimator presents a distinct advantage in this situation.

C. Remarks

We end this section by noting that these results do not cover cases where r, in particular $r = 1$, is straddled by the accrual rates: $\underline{\rho} \leq r \leq \overline{\rho}$. This is due to the fact that the limiting operation in the definition of accrual rate does not provide enough maneuverability, in the same way that ratio tests and root tests in calculus are not informative about series convergence at the decision boundary. However, the framework can be made to extend if more information is available about the limits. For example, say we know that there exists x_0, such that for all $x < x_0$, we have $F(x) \leq \gamma x$. In particular, this applies to the geometric distribution. Then the proof of theorem 1 carries through and we find out, for instance, that we have $\mathbf{E}[M_n] = O(\log n/n)$. Other extensions are also possible, depending on the nature of the available information.

IV. CONCLUSION

In this paper, we considered the problem of estimation of the missing mass, with the valuation of an estimator's performance based on bias and concentration. We presented the popular Good-Turing estimator, and compared its performance with that of trivial estimators, those that do not depend on the observation. We introduced the notion of accrual function and accrual rates of a discrete distribution, and showed that they govern the asymptotic behavior of the expected value of the missing mass.

Using these results, we divided distributions into two categories: those with accrual rates greater than one, and those with accrual rates less than one. For the first, we showed that the performance of the Good-Turing estimator can be matched by a trivial estimator, and thus Good-Turing estimation offers no advantage. For the second, we showed that any trivial estimator can be adversarially paired with a distribution that puts it at a disadvantage compared to the Good-Turing estimator, making the latter distinctly nontrivial.

Distributions with accrual rates larger than one are heavy-tailed. One domain of application where such distributions appear extensively is language modeling. Zipf was one of the earliest researchers in that field to bring out this fact, as he wrote in 1949 [4]: "If we multiply the frequency of an item in a ranked frequency list of vocabulary by its rank on the frequency list, the result is a constant figure." This came to be known as Zipf's law, and describes a family of distributions with power law probability decay, relative to an integer order index.

Our method of accrual function and accrual rates offers a characterization of these and related laws intrinsically, that is without the use of an arbitrary index. Furthermore, the proof that Good-Turing estimation works precisely for such distributions and not for others, can explain why this estimator has been so successful in natural language processing [5], but has not been adopted widely in other disciplines. It also predicts that fields where practitioners are likely to apply it with success are those where such distributions arise, such as economics, networks, etc.

We conclude with an observation that may shed new light on the inner workings of the Good-Turing estimator. Since the accrual rate ρ dictates the asymptotic behavior of $\mathbf{E}[M_n]$ to be approximately $1/n^\rho$, it is tempting to estimate ρ from observation and use that expression as an estimator for M_n.

Given the empirical distribution $\hat{\mathbb{P}}_n$, one can construct the empirical accrual function $\hat{F}_n(x)$. By the definition of ρ, we need to take a limit as x tends to 0. However, the smallest value of x where $\hat{F}_n$ is informative, is $1/n$. Using that, $\hat{\rho} = -\log \hat{F}_n(1/n)/\log n$ is a plausible estimator for the accrual rate. Carrying out the above suggestion, we get $\hat{M}_n = 1/n^{\hat{\rho}} = \hat{F}_n(1/n)$ as an estimator for the missing mass. But notice that $\hat{F}_n(1/n)$ is nothing but the Good-Turing estimator G_n. Whether this is more than just a coincidence is worth investigating.

REFERENCES

[1] D. McAllester and R.E. Schapire. On the convergence rate of Good Turing estimators. *13^{th} Annual Conference on Computational Learning Theory*, 2000.

[2] D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895-911, 2003.

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

[4] G. Zipf. Human behavior and the principle of least effort: An introduction to human ecology. New York: Hafner, 1949.

[5] W. A. Gale and G. Sampson. Good-Turing frequency estimation without tears. Journal of Quantitative Linguistics, 2(3):217-237, 1995.