On Inference about Rare Events

by

Mesrob I. Ohannessian

B.Eng., Computer and Communications Engineering,
American University of Beirut, 2002

S.M., Information Technology,
Massachusetts Institute of Technology, 2005

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

February 2012

Author: ______________________________
Department of Electrical Engineering and Computer Science
January 31, 2012

Certified by: ______________________________
Professor Munther A. Dahleh
Associate Department Head, Professor of Electrical Engineering
Thesis Supervisor

Certified by: ______________________________
Professor Sanjoy K. Mitter
Professor of Electrical Engineering and Engineering Systems
Thesis Supervisor

Accepted by: ______________________________
Professor Leslie A. Kolodziejski
Professor of Electrical Engineering
Chair, Department Committee on Graduate Students

On Inference about Rare Events

by Mesrob I. Ohannessian

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2012, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Despite the increasing volume of data in modern statistical applications, critical patterns and events have often little, if any, representation. This is not unreasonable, given that such variables are critical precisely because they are rare. We then have to raise the natural question: when can we infer something meaningful in such contexts?

The focal point of this thesis is the archetypal problem of estimating the probability of symbols that have occurred very rarely, in samples drawn independently from an unknown discrete distribution. Our first contribution is to show that the classical Good-Turing estimator that is used in this problem has performance guarantees that are asymptotically non-trivial only in a heavy-tail setting. This explains the success of this method in natural language modeling, where one often has Zipf law behavior.

We then study the strong consistency of estimators, in the sense of ratios converging to one. We first show that the Good-Turing estimator is not universally consistent. We then use Karamata's theory of regular variation to prove that regularly varying heavy tails are sufficient for consistency. At the core of this result is a multiplicative concentration that we establish both by extending the McAllester-Ortiz additive concentration for the missing mass to all rare probabilities and by exploiting regular variation. We also derive a family of estimators which, in addition to being strongly consistent, address some of the shortcomings of the Good-Turing estimator. For example, they perform smoothing implicitly. This framework is a close parallel to extreme value theory, and many of the techniques therein can be adopted into the model set forth in this thesis.

Lastly, we consider a different model that captures situations of data scarcity and large alphabets, and which was recently suggested by Wagner, Viswanath and Kulkarni. In their rare-events regime, one scales the finite support of the distribution with the number of samples, in a manner akin to high-dimensional statistics. In that context, we propose an approach that allows us to easily establish consistent estimators for a large class of canonical estimation problems. These include estimating entropy, the size of the alphabet, and the range of the probabilities.

Thesis Supervisor: Professor Munther A. Dahleh
Title: Associate Department Head, Professor of Electrical Engineering
Thesis Supervisor: Professor Sanjoy K. Mitter
Title: Professor of Electrical Engineering and Engineering Systems

To Իսկուհի *– Grandmother, Storyteller, Scholar*

Acknowledgments

I would first like to thank my advisors, Professor Munther A. Dahleh and Professor Sanjoy K. Mitter. I have been privileged to have Munther as a mentor. With his no-nonsense yet jovial attitude, he has taught me how to separate the wheat from the chaff and allowed me to garner a crisper perspective on research, also sometimes on life. Sanjoy was the first to welcome me into the Laboratory for Information and Decision Systems (LIDS). Ever since, he has been an unwavering beacon of knowledge and insight, as well as an exemplary guide. I could also talk to him in at least two languages, not counting mathematics. I would also like to thank my committee member, Professor Devavrat Shah. With his energetic and inspiring personality, he has shown me how one ought to have fun with research.

LIDS has been a wonderful workplace. The environment, professors, students, and administration have all made the lab a great place to call home during my graduate studies. I have come to know well so many good people that it is hard to list them all without forgetting one or two. So if you know me well enough to have seen me in flip-flops, know that I appreciate your friendship! That said, I must mention that I regret introducing my good friend, Parikshit Shah, to coffee. He bested me before, and now I have no chance of catching up.

I am thankful to the continued friendship of Ghinwa Choueiter. I always know I can rely on her, just as much as she can rely on me. Also, Ghinwa and Ari Shapiro have become synonymous with fun and great company, and to that I am very grateful. I must also thank Michael Danziger, whom I do not see nearly often enough, but has anyway remained a great virtual friend.

My parents, and my family as a whole, who I'm sure have been impatient for me to complete my PhD, have nevertheless given me unconditional support throughout. I owe them the greatest debt, which I can only hope to pay forward.

Last but not least, I am eternally grateful to have found Tekla, the most wonderful partner and friend. We met, coincidentally, around the time when I started working on this topic. Unlike this thesis, however, I cannot imagine my days without her.

This research was supported by: NSF Grant CCF-0325774 – AFOSR through the University of California Sub Grant 0205 G GB247 – ARO Grant W911NF-07-1-0568 – NSF Grant EFRI-0735956 – NSF Grant CMS-0625635 – AFOSR Grant FA9550-08-1-0085 – ONR Grant through Georgia Institute of Technology Grant R0897-G15 – Los Alamos National Labs Subcontract 67870-001-08.

Contents

Chapter 1

Introduction

1.1 Motivation

IN modern statistical applications, one is often showered with such large amounts of data that invoking the descriptor ‘rare’ seems misguided. Yet, despite the increasing volume of data, critical patterns and events have often very little, if any, representation. This is not unreasonable, given that such variables are critical precisely because they are rare. We then have to raise the natural question: when can we infer something meaningful in such contexts?

In applications where the underlying phenomena are continuous, such as in the earth sciences or financial and actuarial measurements, one is frequently interested in inference about extreme events. Namely, one would like to estimate the probability of exceeding very large, or falling below very low, thresholds. Such events are often underrepresented in historical data, but they remain eminently relevant because of the catastrophic consequences that may ensue if they were indeed to occur: floods, bankruptcies, etc. A mature mathematical – probabilistic – theory of extremes [1] has been developed as a consequence, and has had relative success in addressing many of the modeling challenges on this front. One of the central tenets that has emerged from this theory is the importance of structure, in particular in the form of regularity of the tails of the underlying distributions. If we interpret the estimation of exceedance probabilities beyond the range of the data as an extrapolation task, we can perhaps more easily accept the importance of proper assumptions about the tail, in order to perform this task in a principled manner. Conversely, if no assumptions are made, then we would be harried by the task of extrapolate into the void with neither compass nor bearing.

On a rather different front, many applications are fundamentally discrete, ranging from natural language modeling to business analytics. One is then interested in inference about objects, for example words and syntactic rules or connectivities and preferences, that have not occurred often in the training data. Once again, we would like to operate on the fringe of the available information, and perhaps go beyond it, and in this sense such problems are qualitatively similar to the problem of extremes. Despite this fact, no attempt has been made, to the best of our knowledge, to connect these two problems. A core motivation of this thesis is to do precisely that, in the hope of better understanding when one can make inferences about discrete rare events, and how one can best do so. In particular, we focus on the main technique that, when not used directly, forms the basis of many discrete rare event estimation approaches: the Good-Turing estimator [17]. Published in 1953, this estimator emerged from British efforts to estimate the odd of Nazi Germany choosing a not-yet-cracked Enigma cipher. Coincidentally, the impetus for developing extreme value theory came after the North Sea flood of 1953, when European governments wanted to estimate the probability of such disasters. Yet for all the analysis that the Good-Turing estimator was subjected to since, no tenet paralleling that of tail regularity in extreme value theory was discussed, i.e. no structural requirement was deemed needed to apply the Good-Turing estimator or its variants. One reason why this might be so is that this estimator was examined in the classical framework of additive errors, such as bias, mean square error, asymptotic normality, and additive concentration, and it was found that it performed adequately with no need for additional restrictions. However, rare probabilities decay to zero, and therefore they are better studied multiplicatively.

■ 1.2 Contributions and Organization

In this thesis, we start in Chapter 2 by reexamining some of what was known about the Good-Turing estimator, and find in it evidence that the natural domain of its applicability is in heavy-tailed distributions. We do this by isolating a minimal property of the underlying probability distribution, the accrual rate, that both describes the 'tail' of the distribution and dictates the behavior of the rare probabilities of interest. When the distribution is light-tailed, then probabilities decay so fast that using the Good-Turing estimator is only marginally better than giving a trivial answer, such as 0 for the probability of a never seen object, i.e. the missing mass. When the distribution is heavy-tailed, however, the penalty of such a trivial estimator is too large, and the Good-Turing estimator excels. Thus structure is indeed important.

Using this groundwork, we go on in Chapter 3 to focus on the heavy-tailed regime. We particularly ask for a tougher notion of performance, strong consistency, in the sense of the ratio of the estimate and the true rare probability converging to one. This is a multiplicative property, and thus departs from the classical analysis of the Good-Turing estimator. Indeed, we show that even for a distribution as simple as the geometric, Good-Turing fails to be strongly consistent. On the other hand, we show that a certain characterization of heavy tails, Karamata's regular variation [18], which is a refinement of the accrual rate and is closely related to the tail regularity notion used in extreme value theory, is sufficient for the consistency of the Good-Turing estimator. The core of this development is a fundamental multiplicative concentration result. We derive this by using a technique introduced by McAllester and Ortiz [21] for additive concentration, which we extend to apply to all rare probabilities. We then exploit the special moment growth rates in heavy-tailed distributions, to achieve multiplicative concentration. Lastly, we use these results to construct new estimators, which not only are strongly consistent in the context of regularly varying heavy tails, but also address some of the shortcomings of the Good-Turing estimator. Our new estimators also give a principled approach that justifies some of the ad-hoc algorithms that are used as improved variants of the Good-Turing estimator. Furthermore, the correspondence with extreme value theory opens the door to quantitative cross-fertilization between the two problems, just as we believed it to be qualitatively possible.

In Chapter 4, we explore alternative structures that allow consistent estimation. In this case, we change the probabilistic model from a fixed distribution, to a distribution that changes with sample size, akin to Kolmogorov asymptotics in high-dimensional statistics. When the model was proposed in [31], it was shown that the Good-Turing estimator is consistent in this regime. So we investigate whether more inference can be performed in a systematic manner. We consider a class of interesting problems, which include entropy estimation, alphabet size estimation, and probability range estimation, and we show that it is possible to tackle them all simultaneously. We give an abstract solution methodology, and then build explicit constructions that use the consistency of the Good-Turing estimator as their basis. This effort shows that other structures that permit inference about discrete rare events are possible, and suggest that one could potentially characterize all such structures and perhaps unify them.

Each chapter of the thesis introduces all the background and preliminary material that it requires, and is self-contained. We conclude in Chapter 5 with a summary and directions for future work.

Bibliographical Notes

Chapter 2 was presented in preliminary form at MTNS 2010 [23], and Chapter 4 was presented in its almost final form at Allerton 2011 [24].

Chapter 2

Good-Turing Estimation and its Performance

LET $X_1, \cdots, X_n$ be an observation sequence of random variables drawn independently from an unknown distribution $\mathbb{P} = (p_1, p_2, \cdots)$ over a countable alphabet of symbols, which we denote by the positive integers. An alternative description is in terms of boxes (or urns) and balls, where each sample corresponds to the label of the box which a throw of a ball lands in, randomly with probability $\mathbb{P}$.

We are interested in using the observations to perform probability estimation. In particular, we would like estimators for qualitatively rare events. One concrete class of such events are the subsets $B_{n,r}$ of symbols/boxes which have appeared exactly r times in the observation. For "rare" to be a valid qualifier, we think of r as much smaller than the sample size n. The case $r = 0$, for example, corresponds to the subset of symbols/boxes which do not appear in the observation. Define the *rare probabilities* as:

$$M_{n,r} := \mathbb{P}\{B_{n,r}\}. \tag{2.1}$$

In particular, $M_{n,0}$ denotes the missing mass, the probability of unseen outcomes. We call the estimation of $M_{n,r}$ the *Good-Turing estimation problem*, in reference to the pioneering work of Good [17], who gives due credit to Turing. Their solution to this estimation problem, the *Good-Turing estimator*, is:

$$G_{n,r} := \frac{(r+1)K_{n,r+1}}{n}, \tag{2.2}$$

where $K_{n,r} := |B_{n,r}|$ is the number of distinct symbols/boxes appearing exactly r times in the sample, i.e. the number of boxes containing exactly r balls. The study of the numbers $K_{n,r}$ themselves, and the number $K_n := \sum_{r \geq 1} K_{n,r}$ is known as the occupancy problem. It is natural to ask why we have not used the obvious estimator, the empirical probability of $B_{n,r}$, which would be $\frac{rK_{n,r}}{n}$ in contrast to (2.2). For the case $r = 0$, it is evident that this degenerates to the trivial 0-estimator, and one would expect to do better. But in general, Good showed that (2.2) guarantees a bias of no more than $1/n$ universally, i.e. regardless to the underlying distribution. One can show that this is not true for the empirical estimators.

The Good-Turing estimator has been studied in many forms. Beyond the bias results, there have been results on statistical efficiency, asymptotic normality, concentration, etc. Most of these results put no assumption on the underlying distribution, therefore giving great generality. However, they are often hampered by the unruliness of $\mathbb{P}$. For example, the results of McAllester and Schapire [22] are in complicated analytical expressions, and give somewhat weaker convergence than one would expect (additive, rather than multiplicative, concentration), even for the case of the missing mass.

In this chapter, we focus on the missing mass problem, i.e. $M_{n,0}$. The message here is whether or not the bias guarantee is asymptotically significant depends on how probabilities decay within $\mathbb{P}$. The 0-estimator example motivates this perspective. The decay of the probabilities, or tail behavior, turns out to determine how one has to solve the Good-Turing estimation problem, and also characterizes the Good-Turing estimator itself. Furthermore, by restricting our attention to certain types of tail behavior which are natural to expect, we can more easily analyze existing estimators, propose new estimators, and also develop a keener insight about the nature of this problem and its extension to other rare event problems. This is the topic of the next chapter.

The rest of the current chapter is organized as follows. In section 2.1 we briefly outline two perspectives for the construction of the Good-Turing estimator. In section 2.2 we present the bias and concentration properties on which we will focus for the rest of the development. In section 2.3 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 2.4.1, we use accrual rates to determine the asymptotic behavior of the expected value of the missing mass. We apply these in section 2.4.2 to describe the distribution-dependent performance of the Good-Turing estimator. Lastly we conclude in section 2.5 with descriptive and predictive consequences of our results.

■ 2.1 Construction of the Good-Turing Estimator

The original derivation of Good ([17], pp. 240–241) uses an empirical Bayes approach with a uniform prior, and presumes a finite alphabet. We can arrive to the same construction by employing an alternative perspective. In particular, assume that the probabilities $\{p_j\}_{j\in\mathbb{N}}$ are known as an unlabeled set (or rather multiset), in the sense that we know the numerical values, but not which symbols in $\mathbb{N}$ they map to. This is equivalent to knowing the distribution, but making the i.i.d. observations with an unknown arbitrary relabeling.

If we choose any of the observed symbols, say X_n without loss of generality, then the a priori probability that $X_n = j$ is p_j. However, we have the rest of the observations $X_1, \cdots, X_{n-1}$, and we would like to know the posterior probability given these. More precisely, Good's approach focuses on conditioning on the number of times the value assumed by X_n appears in the observations. If in the entirety of the observations, the value of X_n appears r times, we are then conditioning on it appearing $r-1$ times in the rest of the observations, other than X_n: the event is:

$$E_r := \left\{ \sum_{i=1}^{n-1} \mathbf{1}\{X_i = X_n\} = r-1 \right\}.$$

Note that although X_n is independent from $X_1, \cdots, X_{n-1}$, the event that X_n appears $r-1$ times in $X_1, \cdots, X_{n-1}$ is not independent from the event that X_n takes a certain value. We have:

$$\begin{aligned}
\mathbb{P}\left\{X_n = j \middle| E_r\right\} &= \frac{\mathbb{P}\{X_n = j\}\mathbb{P}\left\{\sum_{i=1}^{n-1} \mathbf{1}_{=j}(X_i) = r-1\right\}}{\sum_{j'} \mathbb{P}\{X_n = j'\}\mathbb{P}\left\{\sum_{i=1}^{n-1} \mathbf{1}_{=j'}(X_i) = r-1\right\}} \\
&= \frac{p_j \cdot \binom{n-1}{r-1} p_j^{r-1}(1-p_j)^{(n-1)-(r-1)}}{\sum_{j'} p_{j'} \cdot \binom{n-1}{r-1} p_{j'}^{r-1}(1-p_{j'})^{(n-1)-(r-1)}} \\
&= \frac{p_j^r(1-p_j)^{n-r}}{\sum_{j'} p_{j'}^r (1-p_{j'})^{n-r}}. \qquad (2.3)
\end{aligned}$$

This is a complete posterior description of the distribution of X_n given how frequently it appears in the observations. However, we still cannot evaluate or approximate it, because we don't really have the p_j values. Instead, we aggregate by taking the posterior expectation. If $X_n = j$, then its probability is $p_{X_n} = p_j$. Therefore, we have:

$$\mathbf{E}[p_{X_n} | E_r] = \sum_j p_j \mathbb{P}\left\{X_n = j \middle| E_r\right\} = \frac{\sum_j p_j^{r+1}(1-p_j)^{n-r}}{\sum_{j'} p_{j'}^r (1-p_{j'})^{n-r}}. \qquad (2.4)$$

Equation (2.4) merges back with Good's derivation ([17], equation (13)). From then on, one should only observe that the expected number of symbols appearing exactly r times out of n is:

$$\mathbf{E}[K_{n,r}] = \sum_j \binom{n}{r} p_j^r (1-p_j)^{n-r},$$

to write:

$$\mathbf{E}[p_{X_n} | E_r] = \frac{r+1}{n+1} \frac{\mathbf{E}[K_{n+1,r+1}]}{\mathbf{E}[K_{n,r}]}. \qquad (2.5)$$

This leads Good to suggest the principal form of the (per symbol) Good-Turing estimator, which does not explicitly require any knowledge of the probabilities:

$$\hat{\mathbf{E}}[p_{X_n} | E_r] = \frac{r+1}{n+1} \frac{K_{n,r+1}}{K_{n,r}}. \qquad (2.6)$$

Since this is the same value for every element of $B_{n,r} \subset \mathbb{N}$, the set of symbols appearing exactly r times, and since $|B_{n,r,}| = K_{n,r}$, Good suggests estimating the total probability of $B_{n,r}$, $M_{n,r} = \mathbb{P}\{B_{n,r}\}$ by (2.2):

$$\frac{r+1}{n+1} K_{n,r+1}.$$

We can also derive the same formula by dropping the Bayesian perspective completely, and instead looking at the true expectation of $M_{n,r}$, over the ensemble of observations, which is easily seen to satisfy:

$$\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1} \mathbf{E}[K_{n+1,r+1}].$$

The literature on the Good-Turing estimator has alternated between these perspectives (e.g. Cohen and Sackrowitz 1990). It would be interesting to reconcile the two interpretations, but this is mostly a philosophical argument.

■ 2.2 Performance Valuation

■ 2.2.1 Bias and Concentration

Consider an estimator $\hat{M}_{n,0}$ of the missing mass.

Definition 2.1. *If there exists a function $f(n)$ such that $|\mathbf{E}[\hat{M}_{n,0}] - \mathbf{E}[M_{n,0}]| = O(f(n))$, we say that $\hat{M}_{n,0}$ has asymptotic bias of order $f(n)$.*

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

Definition 2.2. *If there exist constants a and b, such that for all $\epsilon > 0$ and n we have*

$$\mathbb{P}\{\hat{M}_{n,0} < M_{n,0} - \epsilon\} < a \exp(-b\epsilon^2 n),$$

we say that $\hat{M}_{n,0}$ concentrates above $M_{n,0}$*. If this holds for $\hat{M}_{n,0} > M_{n,0} + \epsilon$, then we say $\hat{M}_{n,0}$ concentrates* below $M_{n,0}$*. If both hold, then $\hat{M}_{n,0}$ concentrates* around $M_{n,0}$.

Concentration results can have various forms of the exponent. The choice of $\epsilon^2 n$ reflects the results in the literature, such as [22] and [21], based on Chernoff-like bounds.

As mentioned in this chapter's preamble, it was known to Good that $G_{n,0}$ is asymptotically unbiased, with asymptotic bias of the order of $1/n$. More precisely: $\mathbf{E}[M_{n,0}] \leq \mathbf{E}[G_{n,0}] \leq \mathbf{E}[M_{n,0}] + \frac{1}{n}$. Later Robbins [26] showed that the mean squared error also decays faster than $\frac{1}{n+1}$. Most of the bias statements that we will make in this chapter also translate to MSE statements. Many other statistical properties of this estimator, including its asymptotic normality (Esty [12]), its admissibility with respect to mean squared error only with finite support and inadmissibility otherwise, and being a minimum variance unbiased estimator of $\mathbf{E}[M_{n-1,0}]$ (both by Cohen and Sackrowitz [9]), were shown. More recently McAllester and Schapire [22] also showed that $G_{n,0}$ concentrates above $M_{n,0}$.

Note that other estimators of $M_{n,0}$ have also been proposed, with some additional assumptions on the underlying distributions, such as a known number of species, and a uniform distribution. Nevertheless, $G_{n,0}$ performs comparably to those parametric estimators. A good survey of this literature can be found in Gandolfi and Sastri [15].

■ 2.2.2 Trivial Estimators

Let's call an estimator $\hat{M}_{n,0}$ 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. 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 once more 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 $M_{n,0}$. Furthermore, it is easy to show that $\mathbf{E}[M_{n,0}]$ converges to 0, and thus the 0-estimator is also asymptotically unbiased. One manifestation 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 or faster than what $G_{n,0}$ can guarantee, namely $1/n$. In another category, however, $G_{n,0}$ distinctly outperforms this trivial estimator. Such categories turn out to depend on how probability decays within $\mathbb{P}$.

■ 2.3 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 2.3. *We define the accrual function of a distribution $\mathbb{P}$ as:*

$$A(x) = \sum_{p_i \le x} p_i.$$

Note that $A(x)$ 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.

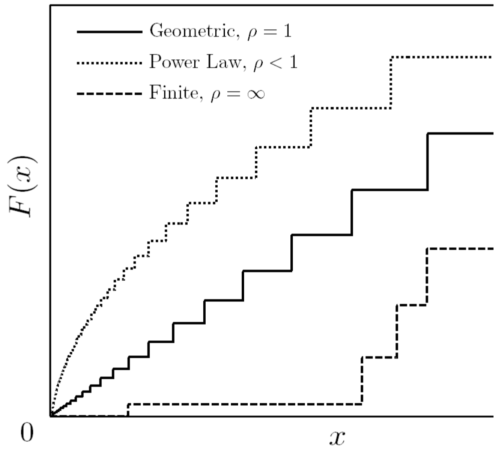

Figure 2.1. Accrual functions for geometric, power law, and finite distributions.

Definition 2.4. *Let the distribution $\mathbb{P}$ have an accrual function $A(x)$. We define the lower and upper accrual rates of $\mathbb{P}$ respectively as:*

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

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

We illustrate this with three examples, as shown in Figure 2.3. 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 $A(x) = x/q$, 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 $A(x) = \sum 6/(\pi i)^2$, for $i \geq \pi^{-1}\sqrt{6/x}$. Near $x = 0$, we can approximate the sum with an integral, and find $A(x) \approx C \cdot \sqrt{x}$. Therefore, the accrual rate is 1/2.

These examples suggest a taxonomy that relates accrual rate to conventional notions of decay. First, the finite support case is characterized, formally, by an infinite rate. When the rate is bounded, by construction, it cannot be above 1. Exponentially decaying tails have an accrual rate of one, but so do tails that behave, for example, roughly as e^{-j^β} for some $\beta \neq 1$. Below one, we have a distinct regime where the decay is roughly a power law. When we say 'roughly', we mean up to the logarithmic ratio inherent in the definition of accrual rate. It is interesting that one arrives at these notions, without an explicit indexing in the definition of the accrual function, although in the next chapter we shall see that a stronger characterization of the accrual is also equivalent to an indexed decay.

2.4 Distribution Dependence

2.4.1 Behavior of $\mathbf{E}[M_{n,0}]$

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

$$(1-x)^n A(x) \leq \mathbf{E}[M_{n,0}] \leq (1-y)^n + A(y). \tag{2.7}$$

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 $M_{n,0} = \sum_i p_i Y_i$ and consequently that:

$$\mathbf{E}[M_{n,0}] = \sum_i p_i \mathbf{E}[Y_i] = \sum_i p_i (1-p_i)^n.$$

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

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

And for the upper bound:

$$\mathbf{E}[M_{n,0}] \quad \leq \quad \sum_{p_i \leq y} p_i \cdot 1 + (1-y)^n.$$

The lemma then follows from the definition of A. □

Theorem 2.2. *Let $\mathbb{P}$ have lower and upper accrual rates $0 < \underline{\rho} \leq \overline{\rho} < \infty$. Then for every $\delta > 0$ there exists n_0 such that for all $n > n_0$ we have:*

$$n^{-(\overline{\rho}+\delta)} \leq \mathbf{E}[M_{n,0}] \leq n^{-(\underline{\rho}-\delta)}$$

or, equivalently, for every $\delta > 0$ we have that $\mathbf{E}[M_{n,0}]$ 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 $\log A(x)/\log x \leq \alpha$, or alternatively $A(x) \geq x^\alpha$.

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 $A(x)$ together with the left-hand inequality in (2.7) to write:

$$\begin{array}{rcl}\mathbf{E}[M_{n,0}] & \geq & (1-\overline{x})^n A(\overline{x}) \\ & \geq & (1-\overline{x})^n \overline{x}^\alpha = \left(1 - \frac{\alpha}{n+\alpha}\right)^n \frac{\alpha^\alpha}{(n+\alpha)^\alpha} \\ & \geq & e^{-\alpha}\alpha^\alpha \frac{1}{(n+\alpha)^\alpha},\end{array}$$

therefore there exists n_1, such that for all $n > n_1$ we have $\mathbf{E}[M_{n,0}] \geq n^{-(\overline{\rho}+\delta)}$.

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 $\log A(x)/\log x \geq \beta$, or alternatively $A(x) \leq x^\beta$.

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 $e^z \geq 1 + z$ with $z = \frac{\beta}{n} \log \frac{\beta}{n}$ we can show that $\underline{x} \leq \frac{\beta}{n} \log \frac{n}{\beta}$, and thus for n large enough we will have $\underline{x} < x_2$. Using the bound for $A(x)$ with the right-hand inequality in (2.7), we can write:

$$\begin{aligned} \mathbf{E}[M_{n,0}] &\leq (1-\underline{x})^n + A(\underline{x}) \\ &\leq (1-\underline{x})^n + \underline{x}^\beta = \left(\frac{\beta}{n}\right)^\beta + \left(1 - \left(\frac{\beta}{n}\right)^{\frac{\beta}{n}}\right)^\beta \\ &\leq \left(\frac{\beta}{n}\right)^\beta + \left(\frac{\beta}{n} \log \frac{n}{\beta}\right)^\beta, \end{aligned}$$

therefore there exists n_2, such that for all $n > n_2$ we have $\mathbf{E}[M_{n,0}] \leq n^{-(\underline{\rho}-\delta)}$. Finally, set $n_0 = n_1 \vee n_2$. □

■ 2.4.2 Performance of $G_{n,0}$

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 2.3. *Let $r \in [0,1]$ be given. Then there exists a trivial estimator, namely $\hat{M}_{n,0} = 1/n^r$, 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 $\overline{\rho} < r$ such that for all n_0 and $c > 0$, there exists $n > n_0$ where the bias is larger than c/n^r. Therefore the given trivial estimator does not have asymptotic bias of order $1/n^r$.

Proof. Consider the forward statement. Set $\hat{M}_{n,0} = 1/n^r$, and let $\mathbb{P}$ be any distribution such that $\underline{\rho} > r$. From Theorem 2.2, we know that for large enough n we have $\mathbf{E}[M_{n,0}] \leq 1/n^r$. It immediately follows that $0 < \hat{M}_{n,0} - \mathbf{E}[M_{n,0}] < 1/n^r$, demonstrating the bias claim.

For the converse, let's take $c = 1$ without loss of generality. 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 $|\mathbf{E}[M_{n,0}] - \hat{M}_{n,0}| < 1/n^r$.

Now, pick two distributions. First, let $\mathbb{P}$ be such that $0 < \underline{\rho} \leq \overline{\rho} < r$. 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:

- $\mathbf{E}[M_{n,0}] - \hat{M}_{n,0} < 1/n^r$ for $\mathbb{P}'$ by our assumption,
- $\mathbf{E}[M_{n,0}] > 1/n^t$ for $\mathbb{P}'$ by Theorem 2.2,
- $n^{s-t} > 2$, and $n^{r-s} > 1$.

In particular, we get $\hat{M}_{n,0} > \mathbf{E}[M_{n,0}] - 1/n^r > 1/n^t - 1/n^r$.

Now, moving on to $\mathbb{P}$, choose $n_1 > n_0$ such that $\mathbf{E}[M_{n,0}] < 1/n^s$, by Theorem 2.2. Then, for all $n > n_1$ we have:

$$\begin{array}{rcl} \hat{M}_{n,0} - \mathbf{E}[M_{n,0}] & > & 1/n^t - 1/n^s - 1/n^r \\ & = & \left(n^{r-s}(n^{s-t} - 1) - 1\right)/n^r > 1/n^r. \end{array}$$

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

This immediately results in the following corollaries.

Corollary 2.4. *If $\mathbb{P}$ is finite, i.e. has infinite accrual rate, then the trivial estimator $\hat{M}_{n,0} = 1/n$ matches the performance of the Good-Turing estimator asymptotically.*

Proof. The bias claim follows from the forward part of Theorem 2.3, by setting $r = 1$. For the concentration result, first note that from Theorem 2.2 we know that for large enough n, we have $\mathbf{E}[M_{n,0}] \leq 1/n = \hat{M}_{n,0}$, then invoke the fact that $M_{n,0}$ concentrates around its own mean. Namely, there exist constants a and b such that for every $\epsilon > 0$ and n we have:

$$\mathbb{P}\{\mathbf{E}[M_{n,0}] < M_{n,0} - \epsilon\} < a \exp(-b\epsilon^2 n), \tag{2.8}$$

This was shown first by McAllester and Schapire [22], and later with tighter constants (in addition to concentration from below) by McAllester and Ortiz [21]. From here, one needs only to observe that for large enough n, an event of the form $\{\hat{M}_{n,0} < Z\}$ is a subset of event $\{\mathbf{E}[M_{n,0}] < Z\}$, and particularly

$$\mathbb{P}\{\hat{M}_{n,0} < M_{n,0} - \epsilon\} \leq \mathbb{P}\{\mathbf{E}[M_{n,0}] < M_{n,0} - \epsilon\},$$

and asymptotic concentration follows from (2.8). □

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

Corollary 2.5. *If $\mathbb{P}$ has accrual rate 1, then the trivial estimator $\hat{M}_{n,0} = 1/n$ has an asymptotic bias of order $L(n)/n$, and the concentration property:*

$$\mathbb{P}\{L(n)\hat{M}_{n,0} < M_{n,0} - \epsilon\} < a \exp(-b\epsilon^2 n), \tag{2.9}$$

where $L(n)$ is a slowly varying function, i.e. $L(cn)/L(n) \to 1$ for all $c > 0$, as $n \to \infty$.

The proof of this corollary follows the same line of argument as the previous one. Examples of $L(n)$ can vary from constants to iterated and poly logarithms, and are functions that grow slower than any finite power. For instance, one can show that for geometric distributions $L(n)$ can be chosen to be a constant (universally for all q smaller than some q_0), or $\log n$ (universally over all geometric distributions). The description of Corollary 2.5 is coarser than, but alludes to, the regular variation context which we elaborate in Chapter 3.

It is not surprising that Good-Turing estimation does not have asymptotic advantage for finite distributions. However, as we illustrated in section 2.3, an accrual rate of 1 is characteristic of exponential-like light tails. Distributions in many applications fall under this category, and it is instructive to see that the (asymptotic) advantage in using the Good-Turing estimator in such situations is marginal.

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 $\mathbf{E}[M_{n,0}]$, and uses it as a trivial estimator, then bias would be zero, and concentration would follow from (2.8). Of course, this would constitute much more than a partial knowledge about accrual rate. In fact, the converse part of Theorem 2.3 formalizes how lack of such precise knowledge dooms trivial estimators to failure in this case. We restate this as a corollary.

Corollary 2.6. *Given any trivial estimator $\hat{M}_{n,0}$, there is always some distribution $\mathbb{P}$ with accrual rates less than* 1*, for which $\hat{M}_{n,0}$ fails to match the performance of the Good-Turing estimator.*

Accrual rates above one are characteristic of heavy tails. We have thus shown that here, in contrast to the light tail case, one cannot construct a single trivial estimator that works as well as the Good-Turing estimator, without being significantly penalized. Therefore, the Good-Turing estimator presents a distinct advantage in this situation.

Remark

One may argue that the comparison we have presented in this chapter is not entirely fair. More precisely, we are comparing the Good-Turing estimator's bias that is *(a) worst-case* over all underlying distributions and *(b) non-asymptotic*, to the bias of a trivial estimator that is *(a) specific* to all distributions with a range of accrual rates and *(b) asymptotic.* This is certainly a valid argument, and shows that no trivial estimator can universally replace the Good-Turing estimator. However, our exposition is simply an analytic tool to probe the significance of the performance guarantees of the Good-Turing estimator. In particular, we have shown that with heavy-tailed distributionw the Good-Turing estimator triumphs despite this unfair comparison, which is indicative that this is the regime where one can effectively learn rare probabilities. In the next chapter, Chapter 3, we shift our attention to a different performance metric, consistency, and show more explicitly the importance and role of heavy tails.

■ 2.5 Summary

In this chapter, 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 of one or infinity, and those with accrual rates less than one. For the first, we showed that the performance of the Good-Turing estimator can be rivaled by a trivial estimator, and thus Good-Turing estimation offers no great 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 non-trivial.

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 [33]: "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. This turns out to be closely associated with regular variation, as we will highlight in the next chapter, Chapter 3. 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 [14], 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.

Chapter 3

Probability Estimation under Regular Variation

THIS chapter moves the focus entirely to heavy tailed distributions. We use Karamata's theory of regular variation to more precisely describe such distributions. Also, whereas the previous chapter focuses on bias and concentration properties of the Good-Turing estimator, this chapter studies consistent estimation of the probability of rare events, in the sense of a ratio converging to one, probabilistically. In particular we ask for strong consistency, where this convergence holds almost surely. We first show that in general the Good-Turing estimator is not consistent. However, under regular variation, we show that the probabilities themselves obey a strong law. We do so by extending the McAllester-Ortiz additive concentration results for the missing mass to all of the rare probabilities, and then using the regular variation property to show multiplicative concentration. This is a more natural and useful property, since all quantities of interest converge to zero. The strong laws allow us to show that Good-Turing estimation is consistent in this regime. Furthermore, we construct new families of estimators that address some of the other shortcomings of the Good-Turing estimator. For example, they perform smoothing implicitly. This framework is a close parallel to extreme value theory, and many of the techniques therein can be adopted into the model set forth in this thesis.

The rest of this chapter is organized as follows. In 3.1, we give the basic definition of consistency and show the failure of Good-Turing to be consistent for geometric distributions. In 3.2, we give the preliminary background for this development, including the notion of regular variation, an exposition to the exponential moment method, and the property of negative association. Using these tools, in 3.3 we extend the additive concentration results of McAllester, Schapire and Ortiz [21, 22] to all rare probabilities. We then use a parallel approach in 3.4 to derive multiplicative concentration under regular variation. Finally, in 3.5, we use these results to obtain strong laws, to show the consistency of Good-Turing estimation under heavy-tailed regular variation, and to construct a family of new consistent estimators.

Notation

Throughout the chapter, it is convenient to use the limiting notation $f \sim g$ to mean $f/g \to 1$. We also use the subscript $_{\text{a.s.}}$ to indicate almost sure convergence when quantities are random.

3.1 Consistency of Probability Estimation

3.1.1 Definition

We say that an estimator $\hat{M}_{n,r}$ of $M_{n,r}$ is consistent if $\hat{M}_{n,r}/M_{n,r} \to 1$ in probability. We say that it is strongly consistent if $\hat{M}_{n,r}/M_{n,r} \to 1$ almost surely.

3.1.2 The Good-Turing Estimator is not Universally Consistent

Consider a geometric distribution given by $p_j = (1-q)q^j$ for $j \in \mathbb{N}_0$, parametrized by $q \in (0,1)$. We can show the following precise behavior for the counts of symbols seen exactly once.

Lemma 3.1. *For the subsequence* $n_i = \lfloor c/p_i \rfloor = \lfloor cq^{-i}/(1-q) \rfloor$*, with* $c > 0$*, we have:*

$$\mathbf{E}[K_{n_i,1}] \to h(c,q), \tag{3.1}$$

where

$$h(c,q) := \sum_{m=-\infty}^{\infty} cq^m e^{-cq^m}. \tag{3.2}$$

Proof. In general, by Poissonization (e.g. [16], Lemma 1) or using $\left(1 - \frac{s}{n}\right)^n \uparrow e^{-s}$ and the dominated convergence theorem, one has that as $n \to \infty$:

$$\left| \sum_{j=0}^{\infty} np_j(1-p_j)^n - \sum_{j=0}^{\infty} np_j e^{-np_j} \right| \to 0. \tag{3.3}$$

This limit is not in general a bounded constant. It can grow unbounded, or can be bounded but oscillatory. However, in the geometric case, we can obtain a bounded constant limit by restricting our attention to a subsequence, such as the one we hypothesized, $n_i = \lfloor c/p_i \rfloor$.

Assume that there exists a rather convenient sequence $j_i \to \infty$ slow enough such that $i - j_i \to \infty$ and

$$\sum_{j=0}^{j_i} c\frac{p_j}{p_i} e^{-c\frac{p_j}{p_i}} \to 0. \tag{3.4}$$

This gives us enough approximation leeway to show first that we can replace n_i by c/p_i in Equation (3.3), without altering the limit:

$$\left| \sum_{j=0}^{\infty} n_i p_j e^{-n_i p_j} - \sum_{j=0}^{\infty} c\frac{p_j}{p_i} e^{-c\frac{p_j}{p_i}} \right| \quad \leq \quad \sum_{j=0}^{\infty} \left| \frac{\lfloor \frac{c}{p_i} \rfloor}{\frac{c}{p_i}} e^{\left(\frac{c}{p_i} - \lfloor \frac{c}{p_i} \rfloor\right) p_j} - 1 \right| c\frac{p_j}{p_i} e^{-c\frac{p_j}{p_i}}$$

$$
\begin{aligned}
\leq \quad & (e+1)\sum_{j=0}^{j_i} c\frac{p_j}{p_i}e^{-c\frac{p_j}{p_i}} \qquad \to 0 \\
& + \left(\frac{p_i}{c} \vee (e^{p_{j_i}}-1)\right) \sum_{j=j_i+1}^{\infty} c\frac{p_j}{p_i}e^{-c\frac{p_j}{p_i}} \qquad \to 0,
\end{aligned}
$$

and second that we can remove the dependence on i from the limit, using the fact that:

$$
\begin{aligned}
\sum_{j=0}^{\infty} c\frac{p_j}{p_i}e^{-c\frac{p_j}{p_i}} &= \sum_{j=j_i}^{\infty} c\frac{p_j}{p_i}e^{-c\frac{p_j}{p_i}} \\
&= \sum_{j=j_i}^{\infty} cq^{-(j-i)}e^{-cq^{-(j-i)}} \\
&= \sum_{m=-\infty}^{i-j_i} cq^m e^{-cq^m} \to \sum_{m=-\infty}^{\infty} cq^m e^{-cq^m}.
\end{aligned}
$$

Therefore, to complete the proof, we construct j_i as desired. In particular, let $j_i = i - \left\lceil \log_{q^{-1}}\left(\frac{2}{c}\log\frac{1}{p_i}\right)\right\rceil$. First note that the subtracted term is of the order of $\log i$, and thus $j_i \to \infty$ yet $i - j_i \to \infty$. Then, by the fact that $\sum p_j = 1$ and $e^{-c\frac{p_{j_i}}{p_i}}$ is the largest of the lot since p_{j_i} is the smallest, we have:

$$
\sum_{j=0}^{j_i} c\frac{p_j}{p_i}e^{-c\frac{p_j}{p_i}} \leq \frac{c}{p_i}e^{-c\frac{p_{j_i}}{p_i}} = \frac{c}{p_i}e^{-cq^{-\left\lceil \log_{q^{-1}}\left(\frac{2}{c}\log\frac{1}{p_i}\right)\right\rceil}} \leq \frac{c}{p_i}e^{-cq^{-\log_{q^{-1}}\left(\frac{2}{c}\log\frac{1}{p_i}\right)}} = cp_i \to 0.
$$

□

Recalling that $\mathbf{E}[K_{n,1}] = n\mathbf{E}[M_{n,0}]$, note that this agrees with the results of the previous chapter, in particular Corollary 2.5. Using this more precise characterization over a subsequence, however, we can establish the following negative result.

Lemma 3.2. *For a geometric distribution with small enough q, the Good-Turing estimator of the missing mass is not consistent.*

Proof. For any real-valued non-negative random variable W, we can use Markov's inequality to establish that:

$$
\begin{aligned}
\mathbb{P}\left\{W < \frac{\mathbf{E}[W]}{1-\eta}\right\} &= 1 - \mathbb{P}\left\{W \geq \frac{\mathbf{E}[W]}{1-\eta}\right\} \\
&\geq 1 - \frac{\mathbf{E}[W]}{\frac{\mathbf{E}[W]}{1-\eta}} = \eta.
\end{aligned}
$$

Let's assume that for some $c > 0$ and $q_0 > 0$, and for all $0 < q < q_0$, we have $h(c,q) < 1$. Choose any such q, then choose any $\eta \in (0, 1-h(c,q))$, and let i_0 be large enough such that for all $i > i_0$ we have $\mathbf{E}[K_{n_i,1}] < 1 - \eta$. Since $K_{n_i,1}$ takes integer values, it follows that for all $i > i_0$:

$$\mathbb{P}\left\{K_{n_i,1} < \frac{\mathbf{E}[K_{n,1}]}{1-\eta}\right\} = \mathbb{P}\left\{K_{n_i,1} = 0\right\} \geq \eta.$$

This means that there's always a positive, bounded away from zero, probability that $K_{n_i,1} = 0$, implying $G_{n_i,0} = 0$. Since $M_{n_i,0} > 0$ for every sample, it follows that with positive probability $G_{n_i,0}/M_{n_i,0} = 0$, and therefore $G_{n,0}/M_{n,0} \nrightarrow_p 1$, for any geometric distribution with $q < q_0$.

To complete the proof, we show that our assumption about h is true. We could argue for it abstractly, but we give a concrete bound instead. In particular, using the fact that $xe^{-x} < \min\{x, 1/x\}$ for all $x > 0$, we have

$$\begin{aligned} h(c,q) &= \sum_{m=-\infty}^{\infty} cq^m e^{-cq^m} \\ &< \sum_{m=-\infty}^{-1} 1/(cq^m) + ce^{-c} + \sum_{m=1}^{\infty} cq^m \\ &= ce^{-c} + \left(\frac{1}{c} + c\right)\frac{q}{1-q} \end{aligned}$$

Let $c = 1$ and $q_0 = (1 - e^{-1})/(3 - e^{-1})$. Then it is easy to verify that $ce^{-c} + \left(\frac{1}{c} + c\right)\frac{q}{1-q}$ is continuous, monotonically increasing, and at q_0 it evaluates to 1. Therefore, for all $q < q_0$, we have that $h(1,q) < 1$ as desired. □

This shows that even in the case of accrual rate 1, and a fairly well behaved distribution like the geometric, the Good-Turing estimator does not result in a consistent estimator for the missing mass. This motivates us to ask whether the same holds true for accrual rates less than 1. As we show next, it turns out that heavy tailed distributions are much better behaved, especially if we make their description more precise by using regular variation.

■ 3.2 Preliminaries

■ 3.2.1 Regularly Varying Distributions

In this section we refine the notion of accrual rate, by using Karamata's theory of regular variation, which was developed originally in [18], with the standard reference being Bingham, Goldie and Teugels [4]. The application we have here is based on the early work of Karlin [19], which was recently given an excellent exposition by Gnedin, Hansen, and Pitman [16]. We follow the notational convention of the latter.

It is first useful to introduce the following counting measure on $[0,1]$:

$$\nu(\mathrm{d}x) := \sum_j \delta_{p_j}(\mathrm{d}x), \tag{3.5}$$

where δ_x is a Dirac delta at x.

Using ν, we define the following function, which was used originally by Karlin [19] to define what is meant by a regularly varying distribution, and which is a cumulative count of all symbols having no less than a certain probability mass:

$$\vec{\nu}(x) := \nu[x, 1]. \tag{3.6}$$

We also define the following family of measures, parametrized by $r = 1, 2, \cdots$:

$$\nu_r(\mathrm{d}x) := x^r \nu(\mathrm{d}x) = \sum_j p_j^r \delta_{p_j}(\mathrm{d}x). \tag{3.7}$$

In particular, note that the accrual function can be written in this notation as $A(x) = \nu_1[0, x]$. Karlin did not use the accrual function for his definition, but we see in Theorem 3.3 that there's an equivalent definition based on $A(x)$.

Definition 3.1. *Following Karlin [19], we say that* $\mathbb{P}$ *is* regularly varying *with* regular variation index $\alpha \in (0,1)$*, if the following holds:*

$$\vec{\nu}(x) \sim x^{-\alpha}\ell(1/x), \quad \text{as} \quad x \downarrow 0, \tag{3.8}$$

where $\ell(t)$ is a slowly varying function, in the sense that for all $c > 0$, $\ell(ct)/\ell(t) \to 1$ as $t \to \infty$.

It is possible to study the cases $\alpha = 0$ and $\alpha = 1$ too, with additional care. However, we do not tackle these situations in this thesis.

We now give the very useful characterizations which were established to be equivalent to Definition 3.1 by Gnedin et al. [16], in the following portmanteau theorem.

Theorem 3.3. *Equation* (3.8) *is equivalent to any (and therefore all) of the following deterministic conditions:*

Accrual function:

$$A(x) = \nu_1[0, x] \sim \frac{\alpha}{1-\alpha} x^{1-\alpha}\ell(1/x), \quad \text{as} \quad x \downarrow 0, \tag{3.9}$$

Expected number of distinct observed symbols:

$$\mathbf{E}[K_n] \sim \Gamma(1-\alpha) n^{\alpha}\ell(n), \quad \text{as} \quad n \to \infty, \tag{3.10}$$

Expected number of symbols observed exactly once:

$$\mathbf{E}[K_{n,1}] \sim \alpha\Gamma(1-\alpha) n^{\alpha}\ell(n), \quad \text{as} \quad n \to \infty, \tag{3.11}$$

We also have the following probabilistic equivalent conditions:

Number of distinct observed symbols:

$$K_n \sim_{\text{a.s.}} \Gamma(1-\alpha) n^{\alpha} \ell(n), \quad \text{as} \quad n \to \infty, \tag{3.12}$$

Number of symbols observed exactly once:

$$K_{n,1} \sim_{\text{a.s.}} \alpha\Gamma(1-\alpha) n^{\alpha} \ell(n), \quad \text{as} \quad n \to \infty, \tag{3.13}$$

Finally any of the above implies the following for all $r > 1$:

$$\nu_r[0,x] \sim \frac{\alpha}{1-\alpha} x^{r-\alpha} \ell(1/x), \quad \text{as} \quad x \downarrow 0, \tag{3.14}$$

$$\mathbf{E}[K_{n,r}] \sim \frac{\alpha\Gamma(r-\alpha)}{r!} n^{\alpha} \ell(n), \quad \text{as} \quad n \to \infty, \tag{3.15}$$

$$K_{n,r} \sim_{\text{a.s.}} \frac{\alpha\Gamma(r-\alpha)}{r!} n^{\alpha} \ell(n), \quad \text{as} \quad n \to \infty, \tag{3.16}$$

It is worth noting that Equation (3.9) implies the accrual rate condition of the previous chapter, namely Definition 2.4, with $\rho = 1 - \alpha$. However, the converse is not necessarily true.

Theorem 3.3 shows how the regularly varying case is very well behaved, especially in terms of the strong laws on the occupancy numbers $K_{n,r}$, which are the elementary quantities for Good-Turing estimation. In fact, from Equations (3.11), (3.13), (3.15), and (3.16), we have that for all $r \geq 1$, $K_{n,r}/\mathbf{E}[K_{n,r}] \to_{\text{a.s.}} 1$. We will harness this fact throughout this chapter to study probability estimation in this regime.

3.2.2 Exponential Moment Method and the Gibbs Variance Lemma

We adhere closely to the exposition of McAllester and Ortiz [21]. The exponential moment method takes its name from the use of Markov's inquality with an exponentiated random variable. It is embodied in the following statement, due originally to Chernoff [8].

Theorem 3.4. *Let W be a real-valued random variable with finite mean $\mathbf{E}[W]$. Associate with W the function $S : \mathbb{R}^+ \to \mathbb{R}$, $w \mapsto S(W,w)$, where:*

$$S(W,w) = \sup_{\beta \in R} w\beta - \log Z(W,\beta),$$

with

$$Z(W,\beta) = \mathbf{E}[e^{\beta W}].$$

Then, the lower and upper deviation probabilities are given by:

$$\mathbb{P}\{W > \mathbf{E}[W]| + \epsilon\} \leq e^{-S(W,\mathbf{E}[W]+\epsilon)},$$

and

$$\mathbb{P}\{W < \mathbf{E}[W] - \epsilon\} \leq e^{-S(W,\mathbf{E}[W]-\epsilon)}.$$

We now give some background about Gibbs measures, which have distinct roots in statistical mechanics, but are also an integral part of the exponential moment method. It turns out that for any W, there's always a largest open interval (β_-, β_+) over which $Z(W, \beta)$ is finite. In this exposition this interval is always the whole real line. Denote the law of W by μ, then with each $\beta \in (\beta_-, \beta_+)$, we can associate the *Gibbs measure*:

$$\mu_\beta(\mathrm{d}w) = \frac{e^{\beta w}}{\mathbf{E}[e^{\beta W}]}\mu(\mathrm{d}w).$$

Denote by $\mathbf{E}_\beta$ any expectation carried out with respect to the new measure. In particular denote the variance of W under the Gibbs measure by $\sigma^2(W, \beta) := \mathbf{E}_\beta[(W - \mathbf{E}_\beta[W])^2]$. Note that $\mathbf{E}_\beta[W]$ is continuous and monotonically increasing as β varies in (β_-, β_+). Denote its range of values by (w_-, w_+), and let $\beta(w)$, for any $w \in (w_-, w_+)$ refer to the unique value $\beta(\beta_-, \beta_+)$ satisfying $\mathbf{E}_\beta[W] = w$.

McAllester and Ortiz [21] distill a particular result out of Chernoff's original work, and dub it the Gibbs variance lemma.

Lemma 3.5. *For any $w \in (w_-, w_+)$ and $\beta \in (\beta_-, \beta_+)$, we have that:*

$$\begin{aligned} S(W, w) &= w\beta(w) - \log Z(W, \beta(w)) \\ &= D\left(\mu_{\beta(w)} \| \mu\right) \\ &= \int_{\mathbf{E}[W]}^{w} \int_{\mathbf{E}[W]}^{v} \frac{1}{\sigma^2(W, \beta(u))} \mathrm{d}u\mathrm{d}v \qquad (3.17) \\ \log(Z(W, \beta)) &= \mathbf{E}[W]\beta + \int_0^\beta \int_0^\alpha \sigma^2(W, \gamma)\mathrm{d}\gamma\mathrm{d}\alpha \qquad (3.18) \end{aligned}$$

The importance of this lemma is that it showcases how we can establish concentration by controlling the variance $\sigma^2(W, \beta)$ in (3.17) and (3.18). The following two lemmas (reproducing Lemmas 9 and 11 in [21], with the exception of part *(ii)* below) are established by doing precisely that.

Lemma 3.6. *Let W be an arbitrary real-valued random variable.*

(i) If for a $\overline{\beta} \in (0, \infty]$, we have $\sup_{0 \leq \beta \leq \overline{\beta}} \sigma^2(W, \beta) \leq \overline{\sigma}^2$, then for all $\epsilon \in [0, \overline{\beta}\overline{\sigma}^2]$:

$$S(W, \mathbf{E}[W] + \epsilon) \geq \frac{\epsilon^2}{2\overline{\sigma}^2}. \qquad (3.19)$$

(ii) If for a $\underline{\beta} \in [-\infty, 0)$, we have $\sup_{\underline{\beta} \leq \beta \leq 0} \sigma^2(W, \beta) \leq \underline{\sigma}^2$, then for all $\epsilon \in [0, -\underline{\beta}\underline{\sigma}^2]$:

$$S(W, \mathbf{E}[W] - \epsilon) \geq \frac{\epsilon^2}{2\underline{\sigma}^2}. \qquad (3.20)$$

We can specialize part *(ii)* to the following case:

Lemma 3.7. *If $W = \sum_j b_j W_j$, where $b_j > 0$ and W_j are independent Bernoulli with parameter q_j, then for all $\epsilon > 0$, we have:*

$$S(W, \mathbf{E}[W] - \epsilon) \geq \frac{\epsilon^2}{2\sum_j b_j^2 q_j}. \tag{3.21}$$

■ 3.2.3 Negative Association

We now introduce the concept of negatively associated random variables. We start with the definition, then give a few lemmas that facilitate establishing the property. Finally we illustrate the usefulness of this concept within the framework of the exponential moment method. All these statements and their proofs can be found in the exposition by Dubhashi and Ranjan [11], and are also outlined in McAllester and Ortiz [21]. We present the definitions and results in terms of finite collections of random variables, but everything extends to countable collections with some additional care.

Definition 3.2. *Real-valued random variables $W_1, \cdots, W_k$ are said to be* negatively associated*, if for any two disjoint subsets A and B of $\{1, \cdots, k\}$, and any two real-valued functions $f : \mathbb{R}^{|A|} \to \mathbb{R}$, and $g : \mathbb{R}^{|B|} \to \mathbb{R}$ that are both either coordinatewise non-increasing or coordinatewise non-decreasing, we have:*

$$\mathbf{E}[f(W_A) \cdot g(W_B)] \leq \mathbf{E}[f(W_A))] \cdot \mathbf{E}[g(W_B)].$$

We now show one way to construct new negatively associated random variables, starting from an existing collection.

Lemma 3.8. *If $W_1, \cdots, W_k$ are negatively associated, and $f_1, \cdots, f_k$ are real-valued functions on the real line, that are either all non-increasing or all non-decreasing, then $f_1(W_1), \cdots, f_k(W_k)$ are also negatively associated.*

The elementary negatively associated random variables in our context are the counts of each particular symbol, or equivalently the components of the empirical measure.

Lemma 3.9. *Let $\mathbb{P} = (p_1, \cdots, p_k)$ define a probability distribution on $\{1, \cdots, k\}$. Let $X_1, \cdots, X_n$ be independent samples from $\mathbb{P}$, and define, for each $j \in \{1, \cdots, k\}$:*

$$C_{n,j} := \sum_{i=1}^{n} \mathbf{1}\{X_i = j\}.$$

Then the random variables $C_{n,1}, \cdots, C_{n,k}$ are negatively associated.

The reason why negative association is useful is the following lemma, which shows that sums of negatively associated random variables can be treated just like the sum of independent random variables with identical marginals, for the purpose of the exponential moment method.

Lemma 3.10. *Let $W = \sum_{j=1}^{k} W_j$, where $W_1, \cdots, W_k$ are negatively associated. If $\tilde{W}_1, \cdots, \tilde{W}_k$ are independent real-valed random variables such that for each $j \in \{1, \cdots, k\}$ we have $\mathcal{L}(\tilde{W}_j) = \mathcal{L}(W_j)$, then for all w, we have:*

$$S(W, w) \geq S(\tilde{W}, w),$$

where $\tilde{W} = \sum_{j=1}^{k} \tilde{W}_j$.

It is worth noting that this approach is not very different from the Poissonization technique used by Karlin [19], Gnedin et al [16], and others who have studied the occupancy problem. Instead of randomizing the sampling epochs to make counts independent, which creates independence at the cost of distorting the binomial distributions into Poisson distributions, the negative association method enforces only independence. Of course, just like de-Poissonization which allows one to reconstruct results in terms of the original variables, here too we need such inversion theorems, and Lemma 3.10 does precisely that.

■ 3.3 Additive Concentration

In this section, we follow the methodology of McAllester and Ortiz [21] in order to extend their additive concentration results for the missing mass $M_{n,0}$, to all of $M_{n,r}$. For this, we use the exponential moment method and the Gibbs variance lemma, and negative association. These results are valid for all distributions, whereas the next section we specialize to regularly varying distributions, and show multiplicative concentration. The main theorem of this section is as follows.

Theorem 3.11. *Consider an arbitrary $\mathbb{P}$. Then, for every $r = 0, 1, 2, \cdots$, there exist constants $a_r, b_r, \epsilon_r > 0$ such that for every $n > 2r$, and for all $0 < \epsilon < \epsilon_r$, we have*

$$\mathbb{P}\left\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\right\} \leq a_r e^{-b_r \epsilon^2 n}. \tag{3.22}$$

We cannot directly parallel the McAllester-Ortiz [21] proofs of additive concentration for the missing mass, because unlike $M_{n,0}$, $M_{n,r}$ cannot be expressed as a sum of negatively associated random variables. However, we can work instead with a quantity that can be expressed as such, namely the total probability of all symbols appearing no more than r times:

$$M_{n,0\to r} := \sum_{k=0}^{r} M_{n,k}. \tag{3.23}$$

Indeed, we can express it as follows:

$$M_{n,0\to r} = \sum_{j} p_j Z_{n,j,r}, \tag{3.24}$$

where $Z_{n,j,r} = \mathbf{1}\{C_{n,j} \leq r\}$ are indicator random variables associated with each symbol, in order to contribute its probability mass only when it appears no more than r times in the observation. Note that each $Z_{n,j,r}$ is a non-increasing function of the corresponding count $C_{n,j}$. Since $\{C_{n,j}\}_{j\in\mathbb{N}}$ are negatively associated by Lemma 3.9, then by Lemma 3.8 so are $\{Z_{n,j,r}\}_{j\in\mathbb{N}}$. The following lemma shows that it's sufficient to prove additive concentration for every $M_{n,0\to r}$, in order to prove Theorem 3.11.

Lemma 3.12. *If for every $r = 0, 1, 2, \cdots$, there exist constants $\tilde{a}_r, \tilde{b}_r, \tilde{\epsilon}_r > 0$ such that for every $n > 2r$, and for all $0 < \epsilon < \tilde{\epsilon}_r$, we have*

$$\mathbb{P}\left\{|M_{n,0\to r} - \mathbf{E}[M_{n,0\to r}]| > \epsilon\right\} \leq \tilde{a}_r e^{-\tilde{b}_r \epsilon^2 n}. \tag{3.25}$$

then for every $r = 0, 1, 2, \cdots$, there exist constants $a_r, b_r, \epsilon_r > 0$ such that for every $n > 2r$, and for all $0 < \epsilon < \epsilon_r$ the concentration (3.22) *holds.*

Proof. Define the events:

$$A = \{\mathbf{E}[M_{n,0\to r}] - \epsilon/2 \leq M_{n,0\to r} \leq \mathbf{E}[M_{n,0\to r}] + \epsilon/2\}, \quad \text{and}$$

$$B = \{\mathbf{E}[M_{n,0\to r-1}] - \epsilon/2 \leq M_{n,0\to r-1} \leq \mathbf{E}[M_{n,0\to r-1}] + \epsilon/2\}.$$

Then $A \cap B \subset \{\mathbf{E}[M_{n,r}] - \epsilon \leq M_{n,r} \leq \mathbf{E}[M_{n,r}] + \epsilon\}$, and thus:

$$\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\} \subset A^c \cup B^c.$$

Therefore we can use our hypothesis and the union bound to write that for every $n > 2r$, $0 < \epsilon < \tilde{\epsilon}_r \wedge \tilde{\epsilon}_{r-1}$:

$$\begin{aligned}
\mathbb{P}\left\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\right\} &\leq \mathbb{P}(A^c \cup B^c) \\
&\leq \mathbb{P}(A^c) + \mathbb{P}(B^c) \\
&\leq \tilde{a}_r e^{-\tilde{b}_r \epsilon^2 n/4} + \tilde{a}_{r-1} e^{-\tilde{b}_{r-1} \epsilon^2 n/4} \\
&\leq (\tilde{a}_r + \tilde{a}_{r-1}) e^{-(\tilde{b}_r \wedge \tilde{b}_{r-1}) \epsilon^2 n/4}.
\end{aligned}$$

This establishes Equation 3.22, with $a_r = \tilde{a}_r + \tilde{a}_{r-1}$, $b_r = (\tilde{b}_r \wedge \tilde{b}_{r-1})/4$, and $\epsilon_r = \tilde{\epsilon}_r \wedge \tilde{\epsilon}_{r-1}$. □

Proof of Theorem 3.11. By Lemma 3.12, we can now work with $M_{n,0\to r}$ rather than $M_{n,r}$. Because of the negative association of $\{Z_{n,j,r}\}_{j\in\mathbb{N}}$, it follows from Lemma 3.10 that in order to establish concentration for $M_{n,0\to r}$ it suffices to show concentration for the quantity:

$$\tilde{M}_{n,0\to r} = \sum_j p_j \tilde{Z}_{n,j,r}, \tag{3.26}$$

where $\tilde{Z}_{n,j,r}$ are independent and have marginal distributions identical to $Z_{n,j,r}$, namely Bernoulli with parameter $q_j = \sum_{k=0}^{r} \binom{n}{k} p_j^k (1-p_j)^{n-k}$. We would therefore like to use Lemma 3.6, with $W = \tilde{M}_{n,0\to r}$.

To obtain the lower exponent, it is easiest to use Lemma 3.7, with $W_j = \tilde{Z}_{n,j,r}$ and $b_j = q_j$. We have:

$$\begin{aligned}
\sum_j p_j^2 q_j &= \sum_{k=0}^{r} \sum_j p_j \binom{n}{k} p_j^{k+1}(1-p_j)^{n-k} \\
&= \sum_{k=0}^{r} \frac{\binom{n}{k}}{\binom{n+1}{k+1}} \underbrace{\sum_j p_j \binom{n+1}{k+1} p_j^{k+1}(1-p_j)^{(n+1)-(k+1)}}_{\mathbf{E}[M_{n+1,k+1}]} \\
&\leq \sum_{k=0}^{r} \frac{r+1}{n+1}\mathbf{E}[M_{n+1,k+1}] = \frac{r+1}{n+1}\mathbf{E}[M_{n+1,1\to r+1}] \\
&\leq \frac{r+1}{n+1}
\end{aligned}$$

Adapting this bound to Equation (3.21), we therefore have the lower exponent:

$$S(W, \mathbf{E}[W]-\epsilon) \geq \epsilon^2 n/[2(r+1)].$$

To obtain the upper exponent, we would like to use part *(i)* of Lemma 3.6. Thanks to independence and separation, the Gibbs measure for $W = \tilde{M}_{n,0\to r}$ remains a sum of independent Bernoulli random variables. However, rather than q_j, these are parametrized by the following:

$$q_j(\beta) := \frac{q_j e^{\beta p_j}}{q_j e^{\beta p_j} + 1 - q_j}.$$

Therefore, the Gibbs variance is given by:

$$\sigma^2(W,\beta) = \sum_j p_j^2 q_j(\beta)\left(1-q_j(\beta)\right) \leq \sum_j p_j^2 q_j(\beta).$$

For $\beta \geq 0$, we have $q_j e^{\beta p_j} + 1 - q_j \geq 1$. Using this and the fact that $e^{\beta p_j} \leq (1-p_j)^{-\beta}$, we can focus our attention to $\beta \leq n-r$, and write:

$$\begin{aligned}
\sigma^2(W,\beta) &\leq \sum_j p_j^2 q_j (1-p_j)^{-\beta} \\
&= \sum_{k=0}^{r} \sum_j p_j \binom{n}{k} p_j^{k+1}(1-p_j)^{n-\beta-k} \\
&= \sum_{k=0}^{r} \frac{\binom{n}{k}}{\binom{n-\beta+1}{k+1}} \underbrace{\sum_j p_j \binom{n-\beta+1}{k+1} p_j^{k+1}(1-p_j)^{(n-\beta+1)-(k+1)}}_{:=\, \zeta_{n,k+1}(\beta)} \\
&= \sum_{k=0}^{r} \frac{k+1}{n-\beta+1} \frac{\binom{n}{k}}{\binom{n-\beta}{k}} \zeta_{n,k+1}(\beta)
\end{aligned}$$

Here we have used the usual extension of the binomial, to arbitrary real arguments, which can be expressed in terms of the Γ function, or falling products. For every $\beta \le n-r$, the $\zeta_{n,k}(\beta)$ define a (defective) probability mass function on the non-negative integers k (just as $\mathbf{E}[M_{n+1,k+1}]$ did in the lower exponent derivation), in the sense that $0 \le \zeta_{n,k}(\beta) \le 1$ for every k, and $\sum_k \zeta_{n,k}(\beta) \le 1$. Therefore, if we bound every summand, we can use the largest bound to bound the entire sum. We have:

$$\frac{\binom{n}{k}}{\binom{n-\beta}{k}} \le \frac{\left(\frac{ne}{k}\right)^k}{\left(\frac{n-\beta}{k}\right)^k} = e^k \left(1-\beta/n\right)^{-k}.$$

Therefore the largest summand bound is that at $k=r$:

$$\sigma^2(W,\beta) \le \frac{r+1}{n-\beta+1} e^r \left(1-\beta/n\right)^{-r} = \frac{(r+1)e^r}{n} \left(1-\beta/n\right)^{-(r+1)} \tag{3.27}$$

Now select $\overline{\beta} = n/(r+2)$ and $\overline{\sigma}^2 = \frac{(r+1)e^{r+1}}{n}$. First observe that $\overline{\beta} < n-r$ since $n > 2r$. Then, using the fact that $x \le 1/(m+1)$ implies $(1-x)^{-m} \le e$, it follows from Equation (3.27) that for all $0 < \beta < \overline{\beta}$ we indeed have $\sigma^2(W,\beta) \le \overline{\sigma}^2$. Therefore, part *(i)* of Lemma 3.6 applies, and we deduce that for all $0 < \epsilon \le 1 < \frac{r+1}{r+2}e^{r+1} \equiv \overline{\beta}\overline{\sigma}^2$ we have:

$$S(W, \mathbf{E}[W]+\epsilon) \ge \frac{\epsilon^2}{2\overline{\sigma}^2}.$$

By combining the lower and upper exponents using a union bound, we get that for $\tilde{a}_r = 2$, $\tilde{b}_r = 1/[2(r+1)e^{r+1}]$, $\epsilon_r = 1$, we have that for every $n > 2r$ and for all $0 < \epsilon < 1$, the additive concentration for $\tilde{M}_{n,0\to r}$ and consequently for $M_{n,0\to r}$ as given by Equation (3.25) holds, and by Lemma 3.12, so does the additive concentration for $M_{n,r}$ as given by Equation (3.22). □

■ 3.4 Multiplicative Concentration

Our objective in this section is to establish strong laws for $M_{n,r}$ for all $r = 0, 1, \cdots$, which is an extension of the known result for the case of the missing mass ($r=0$), which was previously established (without explicit proof) by Karlin [19] (Theorem 9) and (with an explicit proof) by Gnedin et al [16] (Proposition 2.5). The results we present here differ from the latter in two important ways: they use power (Chebyshev) moments and concentration whereas we use exponential (Chernoff) moments and concentration, and they use the Poissonization method whereas we use negative association instead. The derivation of multiplicative concentration parallels that of additive concentration as in the previous section, with the use of regular variation to more tightly bound moment growth. The main theorem of this section is as follows.

Theorem 3.13. *Assume $\mathbb{P}$ is regularly varying with index $\alpha \in (0,1)$, as in Definition 3.1. Then for every $r = 0, 1, 2, \cdots$, there exists an absolute constant a_r, and distribution* *specific constants $b_r > 0$, $n_r < \infty$ and $\delta_r > 0$, such that for all $n > n_r$ and for all $0 < \delta < \delta_r$, we have:*

$$\mathbb{P}\left\{\left|\frac{M_{n,r}}{\mathbf{E}[M_{n,r}]} - 1\right| > \delta\right\} \le a_r e^{-b_r \delta^2 n^\alpha \ell(n)}. \tag{3.28}$$

We cannot deduce the multiplicative concentration of Theorem 3.13 directly from the additive concentration of Theorem 3.11, because the latter uses a worst case bound on the Gibbs variance, whereas regular variation allows us to give more information about how this variance behaves. This is what we harness in the proof.

Proof of Theorem 3.13. Throughout this proof, $\eta > 0$ is an arbitrary constant. For clarity of exposition, we repeat parts of the additive concentration proof, and use regular variation whenever it enters into play. Once more, let's first work with $M_{n,0\to r}$ rather than $M_{n,r}$. Again, because of the negative association of $\{Z_{n,j,r}\}_{j\in\mathbb{N}}$, it follows from Lemma 3.10 that in order to establish (additive) concentration for $M_{n,0\to r}$ it suffices to show concentration for the quantity:

$$\tilde{M}_{n,0\to r} = \sum_j p_j \tilde{Z}_{n,j,r},$$

where $\tilde{Z}_{n,j,r}$ are independent and have marginal distributions identical to $Z_{n,j,r}$, namely Bernoulli with parameter $q_j = \sum_{k=0}^r \binom{n}{k} p_j^k (1-p_j)^{n-k}$. We would therefore like to use Lemma 3.6, with $W = \tilde{M}_{n,0\to r}$.

To obtain the lower exponent, we use the specialized Lemma 3.7 instead, with $W_j = \tilde{Z}_{n,j,r}$ and $b_j = q_j$. We have:

$$\begin{aligned}
\sum_j p_j^2 q_j &= \sum_{k=0}^r \sum_j p_j \binom{n}{k} p_j^{k+1} (1-p_j)^{n-k} \\
&= \sum_{k=0}^r \frac{\binom{n}{k}}{\binom{n+1}{k+1}} \underbrace{\sum_j p_j \binom{n+1}{k+1} p_j^{k+1} (1-p_j)^{(n+1)-(k+1)}}_{\mathbf{E}[M_{n+1,k+1}]} \\
&= \frac{r+1}{n+1} \sum_{k=0}^r \mathbf{E}[M_{n+1,k+1}]
\end{aligned}$$

At this point, we diverge from the additive concentration derivation, to use regular variation. Let $\mathbb{P}$ be regularly varying with index α. Then there exists a sample size $n_{r,1}(\eta) > 2r$ that depends only on $\mathbb{P}$, r, and η, such that for all $n > n_{r,1}(\eta)$ we have:

$$\frac{r+1}{n+1} \sum_{k=0}^r \mathbf{E}[M_{n+1,k+1}] = \frac{r+1}{n+1} \sum_{k=0}^r \frac{k+2}{n+2} \mathbf{E}[K_{n+2,k+2}]$$

$$\begin{aligned}
&\leq (1+\eta)\frac{r+1}{n+1}\sum_{k=0}^{r}\frac{k+2}{n+2}\frac{\alpha\Gamma(k+2-\alpha)}{(k+2)!}n^{\alpha}\ell(n) \\
&= (1+\eta)(r+1)\sum_{k=0}^{r}\frac{\alpha\Gamma(k+2-\alpha)}{(k+1)!}n^{-(2-\alpha)}\ell(n). \quad (3.29)
\end{aligned}$$

Now observe that $\sum_{k=0}^{r}\frac{\alpha\Gamma(k+2-\alpha)}{(k+1)!} < (r+1)$, since $\alpha \in (0,1)$. Therefore, for $c_{r,1} := (r+1)^{-2}$, we have $\sum_j p_j^2 q_j \leq \left(c_{r,1}n^{2-\alpha}\right)^{-1}(1+\eta)\ell(n)$. Adapting this to Equation (3.21), we therefore have, for all $n > n_{r,1}(\eta)$ and all $\epsilon > 0$, the lower exponent:

$$S(W,\mathbf{E}[W]-\epsilon) \geq \frac{c_{r,1}}{2}\epsilon^2 n^{2-\alpha}\frac{1}{(1+\eta)\ell(n)}.$$

We follow a similar track for the upper exponent. Once again, we would like to use part *(i)* of Lemma 3.6. Recall that the Gibbs measure for $W = \tilde{M}_{n,0\to r}$ remains a sum of independent Bernoulli random variables, with modified parameters $q_j(\beta) := \frac{q_j e^{\beta p_j}}{q_j e^{\beta p_j}+1-q_j}$. For any $0 \leq \beta \leq n-r$, we bound the Gibbs variance as before and write:

$$\begin{aligned}
\sigma^2(W,\beta) &\leq \sum_j p_j^2 q_j (1-p_j)^{-\beta} \\
&= \sum_{k=0}^{r}\sum_j p_j \binom{n}{k} p_j^{k+1}(1-p_j)^{n-\beta-k} \\
&= \sum_{k=0}^{r}\frac{\binom{n}{k}}{\binom{n-\beta+1}{k+1}}\underbrace{\sum_j p_j\binom{n-\beta+1}{k+1}p_j^{k+1}(1-p_j)^{(n-\beta+1)-(k+1)}}_{\zeta_{n,k+1}(\beta)} \\
&= \sum_{k=0}^{r}\frac{k+1}{n-\beta+1}\frac{\binom{n}{k}}{\binom{n-\beta}{k}}\zeta_{n,k+1}(\beta) \\
&\leq \frac{(r+1)e^r}{n}(1-\beta/n)^{-(r+1)}\sum_{s=0}^{r+1}\zeta_{n,s}(\beta)
\end{aligned}$$

Recall that we chose $n_{r,1} \geq 2r$. Then, for all $n > n_{r,1}$, we have $n/(r+2) \leq n/2 < (n+1)/2 \leq n-r$. If we again select $\overline{\beta} := n/(r+2)$ then for all $0 \leq \beta \leq \overline{\beta}$ and for all $n > n_{r,1}$, we have:

$$\sigma^2(W,\beta) \leq \frac{(r+1)e^{r+1}}{n}\sum_{s=0}^{r+1}\zeta_{n,s}(\beta) \quad (3.30)$$

We have used here the same bound used in the approximations in the proof of Theorem 3.11, with the exception of preserving the sum of the $\zeta_{n,s}(\beta)$ and adding the $s=0$ term. We do this in order to exploit regular variation. With the addition of the $s=0$ term, $\sum_{s=0}^{r+1}\zeta_{n,s}(\beta)$ becomes a monotonic non-decreasing function over

$\beta \in [0, n-r]$. To see why, note that when β is an integer, this sum represents the expectation of the total rare probabilities of symbols occurring no more than $r+1$ times, out of $n-\beta$ samples. The larger the value of β, the fewer the samples, and there is in average more probability in symbols with small counts.

Assume n is even, without loss of generality, as otherwise we can use $(n+1)/2$ instead. Then, there exists a sample size $n_{r,2}(\eta) > n_{r,1}$ that depends only on $\mathbb{P}$, r, and η, such that for all $n > n_{r,2}(\eta)$ and for all $\beta \leq \overline{\beta}$, we have:

$$\begin{aligned}
\sum_{s=0}^{r+1} \zeta_{n,s}(\beta) &\leq \sum_{s=0}^{r+1} \zeta_{n,s}(n/2) \equiv \sum_{s=0}^{r+1} \mathbf{E}[M_{n/2+1,s}] \\
&= \sum_{s=0}^{r+1} \frac{s+1}{n/2+2} \mathbf{E}[K_{n/2+2,s+1}] \\
&\leq (1+\eta) \sum_{s=0}^{r+1} \frac{s+1}{n/2+2} \frac{\alpha\Gamma(s+1-\alpha)}{(s+1)!} (n/2+2)^{\alpha} \ell(n) \\
&= (1+\eta) \sum_{s=0}^{r+1} 2^{1-\alpha} \frac{\alpha\Gamma(s+1-\alpha)}{s!} n^{-(1-\alpha)} \ell(n) \qquad (3.31)
\end{aligned}$$

Now observe that $(r+1)e^{r+1} \sum_{s=0}^{r+1} 2^{1-\alpha} \frac{\alpha\Gamma(s+1-\alpha)}{s!} < 2e^{r+1}(r+1)(r+2)$, since $\alpha \in (0,1)$. Therefore, using $c_{r,2} := [2e^{r+1}(r+1)(r+2)]^{-1}$, we can combine Equations (3.30) and (3.31), and obtain that for all $\beta \leq \overline{\beta}$, we have $\sigma^2(W, \beta) \leq \overline{\sigma}^2$, where

$$\overline{\sigma}^2 := \left[c_{r,2} n^{2-\alpha}\right]^{-1} (1+\eta)\ell(n).$$

With this bound, part *(i)* of Lemma 3.6 applies, and we deduce that for every $n > n_{r,2}(\eta)$, and for all $0 < \epsilon < \frac{1}{r+2} \left[c_{r,2} n^{1-\alpha}\right]^{-1} (1+\eta)\ell(n) \equiv \overline{\beta}\overline{\sigma}^2$ we have the upper exponent:

$$S(W, \mathbf{E}[W] + \epsilon) \geq \frac{\epsilon^2}{2\overline{\sigma}^2} = \frac{c_{r,2}}{2} \epsilon^2 n^{2-\alpha} \frac{1}{(1+\eta)\ell(n)}.$$

Let $\tilde{a}_r = 2$, $\tilde{b}_r = (c_{r,1} \wedge c_{r,2})/2$, and $\tilde{\epsilon}_r = \tilde{d}_r (1+\eta) n^{\alpha-1} \ell(n)$ where $\tilde{d}_r = \frac{1}{r+2} c_{r,2}^{-1} = 2(r+1)e^{r+1}$. Then, by combining the lower and upper exponents using a union bound, we get that for every $n > n_{r,2}$ and for all $0 < \epsilon < \tilde{\epsilon}_r$, the additive concentration for $\tilde{M}_{n,0\to r}$ and therefore of $M_{n,0\to r}$ holds as follows:

$$\mathbb{P}\left\{|M_{n,0\to r} - \mathbf{E}[M_{n,0\to r}]| > \epsilon\right\} \leq \tilde{a}_r e^{-\tilde{b}_r \epsilon^2 n^{2-\alpha} \frac{1}{(1+\eta)\ell(n)}}. \qquad (3.32)$$

Observe that the range of ϵ depends on n, but that won't be a problem when we switch to multiplicative mode.

By using the same development as the proof of Lemma 3.12, we can deduce that there exist constants $a_r = \tilde{a}_r + \tilde{a}_{r-1}$, $\check{b}_r = (\tilde{b}_r \wedge \tilde{b}_{r-1})/4$, and $\epsilon_r = d_r (1+\eta) n^{\alpha-1} \ell(n)$ with $d_r = \tilde{d}_r \wedge \tilde{d}_{r-1}$, such that for every $n > n_{r,2}$ and for all $0 < \epsilon < \epsilon_r$, we have:

$$\mathbb{P}\left\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\right\} \leq a_r e^{-\check{b}_r \epsilon^2 n^{2-\alpha} \frac{1}{(1+\eta)\ell(n)}}. \qquad (3.33)$$

Now, observe that:

$$\begin{aligned}\mathbf{E}[M_{n,r}] &= \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}] \\ &\sim \frac{\alpha\Gamma(r+1-\alpha)}{r!}n^{\alpha-1}\ell(n). \end{aligned} \tag{3.34}$$

Let's define $m_r := \frac{\alpha\Gamma(r+1-\alpha)}{r!}$ for convenience. It follows from Equation (3.34) that there exists a sample size $n_r(\eta) > n_{r,2}(\eta)$ that depends only on $\mathbb{P}$, r, and η, such that for all $n > n_r(\eta)$ we have:

$$(1+\eta)^{-1}m_r n^{\alpha-1}\ell(n) \leq \mathbf{E}[M_{n,r}] \leq (1+\eta)m_r n^{\alpha-1}\ell(n).$$

Let a_r be as before, $b_r = \breve{b}_r m_r^2/(1+\eta)^3$, and $\delta_r = d_r/m_r$. Then for every $n > n_r(\eta)$ and for all $\delta < \delta_r$, we have:

$$\delta\mathbf{E}[M_{n,r}] \leq \delta_r(1+\eta)m_r n^{\alpha-1}\ell(n) \leq d_r(1+\eta)n^{\alpha-1}\ell(n) = \epsilon_r.$$

Therefore Equation (3.33) applies, and we get:

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{M_{n,r}}{\mathbf{E}[M_{n,r}]}-1\right| > \delta\right\} &= \mathbb{P}\left\{|M_{n,r}-\mathbf{E}[M_{n,r}]| > \delta\mathbf{E}[M_{n,r}]\right\} \\ &\leq a_r e^{-\breve{b}_r\delta^2\mathbf{E}[M_{n,r}]^2 n^{2-\alpha}\frac{1}{(1+\eta)\ell(n)}} \\ &\leq a_r \exp\left\{-\breve{b}_r\delta^2\left[\frac{m_r n^{\alpha-1}\ell(n)}{1+\eta}\right]^2 n^{2-\alpha}\frac{1}{(1+\eta)\ell(n)}\right\} \\ &= a_r\exp\left\{-\frac{\breve{b}_r m_r^2}{(1+\eta)^3}\delta^2 n^\alpha\ell(n)\right\} = a_r e^{-b_r\delta^2 n^\alpha \ell(n)}.\end{aligned}$$

Lastly, note that for fixed $\eta > 0$, b_r and δ_r depend on $\mathbb{P}$, but do so only through α, due to m_r. On the other hand, the sample sizes n_r depend on the particular convergence rates in the regular variation characterization, and to describe them explicitly requires more distribution specific knowledge than simply having α. □

■ 3.5 Consistent Probability Estimation

■ 3.5.1 Strong Laws and Consistency of Good-Turing Estimation

The strong laws for the rare probabilities are easily established using the multiplicative concentration of Theorem 3.13.

Theorem 3.14. *If $\mathbb{P}$ is regularly varying with index $\alpha \in (0,1)$, as in Definition 3.1, then for every $r = 0, 1, 2, \cdots$, we have*

$$\frac{M_{n,r}}{\mathbf{E}[M_{n,r}]} \to_{\text{a.s.}} 1, \tag{3.35}$$

and the asymptotic expression:

$$M_{n,r} \sim_{\text{a.s.}} \mathbf{E}[M_{n,r}] \sim \frac{\alpha\Gamma(r+1-\alpha)}{r!} n^{\alpha-1}\ell(n). \tag{3.36}$$

Proof. For any $\alpha \in (0,1)$, the integral $\int_0^\infty e^{-z^\alpha} \mathrm{d}z = \Gamma\left(1+\frac{1}{\alpha}\right)$, i.e. converges and is bounded. By a change of variable and the integral test, it follows that the right hand side of inequality (3.28) is summable. Therefore, we can apply the Borel-Cantelli lemma in the usual way, to obtain the almost sure convergence of Equation (3.35).

For (3.36), recall that $\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}]$, therefore it follows from Equation (3.15) that:

$$\mathbf{E}[M_{n,r}] \sim \frac{\alpha\Gamma(r+1-\alpha)}{r!} n^{\alpha-1}\ell(n). \qquad \square$$

As a first application of these strong laws, in conjunction with the strong laws for $K_{n,r}$, we prove the strong consistency of the Good-Turing estimator in this regime.

Theorem 3.15. *If $\mathbb{P}$ is regularly varying with index $\alpha \in (0,1)$, as in Definition 3.1, then the Good-Turing estimators are strongly consistent for all $r = 0, 1, \cdots$:*

$$\frac{G_{n,r}}{M_{n,r}} \rightarrow_{\text{a.s.}} 1. \tag{3.37}$$

Proof. Recall that $G_{n,r} = \frac{r+1}{n}K_{n,r+1}$. Therefore by the strong law of the rare counts, i.e. Equations (3.15) and (3.16), we have that

$$\frac{G_{n,r}}{\mathbf{E}[G_{n,r}]} \rightarrow_{\text{a.s.}} 1. \tag{3.38}$$

On the other hand, note that by equation (3.15), we have $\mathbf{E}[K_{n,r}]/\mathbf{E}[K_{n+1,r}] \to 1$ for any r. Since $\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}]$, it follows that

$$\frac{\mathbf{E}[G_{n,r}]}{\mathbf{E}[M_{n,r}]} \to 1. \tag{3.39}$$

Combining the convergences (3.35), (3.38), and (3.39), we obtain (3.37). □

■ 3.5.2 New Consistent Estimators

Since we are now considering a model where regular variation plays a critical role, we can take inspiration from extreme value theory, where regular variation is also pivotal. In particular, we suggest dividing the estimation task into two: estimating the regular variation index, then using it in asymptotic expressions, in order to estimate the quantities of interest. In particular, we shall show that the Good-Turing estimator for the missing mass itself has such a two-stage characterization, but that the concept can be used to develop a richer class of estimators for the missing mass and other rare probabilities.

Estimating the Regular Variation Index

Using Equations (3.12) and (3.13), we have that the ratio of the number of symbols appearing exactly once to the total number of distinct symbols defines a consistent estimator of the regular variation index:

$$\hat{\alpha} := \frac{K_{n,1}}{K_n} \rightarrow_{\text{a.s.}} \alpha. \tag{3.40}$$

Note that this is by no means the only approach to estimating the index. For example, other asymptotic expressions that appear in Theorem 3.3 may be harnessed. Moreover, one may devise methods that are inspired from techniques in extreme value theory [1], such as performing a Gumbelian splitting of the data into M blocks of size N, i.e. $n = M \cdot N$. Then one can perform the naive index estimation of Equation (3.40) in each block, call it $\hat{\alpha}_m$, $m = 1, \cdots, M$, then average out:

$$\hat{\alpha} = \frac{1}{M} \sum_m \hat{\alpha}_m. \tag{3.41}$$

With proper choices of M and N, this empirically shows much less volatility than a straight application of (3.40) (i.e. $M = 1$, $N = n$), as illustrated in Figure 3.5.2. The analytic insight behind this fact is the variance reduction of the averaging process. The bias of the estimator will depend on how fast the regular variation comes into effect, i.e. for any given $\eta > 0$ for what n do we get within a $1 + \eta$ factor of the limiting expression? However this process helps reduce variance within this mode, and if N grows fast enough, then the compromise results in palpable improvement. Ideally, one would want to perform this estimation without explicit splitting, and rather use a principled semi-parametric approach, such as the Hill estimator and its more sophisticated variants [1]. However, we do not elaborate on that in this thesis.

Two Probability Estimator Constructions

We have shown that the Good-Turing estimator is consistent in the regularly varying heavy-tailed setting. But the questions remains whether one could do better by working within this framework explicitly. In this section, we provide two new rare probability estimators, and show that they are consistent. Furthermore, these address some of the shortcomings of the Good-Turing estimator. For example, they incorporate smoothing implicitly.

In particular, our constructions are in two stages. We first assume that we have chosen a consistent estimator $\hat{\alpha} \rightarrow_{\text{a.s.}} \alpha$. This can be given by (3.40), or potentially more powerful estimators. Then, we focus on how to use the estimated index to construct consistent estimators for rare probabilities.

Theorem 3.16. *Consider the following family of estimators, for $r = 1, 2, \cdots$:*

$$\hat{M}^{(1)}_{n,r} := (r - \hat{\alpha}) \frac{K_{n,r}}{n} \tag{3.42a}$$

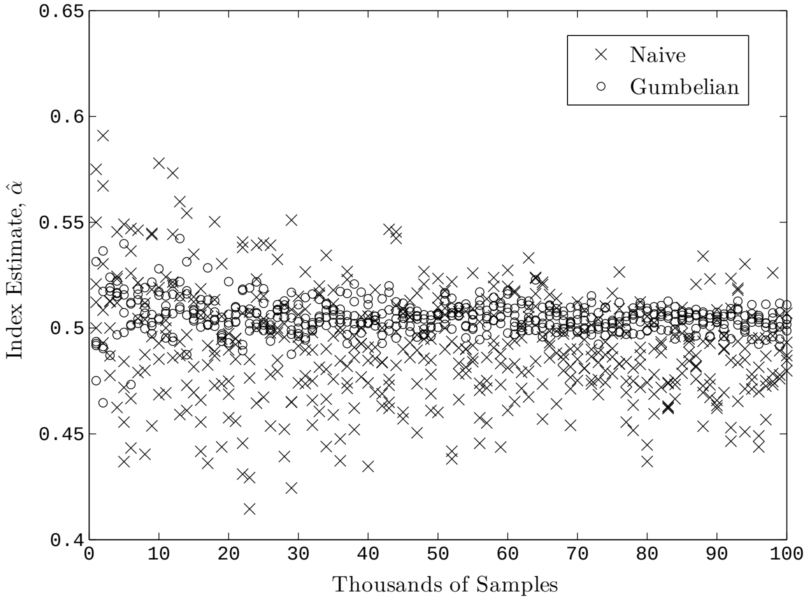

Figure 3.1. Comparison of naive and Gumbelian index estimators. The underlying distribution is regularly varying with index 0.5. For a sample size of n, the Gumbelian estimator is performing averaging of $M = \lfloor\sqrt{n}\rfloor$ blocks of size $N = \lfloor\sqrt{n}\rfloor$, according to Equation (3.41). Observe the considerably improved volatility of the resulting estimator.

and for $r = 0$:

$$\hat{M}^{(1)}_{n,0} := 1 - \sum_{r\geq 1} \hat{M}^{(1)}_{n,r} = \hat{\alpha}\frac{K_n}{n}. \tag{3.42b}$$

If $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3.1. Then $\hat{M}^{(1)}_{n,r}$ are strongly consistent, for all $r = 0, 1, \cdots$:

$$\frac{\hat{M}^{(1)}_{n,r}}{M_{n,r}} \to_{\text{a.s.}} 1. \tag{3.43}$$

Proof. Since $\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}]$, it follows from Equation (3.15) and the strong law for $M_{n,r}$ given by (3.35) that:

$$M_{n,r} \sim_{\text{a.s.}} \frac{\alpha\Gamma(r+1-\alpha)}{r!} n^{\alpha-1}\ell(n).$$

First consider the case $r = 1, 2, \cdots$. Since $\hat{\alpha} \to_{\text{a.s.}} \alpha$, it follows that $(r - \hat{\alpha}) \sim_{\text{a.s.}} (r - \alpha)$. Observing that $\Gamma(r + 1 - \alpha) = (r - \alpha)\Gamma(r - \alpha)$, we can use Equation (3.16) to obtain:

$$\hat{M}^{(1)}_{n,r} = (r - \hat{\alpha})\frac{K_{n,r}}{n} \sim_{\text{a.s.}} \frac{\alpha\Gamma(r+1-\alpha)}{r!} n^{\alpha-1}\ell(n) \sim_{\text{a.s.}} M_{n,r}.$$

For the case $r = 0$, we use Equation (3.12) and $\hat{\alpha} \rightarrow_{\text{a.s.}} \alpha$ to obtain:

$$\hat{M}^{(1)}_{n,0} = \hat{\alpha}\frac{K_n}{n} \sim_{\text{a.s.}} \frac{\alpha\Gamma(1-\alpha)}{r!}n^{\alpha-1}\ell(n) \sim_{\text{a.s.}} M_{n,0}. \qquad \square$$

One motivation for introducing the $\hat{M}^{(1)}_{n,r}$ family is because they have the 'absolute discount' form that is a component of many language learning heuristics, and especially the Kneser-Ney line of state-of-the-art algorithms [7]. Here we have systematically shown the nature of the discount as being the regular variation index. The structure of $\hat{M}^{(1)}_{n,r}$ addresses a peculiarity of the Good-Turing estimator. In particular $G_{n,r}$ will assign a probability of zero to a group of symbols, simply because there are no symbols appearing in the one higher occupancy level $r + 1$, regardless to how many symbols there are in the occupancy level r. Note that here, on the contrary, the estimator will evaluate to 0 if and only if there are no symbols in the occupancy level itself. As such, $M^{(1)}_{n,r}$ has smoothing built-in.

We can push smoothing further by using K_n, the number of distinct observed symbols, rather than relying on the individual occupancy numbers $K_{n,r}$. Since K_n experiences less variability, e.g. it is monotonically increasing with sample size, the resulting estimator inherits some of that robustness.

Theorem 3.17. *Consider the following family of estimators, for* $r = 0, 1, \cdots$*:*

$$\hat{M}^{(2)}_{n,r} := \frac{\hat{\alpha}\Gamma(r+1-\hat{\alpha})}{r!\Gamma(1-\hat{\alpha})}\frac{K_n}{n} \equiv \binom{r-\hat{\alpha}}{r}\hat{\alpha}\frac{K_n}{n}. \tag{3.44}$$

If $\mathbb{P}$ *is regularly varying with index* $\alpha \in (0,1)$*, as in Definition 3.1. Then* $\hat{M}^{(2)}_{n,r}$ *are strongly consistent, for all* $r = 0, 1, \cdots$*:*

$$\frac{\hat{M}^{(2)}_{n,r}}{M_{n,r}} \rightarrow_{\text{a.s.}} 1. \tag{3.45}$$

Proof. For convenience, define:

$$g(\alpha) := \frac{\alpha\Gamma(r+1-\alpha)}{r!\Gamma(1-\alpha)} \equiv \binom{r-\alpha}{r}\alpha.$$

We can thus write, as in the proof of Theorem 3.16:

$$M_{n,r} \sim_{\text{a.s.}} \frac{\alpha\Gamma(r+1-\alpha)}{r!}n^{\alpha-1}\ell(n) = g(\alpha)\Gamma(1-\alpha)n^{\alpha-1}\ell(n).$$

By the continuity of $g(\alpha)$, since $\hat{\alpha} \rightarrow_{\text{a.s.}} \alpha$, we also have that $g(\hat{\alpha}) \rightarrow_{\text{a.s.}} g(\alpha)$. Therefore, using Equation (3.12) we obtain:

$$\hat{M}^{(2)}_{n,r} = g(\hat{\alpha})\frac{K_n}{n} \sim_{\text{a.s.}} g(\alpha)\Gamma(1-\alpha)n^{\alpha-1}\ell(n) \sim_{\text{a.s.}} M_{n,r}. \qquad \square$$

It is worth noting that we always have $\hat{M}^{(1)}_{n,0} = \hat{M}^{(2)}_{n,0}$ by construction, and that if $\hat{\alpha}$ is the naive index estimator as in (3.40), we also have:

$$\hat{M}^{(1)}_{n,1} = (1-\hat{\alpha})\frac{K_{n,1}}{n} = (1-\hat{\alpha})\hat{\alpha}\frac{K_n}{n} = \hat{\alpha}\frac{\Gamma(2-\hat{\alpha})}{\Gamma(1-\hat{\alpha})} = \hat{M}^{(2)}_{n,1}. \tag{3.46}$$

Good-Turing as a Two-Stage Estimator

If we use the naive index estimator $\hat{\alpha}$ as in (3.40), then we have:

$$G_{n,0} = \frac{K_{n,1}}{n} = \frac{K_{n,1}}{K_n}\frac{K_n}{n} = \hat{\alpha}\frac{K_n}{n} = \hat{M}^{(1)}_{n,0} = \hat{M}^{(2)}_{n,0}. \tag{3.47}$$

Therefore, we can interpret the Good-Turing estimator of the missing mass as a two-stage estimator: estimate the regular variation index α, as in (3.40), and then use it to obtain a probability estimator, as in (3.42) or (3.44). However, the advantage of the estimators that we propose is that we can use any alternative index estimator α, for example as suggested by Equation (3.41), in order to benefit from less volatile convergence.

Example

As an illustration of the various convergence results and estimators, we use a simple case where $p_j \propto j^{-1/\alpha}$ is a pure power law. This defines a distribution $\mathbb{P}$ which is regularly varying with index α. In the numerical examples below we use $\alpha = \frac{3}{4}$, motivated by the very heavy tails that appear in natural language word frequencies.

In Figure 3.5.2, we show the decay behavior over up to 100,000 samples, of a sample path of the rare probabilities $M_{n,r}$ and their expectations $\mathbf{E}[M_{n,r}]$, for $r = 0, 1, 2$, and 3. We can qualitatively observe the close correspondence to the theoretical $n^{\alpha-1}$ rates.

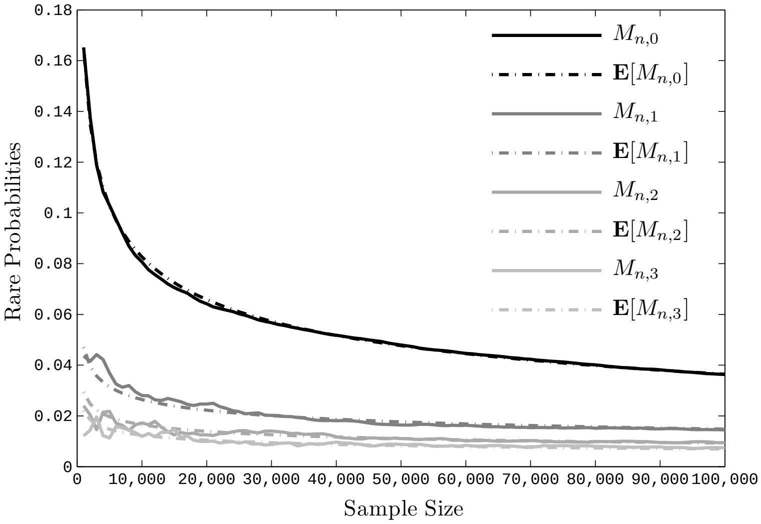

Figure 3.2. Decay behavior of $M_{n,r}$ and $\mathbf{E}[M_{n,r}]$ as n ranges from 0 to 100,000 samples, for $r = 0, 1, 2$, and 3. The underlying distribution is regularly varying, with index $\alpha = \frac{3}{4}$.

In Figure 3.5.2 we illustrate the strong law by plotting the ratio $M_{n,r}/\mathbf{E}[M_{n,r}]$. Though the convergence is far from smooth, nor does it occur at a uniform rate over

r, we can qualitatively see that the sample paths narrow down toward 1 as the sample size increases.

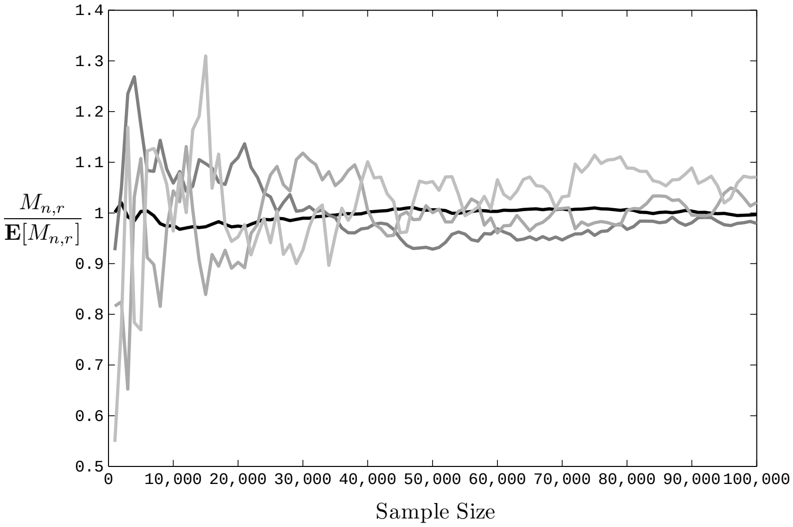

Figure 3.3. Multiplicative concentration of $M_{n,r}$ around $\mathbf{E}[M_{n,r}]$ results in a strong law where $M_{n,r}/\mathbf{E}[M_{n,r}] \to_{\text{a.s.}} 1$. We illustrate this for the same example as Figure 3.5.2.

To showcase the performance of the new estimators and to compare them to the Good-Turing estimator, we plot the general behavior of $G_{n,r}$, $\hat{M}^{(1)}_{n,r}$, and $\hat{M}^{(2)}_{n,r}$ alongside $M_{n,r}$, in the same example. We make two deliberate simulation choices:

- We use the naive estimator for α, as given by Equation (3.40), in order to show that the benefit of the new estimators comes from their structure too, and not only because of better index estimation.
- We purposefully use fewer samples than in Figures 3.5.2 and 3.5.2, in order to emphasize that the improvements appear even at moderate sample sizes. We let n range from 0 to $10,000$.

Since we know that all estimators coincide for the case $r = 0$, and the new estimators coincide for $r = 1$, we only look at the first two distinctive cases: $r = 2$ and 3. First, we show the raw behavior of the estimators in Figure 3.5.2 (for $r = 2$) and Figure 3.5.2 (for $r = 3$). Then, to make the comparison crisper, we also show the behavior of the ratios of each estimator to $M_{n,r}$ in Figure 3.5.2 (for $r = 2$) and Figure 3.5.2 (for $r = 3$). For reference, these figures also show the 1-line and the ratio of the mean itself, i.e. $\mathbf{E}[M_{n,r}]/M_{n,r}$.

Lastly, we make a few qualitative comments. First, it is apparent that $\hat{M}^{(1)}_{n,r}$ outperforms all estimators when it comes to tracking $M_{n,r}$ closely. It is followed by $\hat{M}^{(2)}_{n,r}$ in

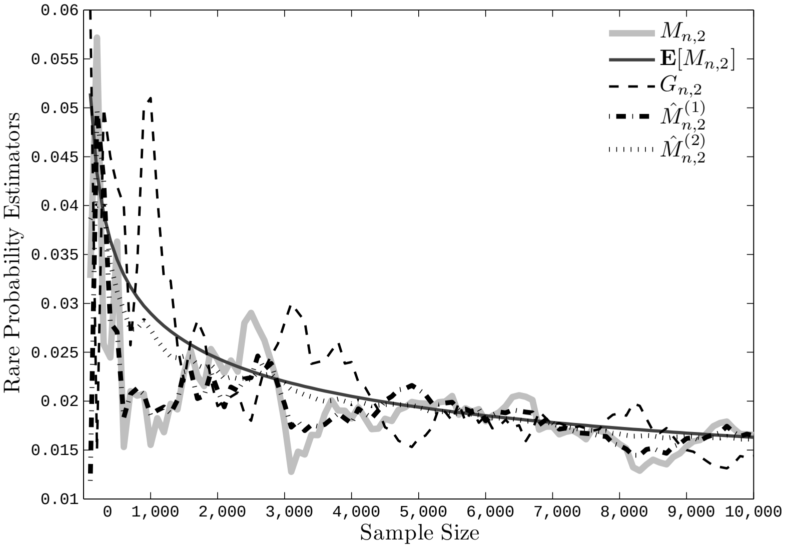

Figure 3.4. Behavior of $G_{n,2}$, $\hat{M}^{(1)}_{n,2}$, and $\hat{M}^{(2)}_{n,2}$, alongside $M_{n,2}$ and $\mathbf{E}[M_{n,2}]$, in the same example as Figure 3.5.2, as n ranges from 0 to 10,000 samples.

performance, while $\hat{G}^{(1)}_{n,r}$ is consistently more volatile than both of the new estimators. Also note that $\hat{M}^{(2)}_{n,r}$ is the smoothest estimator. However, it tracks $\mathbf{E}[M_{n,r}]$ much better than $M_{n,r}$ itself. Asymptotically, this does not matter, however for small samples this might be a feature or a shortcoming depending on whether the focus is on $M_{n,r}$ or $\mathbf{E}[M_{n,r}]$.

■ 3.6 Summary

In this chapter, we studied the problem of rare probability estimation, from the perspective of strong consistency in the sense of ratios converging to one. We first showed that consistency is not to be taken for granted. In particular, even in well behaved distributions such as the geometric, the Good-Turing estimator may not be consistent.

We then focused our attention to heavy tailed distributions, which Chapter 2 suggested as the regime of distinctly non-trivial performance of the Good-Turing estimator. We refined the notion of accrual rate, using Karamata's theory of regular variation [18], following closely the early development of Karlin [19] in the context of infinite urn schemes. We then used the McAllester-Ortiz method [21] to extend their additive concentration results to all rare probabilities. Moreover, in the setting of regularly varying heavy tailed distributions, we showed that one has multiplicative concentration.

We then used the multiplicative concentration to establish strong laws. These allowed us to show that regularly varying heavy tails are sufficient for the consistency of the Good-Turing estimator. We then used the newly established strong laws, in

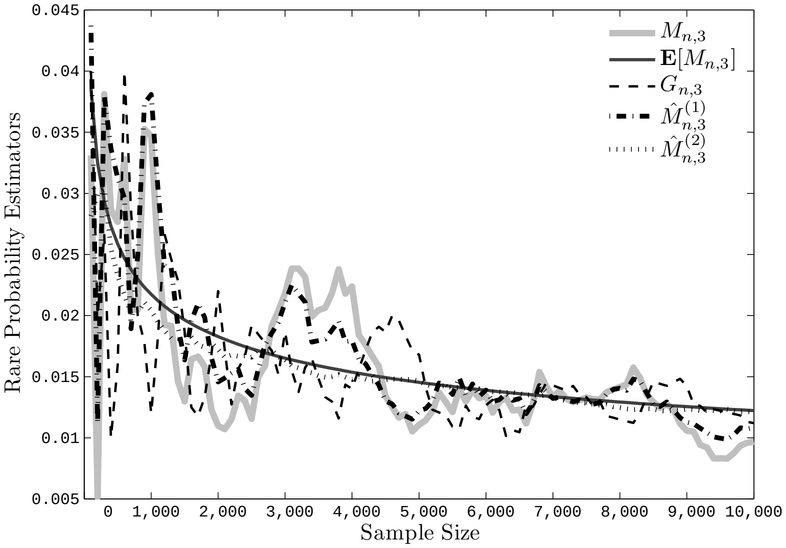

Figure 3.5. Behavior of $G_{n,3}$, $\hat{M}^{(1)}_{n,3}$, and $\hat{M}^{(2)}_{n,3}$, alongside $M_{n,3}$ and $\mathbf{E}[M_{n,3}]$, in the same example as Figure 3.5.2, as n ranges from 0 to $10,000$ samples.

addition those established for the occupancy numbers by Karlin, to construct two new families of consistent rare probability estimators. These new estimators address some of the shortcomings of the Good-Turing estimator. In particular, they have built-in smoothing, and their structure follows closely the 'absolute discounting' form used extensively in computational language modeling heuristics [7]. As such, in addition to a systematic and principled estimation method, our results provide a justification to these algorithms and an interpretation of the discount as the regular variation index. Since our estimators can be split into two parts, first index estimation and then probability estimation, they are closely related to tail estimation techniques in extreme value theory [1]. This correspondence opens the door for modern semiparametric methods to be applied in the present framework. We leave this elaboration to the conclusion, in Chapter 5.

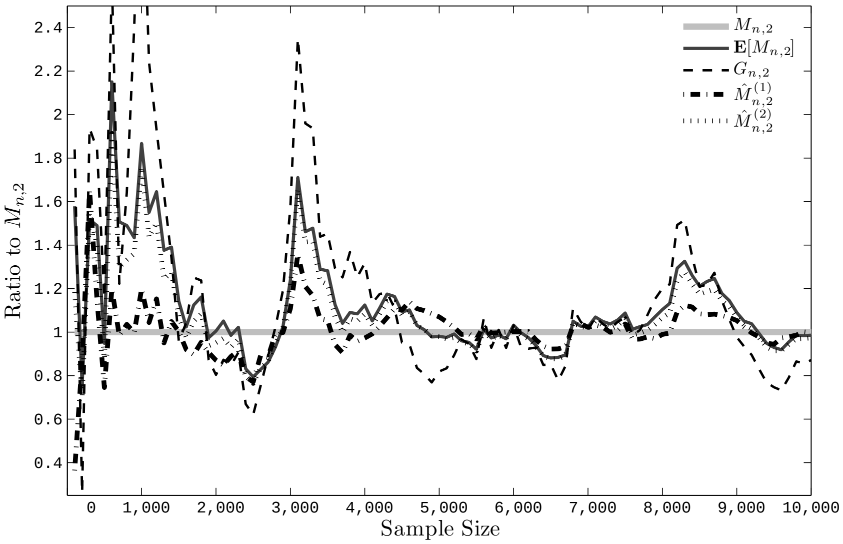

Figure 3.6. Ratios $G_{n,2}/M_{n,2}$, $\hat{M}^{(1)}_{n,2}/M_{n,2}$, and $\hat{M}^{(2)}_{n,2}/M_{n,2}$ alongside $\mathbf{E}[M_{n,2}]/M_{n,2}$, in the same example as Figure 3.5.2, as n ranges from 0 to $10,000$ samples.

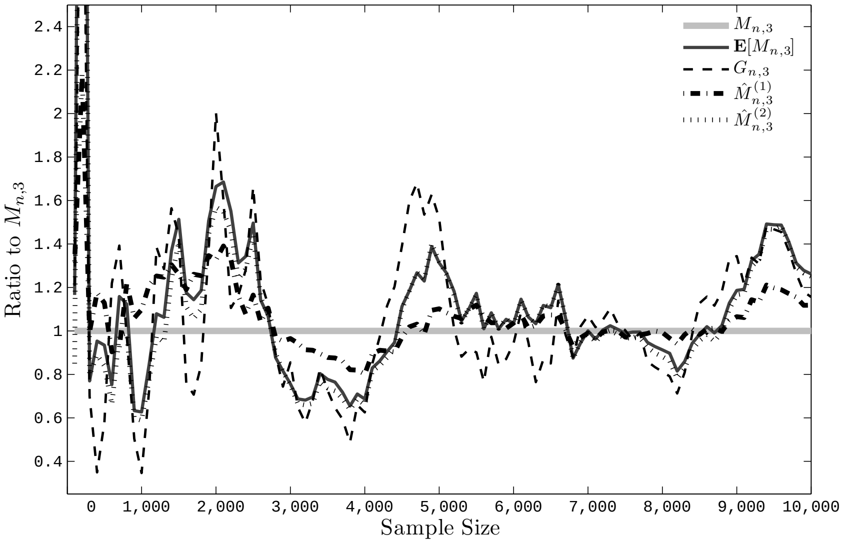

Figure 3.7. Ratios $G_{n,3}/M_{n,3}$, $\hat{M}^{(1)}_{n,3}/M_{n,3}$, and $\hat{M}^{(2)}_{n,3}/M_{n,3}$ alongside $\mathbf{E}[M_{n,3}]/M_{n,3}$, in the same example as Figure 3.5.2, as n ranges from 0 to $10,000$ samples.

Chapter 4

Canonical Estimation in a Scaling Regime

WE now propose a general methodology for performing statistical inference within a 'rare-events regime' that was recently suggested by Wagner, Viswanath and Kulkarni. Our approach allows one to easily establish consistent estimators for a very large class of canonical estimation problems, in a large alphabet setting. These include the problems studied in the original chapter, such as entropy and probability estimation, in addition to many other interesting ones. We particularly illustrate this approach by consistently estimating the size of the alphabet and the range of the probabilities. We start by proposing an abstract methodology based on constructing a probability measure with the desired asymptotic properties. We then demonstrate two concrete constructions by casting the Good-Turing estimator as a pseudo-empirical measure, and by using the theory of mixture model estimation.

■ 4.1 Introduction

We propose a general methodology for performing statistical inference within the 'rare-events regime' suggested by Wagner, Viswanath and Kulkarni in [31], referred to as WVK hereafter. This regime is a scaling statistical model that strives to capture large alphabet settings, and is characterized by the following notion of a *rare-events source.*

Definition 4.1. *Let $\{(A_n, p_n)\}_{n\in\mathbb{N}}$ be a sequence of pairs where each A_n is an alphabet of finite symbols, and p_n is a probability mass function over A_n. Let X_n be a single sample from p_n, and use it to define a 'shadow' sequence $Z_n = np_n(X_n)$. Let P_n denote the distribution of Z_n. We call $\{(A_n, p_n)\}_{n\in\mathbb{N}}$ a* rare-events source, *if the following conditions hold.*

(i) There exists an interval $C = [\check{c}, \hat{c}]$, $0 < \check{c} \leq \hat{c} < \infty$, such that for all $n \in \mathbb{N}$ we have $\frac{\check{c}}{n} \leq p_n(a) \leq \frac{\hat{c}}{n}$ for all $a \in A_n$, or equivalently, P_n is supported on C.

(ii) There exists a random variable Z, such that $Z_n \to Z$ in distribution. Equivalently, there exists a distribution P, such that $P_n \Rightarrow P$ weakly.

To complete the model, we adopt the following sampling scheme. For each n, we draw n independent samples from p_n, and we denote them by $X_{n,1}, \cdots, X_{n,n}$. Using these samples, we are interested in estimating various quantities. WVK consider, among a few others, the following:

- The total (Good-Turing) probabilities of all symbols appearing exactly k times, for each $k \in \mathbb{N}_0$.

- The normalized log-probability of the observed sequence.

- The normalized entropy of the source.

- The relative entropy between the true and empirical distributions.

They also consider two-sequence problems and hypothesis testing, but we focus here on single sequence estimation.

It is striking that many of these quantities can be estimated in such a harsh scaling model, where one cannot hope for the empirical distribution to converge in any traditional sense. However, WVK's estimators have some drawbacks. For example, since they are based on series expansions of the quantities to be estimated, one has to carefully choose the growth rate of partial sums, in order to control convergence properties. More importantly, they are specifically tailored to each individual task. Their consistency is established on a case-by-case basis. What is desirable, and what this chapter contributes to, is a methodology for performing more general statistical inference within this regime. Ideally such a framework would allow one to tackle a very large class of canonical estimation problems, and establish consistency more easily.

We may summarize the fundamental ideas behind our approach and the organization of this chapter as follows. First, in Section 4.2, we isolate the class of estimation problems that we are interested in as those that asymptotically converge to an integral against P. The quantities studied by WVK fall in this category, and so do other interesting problems such as estimating the size of the alphabet. Other problems, such as estimating the range of the probabilities given by the support interval C, can also be studied in this framework.

Next, in Section 4.3, we propose an abstract solution methodology. At its core, we construct a (random) distribution $\tilde{P}_n$ that converges weakly to P for almost every observation sample. This construction immediately establishes the consistency of natural estimators for the abovementioned quantities, if bounds on C are known. If in addition the rate of the convergence of $\tilde{P}_n$ is established, the framework gives consistent estimators even without bounds on C.

To make this methodology concrete, we build on a core result of WVK that establishes the strong consistency of the Good-Turing estimator. In particular, since the role of the empirical measure is lost, we show in Section 4.4 that we can treat the Good-Turing estimator as a pseudo-empirical measure. Once this is established, we can borrow heavily from the theory of mixture models, where inference is done using i.i.d.

samples, and adapt it to our framework. In Section 4.5, we suggest two approaches for constructing $\tilde{P}_n$: one that is based on maximum likelihood, and another that is based on minimum distance. Both constructions guarantee the almost sure weak convergence of $\tilde{P}_n$ to P, but the latter, under some conditions, also provides the desirable convergence rates.

In Section 4.6 we illustrate the methodology with some examples. In particular, we show how one can consistently estimate the entropy of the source and the probability of the sequence as studied by WVK, but we also propose consistent estimators for the size of the alphabet and for the support interval C.

Notation

Throughout, we use $F(.;.)$ to denote the cumulative distribution of the second argument (which is a probability measure on the real line or on the integers) evaluated at the first argument (which is a point on the real line or an integer).

■ 4.2 A General Class of Estimation Problems

■ 4.2.1 Definitions

Given i.i.d. samples $X_{n,1}, \cdots, X_{n,n}$ from the rare-events source (A_n, p_n), we can pose a host of different estimation problems. Since the alphabet is changing, quantities that depend on explicit symbol labels are not meaningful. Therefore, one ought to only consider estimands that are symmetric, in the sense of invariant under re-labeling of the symbols in A_n. In particular, we consider the following class of general estimation problems.

Definition 4.2. *Consider the problem of estimating a sequence $\{Y_n\}_{n\in\mathbb{N}}$ of real-valued random variables using, for every n, the samples $X_{n,1}, \cdots, X_{n,n}$. We call this a* canonical estimation problem *if, for every rare-events source, we have:*

$$\mathbf{E}\left[Y_n\right] = \int_C f_n(x) \mathrm{d}P_n(x). \tag{4.1}$$

for some sequence of $\{f_n\}$ of continuous real-valued functions on $\mathbb{R}^+$ that are uniformly bounded on C and that converge pointwise to a continuous function f.

Observe that this definition corresponds indeed to estimands that are invariant under re-labeling, in expectation. The following lemma characterizes the limit.

Lemma 4.1. *For any canonical estimation problem,*

$$\mathbf{E}[Y_n] \to \int_C f(x) \mathrm{d}P(x). \tag{4.2}$$

Proof. Since $P_n \Rightarrow P$, we can apply Skorokhod's theorem ([3], p. 333), to construct a convergent sequence of random variables $\xi_n \to_{\text{a.s.}} \xi$, where $\xi_n \sim P_n$ and $\xi \sim P$. By continuity, it follows that $f_n(\xi_n) \to_{\text{a.s.}} f(\xi)$. By the bounded convergence theorem, we then have $\mathbf{E}[f_n(\xi_n)] \to \mathbf{E}[f(\xi)]$. Since $\mathbf{E}[Y_n] = \mathbf{E}[f_n(\xi_n)]$, and $\int_C f(x)\mathrm{d}P(x) = \mathbf{E}[f(\xi)]$, the lemma follows. □

It is often more interesting to consider the subclass of canonical problems where there is strong concentration around the mean, and where the Borel-Cantelli lemma applies to give almost sure convergence to the mean.

Definition 4.3. *If a canonical estimation problem further satisfies $|Y_n - \mathbf{E}[Y_n]| \to_{\text{a.s.}} 0$, then call it a* strong canonical problem*. It follows that for strong canonical problems,*

$$Y_n \to_{\text{a.s.}} \int_C f(x)\mathrm{d}P(x). \tag{4.3}$$

Using these definitions, a reasonable estimator will at least agree with the limit set forth in Lemma 4.1. Other modes of convergence may be reasonable, but we would like to exhibit a statistic that almost surely converges to that limit. We make this precise in the following definition.

Definition 4.4. *Given a canonical problem as in Definition 4.2, a corresponding* estimator *is a sequence $\{\hat{Y}_n\}_{n\in\mathbb{N}}$ such that, for each n, $\hat{Y}_n(a_1, \cdots, a_n)$ is a real-valued function on $(A_n)^n$, to be evaluated on the sample sequence $X_{n,1}, \cdots, X_{n,n}$. A* consistent estimator *is one that obeys*

$$\hat{Y}_n(X_{n,1}, \cdots, X_{n,n}) \to_{\text{a.s.}} \int_C f(x)\mathrm{d}P(x). \tag{4.4}$$

For canonical estimation problems that are not necessarily strong, this approach produces an asymptotically unbiased estimator, with asymptotic mean squared error that is no more than the asymptotic variance of the estimand itself. For strong canonical estimation problems, this approach establishes strong consistency, in the sense that the estimator converges to the estimand, almost surely.

■ 4.2.2 Examples

To motivate the setting we have just described, we first note that all of the quantities studied by WVK are strong canonical estimation problems. For each quantity, WVK propose an estimator, and individually establish its consistency by showing almost sure convergence to the limit in Lemma 4.1. In contrast, what we emphasize here is that this can potentially be done *universally* over all strong canonical problems.

To highlight the usefulness of this generalization, we illustrate two important quantities that fall within this framework. We will revisit these in more detail in Section 4.6. The first quantity is the normalized size of the alphabet: $|A_n|/n$. For this, one can show (see, for example, [2]), that $|A_n|/n = \int_C \frac{1}{x}\mathrm{d}P_n(x)$. Therefore we can take $f_n(x) = f(x) = \frac{1}{x}$, and since the estimand is deterministic, we have a strong canonical estimation problem.

The second quantity of interest is the interval C, or equivalently its endpoints $\check{c}$ and $\hat{c}$. Note that, by construction, P is supported on C. Without loss of generality, we may assume that $\check{c}$ and $\hat{c}$ are respectively the essential infimum and essential supremum of $Z \sim P$. Therefore, note that $\left(\int x^{\pm q} \mathrm{d}P(x)\right)^{1/q}$ converges to the essential infimum $(-)$ or supremum $(+)$ as $q \to \infty$. We can therefore consider, for fixed $q \geq 1$, the strong canonical problems that ensue from the choices $f_n(x) = f(x) = x^{-q}$ and $f_n(x) = f(x) = x^q$. These, by themselves, are not sufficient to provide estimates for $\check{c}$ and $\hat{c}$. However if, in addition to consistency, we establish the convergence rates of their estimators, then we can apply our framework to estimate C, as we show in Section 4.6.

■ 4.3 Solution Methodology

Our task now is to exhibit consistent estimators to canonical problems. We present here our abstract methodology, which we demonstrate concretely in Section 4.5. The core of our approach consists of using the samples $X_{n,1}, \cdots, X_{n,n}$ to construct a random measure $\tilde{P}_n$ over $\mathbb{R}^+$, such that for almost every sample sequence, the sequence of measures $\{\tilde{P}_n\}$ converges weakly to P. We write: as $n \to \infty$

$$\tilde{P}_n \Rightarrow_{\text{a.s.}} P. \tag{4.5}$$

If we accomplish this, we can immediately suggest a consistent estimator under certain conditions, as expressed by Lemma 4.2. We will be interested in integrating functions against the measure $\tilde{P}_n$. However, since the support C of P is unknown, we first introduce the notion of a *tapered function* as a convenient way to control the region of integration. Given a real-valued function $g(x)$ on $\mathbb{R}^+$, for every $D \geq 1$ define its D-tapered version as:

$$g_D(x) \equiv \begin{cases} g(D^{-1}) & x < D^{-1} \\ g(x) & x \in [D^{-1}, D] \\ g(D) & x > D \end{cases}$$

If g is continuous on $(0, +\infty)$, then we can think of $g_D(x)$ as a bounded continuous extension of the restriction of g on $[D^{-1}, D]$ to all of $\mathbb{R}^+$.

Lemma 4.2. *Consider a canonical problem characterized by some f. Let the support C of a rare-events source be known up to an interval $[D^{-1}, D] \supseteq C$ for some $D > 1$. Then, if $\tilde{P}_n \Rightarrow_{\text{a.s.}} P$ as $n \to \infty$, we have that*

$$\hat{Y}_n = \int_{\mathbb{R}^+} f_D(x) \mathrm{d}\tilde{P}_n(x) \tag{4.6}$$

is a consistent estimator.

Furthermore, if f is bounded everywhere, we can make the uninformative choice $D = \infty$.

Proof. Since the tapered function f_D is continuous and bounded on $\mathbb{R}^+$, the almost sure weak convergence of $\tilde{P}_n$ to P implies that $\int_{\mathbb{R}^+} f_D \mathrm{d}\tilde{P}_n \to_{\text{a.s.}} \int_{\mathbb{R}^+} f_D \mathrm{d}P$. But since P is supported on C and f_D agrees with f on C, we have $\int_{\mathbb{R}^+} f_D \mathrm{d}P = \int_C f_D \mathrm{d}P = \int_C f \mathrm{d}P$. □

In general, however, we will be interested in problems where we do not have an *a priori* knowledge about the endpoints of C, and where an uninformative choice cannot be made because f is not bounded on $\mathbb{R}^+$, such as $f(x) = \log x$, $1/x$, or x^q. For these problems, we can apply our methodology of integrating against $\tilde{P}_n$ by first establishing a rate for the convergence of equation (4.5). We characterize such a rate using a sequence $K_n \to \infty$, such that:

$$K_n d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0, \tag{4.7}$$

where d_{W} denotes the Wasserstein distance, which can be expressed in its dual forms:

$$\begin{aligned} d_{\mathrm{W}}(\tilde{P}_n, P) &\equiv \int_{\mathbb{R}^+} |F(x; P_n) - F(x; P)| dx \\ &= \sup_{h \in \text{Lipschitz}(1)} \left| \int_{\mathbb{R}^+} h \mathrm{d}P_n - \int_{\mathbb{R}^+} h \mathrm{d}P \right|. \end{aligned} \tag{4.8}$$

In the remainder of the chapter we will particularly focus on K_n of the form n^s for some $s > 0$.

In the following lemma, we describe how we can use convergence rates such as (4.7) to construct consistent estimators that work with no prior knowledge on C, for a large subclass of canonical problems.

Lemma 4.3. *Consider a canonical problem characterized by some f, which is Lipschitz on every closed interval $[a, b]$, $0 < a \leq b < \infty$. If $K_n d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$ as $n \to \infty$, for some $K_n \to \infty$, then we can choose $D_n \to \infty$ such that*

$$\hat{Y}_n = \int_{\mathbb{R}^+} f_{D_n}(x) \mathrm{d}\tilde{P}_n(x) \tag{4.9}$$

is a consistent estimator. The growth of D_n controls the growth of the Lipschitz constant of f_{D_n}, which should be balanced with the convergence rate K_n. More precisely, $\hat{Y}_n$ in (4.9) is consitent for any $D_n \to \infty$ that additionally satisfies

$$\liminf_{n \to \infty} \frac{K_n}{\mathrm{Lip}(f_{D_n})} > 0, \tag{4.10}$$

where $\mathrm{Lip}(g)$ *indicates the Lipschitz constant of g.*

Proof. First note that for any $D \geq (\check{c}^{-1} \vee \hat{c})$, since P is supported on C and f_D agrees with f on C, we have:

$$\int_{\mathbb{R}^+} f_D \mathrm{d}P = \int_C f_D \mathrm{d}P = \int_C f \mathrm{d}P. \tag{4.11}$$

Then, using the fact that for every D, $f_D/\text{Lip}(f_D)$ is Lipschitz(1), we can invoke the dual representation (4.8) of the Wasserstein distance to write:

$$K_n \sup_D \frac{1}{\text{Lip}(f_D)} \left| \int_{\mathbb{R}^+} f_D \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} f_D \mathrm{d}P \right| \to_{\text{a.s.}} 0. \tag{4.12}$$

By combining equations (4.11) and (4.12), it follows that for any sequence $D_n \to \infty$, we have:

$$\frac{K_n}{\text{Lip}(f_{D_n})} \left| \int_{\mathbb{R}^+} f_{D_n} \mathrm{d}\tilde{P}_n - \int_C f \mathrm{d}P \right| \to_{\text{a.s.}} 0. \tag{4.13}$$

If furthermore D_n is chosen such that equation (4.10) is satisfied, then the factor $\frac{K_n}{\text{Lip}(f_{D_n})}$ is eventually bounded away from zero, and can be eliminated from equation (4.13) to lead to the convergence of the estimator. □

Of course, there may be more than one way in which one could construct $\tilde{P}_n$. In this paper, we focus on demonstrating the validity and usefulness of the methodology by providing two possible constructions. The results would remain valid regardless to the specific construction, and other constructions boasting more appealing properties, such as rates of convergence under more lenient assumptions, are welcome future contributions to this framework.

■ 4.4 The Good-Turing Pseudo-Empirical Measure

■ 4.4.1 Definitions and Properties

The platform on which we build our estimation scheme is the Good-Turing estimator, and in particular its strong consistency in this scaling regime, as established by WVK. In this section, we recall the main definition and review the properties relevant to the rest of the development. Once again, let $B_{n,k}$ be the subset of symbols of A_n that appear exactly k times in the samples $X_{n,1}, \cdots, X_{n,n}$. Then we have, for each $k = 0, 1, \cdots, n$, the probability $M_{n,k} = p_n(B_{n,k})$ of all symbols that appear exactly k times. We can write these in the infinite vector form with the notation $M_{n,\cdot} \equiv \{M_{n,k}\}_{k \in \mathbb{N}_0}$, which we pad with zeros for $k > n$. In particular, we restate the Good-Turing estimator:

Definition 4.5. *Recall $K_{n,k} = |B_{n,k}|$ be the number of symbols of A_n that appear k times in $X_{n,1}, \cdots, X_{n,n}$. Write the* Good-Turing estimator *of $M_{n,\cdot}$ as a vector $G_{n,\cdot} \equiv \{G_{n,k}\}_{k \in \mathbb{N}_0}$, for each $k \in \mathbb{N}_0$, where recall that*

$$G_{n,k} = \frac{(k+1)K_{n,k+1}}{n}. \tag{4.14}$$

WVK establish a host of convergence properties for the Good-Turing estimation problem and the Good-Turing estimator. We group these in the following theorem.

Theorem 4.4. *Define the Poisson P-mixture $\lambda \equiv \{\lambda_k\}_{k\in\mathbb{N}_0}$ as, for each $k \in \mathbb{N}_0$:*

$$\lambda_k = \int_C \frac{x^k e^{-x}}{k!} \mathrm{d}P(x). \tag{4.15}$$

We then have the following results that determine the limiting behavior of $M_{n,\cdot}$, and the strong consistency of the Good-Turing estimator $G_{n,\cdot}$:

(i) We have that $M_{n,k} \to_{\text{a.s.}} \lambda_k$ and $G_{n,k} \to_{\text{a.s.}} \lambda_k$, and therefore $|G_{n,k} - M_{n,k}| \to_{\text{a.s.}} 0$, pointwise for each $k \in \mathbb{N}_0$ as $n \to \infty$.

(ii) By Scheffé's theorem ([3], p. 215), it also follows that these convergences hold in L_1 almost surely, in that $\|M_{n,\cdot} - \lambda\|_1 \to_{\text{a.s.}} 0$ and $\|G_{n,\cdot} - \lambda\|_1 \to_{\text{a.s.}} 0$, and therefore $\|G_{n,\cdot} - M_{n,\cdot}\|_1 \to_{\text{a.s.}} 0$, as $n \to \infty$.

■ 4.4.2 Empirical Measure Analogy

The analogy that we would like to make in this section is the following. Assuming λ is given, one could take n i.i.d. samples from it, and form the empirical measure or the type, call it $\hat{\lambda}_n \equiv \{\hat{\lambda}_{n,k}\}_{k\in\mathbb{N}_0}$. Such an empirical measure would satisfy well-known statistical properties, in particular the strong law of large numbers would apply, and we would have $\hat{\lambda}_{n,k} \to_{\text{a.s.}} \lambda_k$. By Scheffé's theorem, L_1 convergence would also follow. It is evident from Theorem 4.4 that despite the fact that we do not have such a true empirical measure, the Good-Turing estimator $G_{n,\cdot}$ behaves as one, and we may be justified to call it a *pseudo-empirical measure*.

Now observe that since, for discrete distributions, the total variation distance is related to the L_1 distance by $\sup_{B\subset\mathbb{N}_0} |\hat{\lambda}_n(B) - \lambda(B)| = \frac{1}{2}\|\hat{\lambda}_n - \lambda\|_1$, the true empirical measure also converges in total variation. As a special case, the Glivenko-Cantelli theorem applies in that $\sup_k |F(k;\lambda) - F(k;\hat{\lambda}_n)| \to_{\text{a.s.}} 0$. Recall that $F(.;.)$ denotes the cumulative of the second argument (a measure) evaluated at the first argument. In light of the above, this remains valid for the pseudo-empirical measure. However, for the classical empirical measure, we also have the *rate* of convergence in the Glivenko-Cantelli theorem, in the form of the Kolmogorov-Smirnov theorem and its variants for discrete distributions, see for example [32]. Such results are often formulated in terms of a convergence in probability of rate $\frac{1}{\sqrt{n}}$. So we next ask whether such rates hold for the pseudo-empirical measure as well.

We first note that the rare-events source model is lenient, in the sense that it does not impose any convergence rate on $P_n \Rightarrow P$. Therefore, convergence results that aim to parallel those of a true empirical measure will depend on assumptions on the rate of this core convergence. In particular, let us assume that we know something about the weak convergence rate of P_n to P in terms of the Wasserstein distance, in that we assume there exists an $r > 0$ such that

$$n^r d_{\mathrm{W}}(P_n, P) \to 0.$$

For example, in Lemma 4.7, we will show that this holds true for a class of rare-events sources suggested by WVK.

Next, note that Lemma 11 in WVK gives the following useful concentration rate for the pseudo-empirical measure around its mean.

Lemma 4.5. *For any $\delta > 0$, $n^{1/2-\delta}\|G_{n,\cdot} - \mathbf{E}[G_{n,\cdot}]\|_1 \to_{\text{a.s.}} 0$.*

In the following statement, we show that a Kolmogorov-Smirnov-type convergence to λ does hold for the pseudo-empirical measure $G_{n,\cdot}$, with a rate that is essentially the slower of that of the concentration of Lemma 4.5 and that of the rare-events source itself.

Theorem 4.6. *Let $r > 0$ be such that $n^r d_{\mathrm{W}}(P_n, P) \to 0$. Then for any $\delta > 0$, we have:*

$$n^{\min\{r,\,1/2\}-\delta} \sup_k |F(k;\lambda) - F(k;G_{n,\cdot})| \to_{\text{a.s.}} 0. \tag{4.16}$$

Proof. For convenience, define $B_k \equiv \{0, \cdots, k\}$. The proof requires three approximations. The first is to approximate $G_{n,\cdot}$ with $\mathbf{E}[G_{n,\cdot}]$. This is already achieved using Lemma 4.5. Since the L_1 distance is twice the total variation distance, and specializing to the subsets B_k, we have that for all $\delta > 0$:

$$n^{1/2-\delta} \sup_k |F(k;\mathbf{E}[G_{n,\cdot}]) - F(k;G_{n,\cdot})| \to_{\text{a.s.}} 0. \tag{4.17}$$

The next two approximations are *(a)* to approximate $\mathbf{E}[G_{n,\cdot}]$ with a Poisson P_n-mixture (using the theory of Poisson approximation), and *(b)* to approximate the latter with λ, which is a Poisson P-mixture (using the convergence in $d_{\mathrm{W}}(P_n, P)$).

Part (a) – For convenience, let π_n be a Poisson(x) P_n-mixture, and let η_n be a Binomial $\left(\frac{x}{n}, n\right)$ P_n-mixture. One can show, as in the proof of Lemma 7 of WVK, that $\mathbf{E}[G_{n,\cdot}]$ is a Binomial $\left(\frac{x}{n}, n-1\right)$ P_n-mixture. We first relate $\mathbf{E}[G_{n,\cdot}]$ to η_n which is the natural candidate for Poisson approximation. We then use Le Cam's theorem to relate η_n to π_n.

We start with a general observation. Let $\mathscr{F} = \{f(\cdot;x) : x \in C\}$ and $\mathscr{G} = \{g(\cdot;x) : x \in C\}$ be two parametric classes of probability mass functions over $\mathbb{N}_0$, e.g. Poisson and Binomial, and let Q be a mixing distribution supported on C. Say that for some subset $B \subset \mathbb{N}_0$, we have the pointwise bound $|f(B;x) - g(B;x)| \leq \ell(x)$. It follows that the mixture of the bound is also a bound on the mixture. More precisely:

$$\left| \int_C f(B;x)dQ(x) - \int_C g(B;x)dQ(x) \right| \leq \int_C \ell(x)dQ(x). \tag{4.18}$$

Note that if the pointwise bound above holds uniformly over B, then the same is true for the mixture bound. We will use this particularly with the subsets B_k, to bound the difference of cumulative distribution functions.

Now let $g_n(k;x)$ be the c.d.f. of a Binomial $\left(\frac{x}{n}, n\right)$ random variable, and let $\tilde{g}_n(k;x)$ be the c.d.f. of a Binomial $\left(\frac{x}{n}, n-1\right)$ random variable. For any given k, we have the following:

$$\begin{aligned}
&\left(1-\frac{x}{n}\right)\tilde{g}_n(k;x) \\
&= \sum_{m=0}^{k} \frac{n-m}{n}\binom{n}{m}\left(\frac{x}{n}\right)^m\left(1-\frac{x}{n}\right)^{n-m} \\
&= g_n(k;x) - \frac{1}{n}\sum_{m=0}^{k} m\binom{n}{m}\left(\frac{x}{n}\right)^m\left(1-\frac{x}{n}\right)^{n-m}.
\end{aligned}$$

Using the facts that the sum is no larger than the mean and that $\tilde{g}_n(k;x) \leq 1$, it follows that for any given k we have:

$$\begin{aligned}
&|g_n(k;x) - \tilde{g}_n(k;x)| \\
&= \left|\frac{1}{n}\sum_{m=0}^{k} m\binom{n}{m}\left(\frac{x}{n}\right)^m\left(1-\frac{x}{n}\right)^{n-m} - \frac{x}{n}\tilde{g}_n(k;x)\right| \\
&\leq \frac{x}{n}
\end{aligned}$$

Note that $\int_C g_n(k;x)\mathrm{d}P_n = F(k;\eta_n)$, the c.d.f. of η_n, and $\int_C \tilde{g}_n(k;x)\mathrm{d}P_n = F(k;\mathbf{E}[G_{n,\cdot}])$, the c.d.f. of $\mathbf{E}[G_{n,\cdot}]$. Using the observation leading to equation (4.18), it follows that:

$$\sup_k |F(k;\mathbf{E}[G_{n,\cdot}]) - F(k;\eta_n)| \leq \frac{1}{n}\int_C x\mathrm{d}P_n(x) \leq \frac{\hat{c}}{n}. \tag{4.19}$$

Using Le Cam's theorem (see, for example, [28]), we know that the total variation distance, and hence the difference of probabilities assigned to any subset $B \subset \mathbb{N}_0$ by a Poisson(x) distribution and a Binomial $\left(\frac{x}{n}, n\right)$ distribution is upper-bounded by $\frac{x^2}{n}$. We apply this to the subsets B_k, and use the observation leading to equation (4.18) once again to extend this result to the respective P_n-mixtures:

$$\sup_k |F(k;\pi_n) - F(k;\eta_n)| \leq \frac{1}{n}\int_C x^2\mathrm{d}P_n(x) \leq \frac{\hat{c}^2}{n}. \tag{4.20}$$

By combining equations (4.19) and (4.20), we deduce that for all $\delta > 0$:

$$n^{1-\delta}\sup_k |F(k;\mathbf{E}[G_{n,\cdot}]) - F(k;\pi_n)| \to 0. \tag{4.21}$$

Part (b) – Now let $h(k;x)$ be the c.d.f. of a Poisson(x) random variable. Observe that:

$$0 \leq \frac{\mathrm{d}}{\mathrm{d}x}h(k;x) = \sum_{m=0}^{k} -\frac{x^m e^{-x}}{m!} + m\frac{x^{m-1}e^{-x}}{m!}$$

$$\leq \frac{1}{x}\sum_{m=0}^{k} m\frac{x^m e^{-x}}{m!} = \frac{1}{x}\mathbf{E}\left[\text{Poisson}(x)\right] = 1.$$

Therefore, when viewed as a function of x, $h(k;x)$ is a Lipschitz(1) function on C for all k. Using the dual representation of the Wasserstein distance, we then have:

$$\begin{aligned}
&\sup_k |F(k;\pi_n) - F(k;\lambda)| \\
&\quad = \sup_k \left| \int_C h(k;x)\mathrm{d}P_n(x) - \int_C h(k;x)\mathrm{d}P(x) \right| \\
&\quad \leq \sup_{h\in\text{Lipschitz}(1)} \left| \int_C h\mathrm{d}P_n - \int_C h\mathrm{d}P \right| = d_W(P_n, P).
\end{aligned}$$

Using the assumption of the convergence rate of P_n to P, it follows that for all $\delta > 0$ we have:

$$n^{r-\delta}\sup_k |F(k;\pi_n) - F(k;\lambda)| \to 0. \tag{4.22}$$

The statement of the theorem follows by combining equations (4.17), (4.21), and (4.22). □

In a practical situation, one would expect that the rare-events source is well-behaved enough that $r > 1/2$, and that the bottleneck of Theorem 4.6 is given by the 1/2 rate, and therefore we have a behavior that more closely parallels a true empirical measure. Indeed, some natural constructions obey this principle. Most trivially, for a sequence of uniform sources, e.g. if $p_n(a) = 1/n$, we have $P_n = P$, and therefore $r = \infty$. More generally, consider the following class of rare-events sources suggested by WVK.

Definition 4.6. *Let g be a density on $[0,1]$ that is continuous Lebesgue almost everywhere, and such that $\check{c} \leq g(w) \leq \hat{c}$ for all $w \in [0,1]$. Let $A_n = \{1, \cdots, \lfloor \alpha n \rfloor\}$ for some $\alpha > 0$, and for every $a \in A_n$ let $p_n(a) = \int_{(a-1)/\lfloor \alpha n \rfloor}^{a/\lfloor \alpha n \rfloor} g(w)\mathrm{d}w$. One can then verify that $\{(A_n, p_n)\}$ is indeed a rare-events source, with P being the law of $g(W)$, where $W \sim g$. We call such a construction a* rare-events source obtained by quantizing *g.*

Lemma 4.7. *Let g be a density as in Definition 4.6, and let $\{(A_n, p_n)\}$ be a rare-events source obtained by quantizing g. If g has finitely many discontinuities, and is Lipschitz within each interval of continuity, then for all $r < 1$:*

$$n^r d_{\mathrm{W}}(P_n, P) \to 0$$

Proof. Without loss of generality, assume $\alpha = 1$, and that the largest Lipschitz constant is 1. Consider the quantized density on $[0,1]$:

$$g_n(w) = n\int_{(\lceil wn \rceil - 1)/n}^{\lceil wn \rceil/n} g(v)\mathrm{d}v,$$

where the integral is against the Lebesgue measure. Then it follows that P_n is the law of $g_n(W_n)$, where $W_n \sim g_n$.

Say g has L discontinuities, and let D_n be the union of the L intervals of the form $[(a-1)/n, a/n]$ which contain these discontinuities. In all other intervals, we have that $|g(w) - g_n(w)| \leq 1/n$, using Lipschitz continuity and the intermediate value theorem. It follows that

$$\begin{aligned}
&\int_{[0,1]} |g(w) - g_n(w)| \mathrm{d}w \\
&\quad = \int_{D_n} |g - g_n| \mathrm{d}w + \int_{[0,1] \setminus D_n} |g - g_n| \mathrm{d}w \leq \frac{L}{n} + \frac{1}{n}.
\end{aligned}$$

For any particular $x \in C$, let $B_x = \{w \in [0,1] : g(w) < x\}$. We then have

$$\begin{aligned}
|F(x; P_n) - F(x; P)| &= \left| \int_{B_x} g(w) - g_n(w) \mathrm{d}w \right| \\
&\leq \int_{B_x} |g(w) - g_n(w)| \mathrm{d}w \leq \frac{L+1}{n}.
\end{aligned}$$

By integrating over all x:

$$d_{\mathrm{W}}(P_n, P) = \int_C |F(x; P_n) - F(x; P)| \mathrm{d}x \leq \frac{(L+1)(\hat{c} - \check{c})}{n}.$$

Therefore the lemma follows. □

We end by remarking that the rare-events sources covered by Lemma 4.7 are rather general in nature. For example, all of the illustrative and numerical examples offered by WVK are special cases (more precisely, they have piecewise-constant g).

■ 4.5 Constructing $\tilde{P}_n$ via Mixing Density Estimation

We would now like to address the task of using $X_{n,1}, \cdots, X_{n,n}$ to construct a sequence of probability measures $\tilde{P}_n$ that, for almost every sample sequence, converges weakly to P, as outlined in Section 4.3. Since we have established the Good-Turing estimator as a pseudo-empirical measure issued from a Poisson P-mixture, in both consistency and rate, this is analogous to a mixture density estimation problem, with the true empirical measure replaced with the Good-Turing estimator $G_{n,\cdot}$.

We start by noting that the task is reasonable, because the mixing distribution in a Poisson mixture is identifiable from the mixture itself. This observation can be traced back to [29] and [13]. Then, the first natural approach is to use non-parametric maximum likelihood estimation. In Section 4.5.1, we use Simar's work in [27] to construct a valid estimator in this framework. Unfortunately, to the best of the authors' knowledge, the maximum likelihood estimator does not have a well-studied rate of convergence on the recovered mixing distribution. In Section 4.5.2 we consider instead a minimum distance estimator, with which Chen gives optimal rates of convergence in [6], albeit by assuming finite support for P.

4.5.1 Maximum Likelihood Estimator

We first define the maximum likelihood estimator in our setting. Despite the fact that it is not, strictly speaking, maximizing a true likelihood, we keep this terminology in light of the origin of the construction.

Definition 4.7. *Given the pseudo-empirical measure (Good-Turing estimator) $G_{n,\cdot}$ the* maximum likelihood estimator *of the mixing distribution is a probability measure $\tilde{P}_n^{ML}$ on $\mathbb{R}^+$ which maximizes the pseudo-likelihood as follows:*

$$\tilde{P}_n^{ML} \in \operatorname*{argmax}_{Q} \sum_{k=0}^{\infty} G_{n,k} \log \left(\int_0^{\infty} \frac{x^k e^{-x}}{k!} \mathrm{d}Q(x) \right). \tag{4.23}$$

It is not immediately clear whether $\tilde{P}_n^{\mathrm{ML}}$ exists or is unique. These questions were answered in the affirmative in [27]. On close examination, it is clear that these properties do not depend on whether we are using a pseudo-empirical measure instead of a true empirical measure. Hence they remain valid in our context. Next, we establish the main consistency statement.

Theorem 4.8. *For almost every sample sequence, the sequence $\{\tilde{P}_n^{ML}\}$ converges weakly to P as $n \to \infty$. We write this as $\tilde{P}_n^{ML} \Rightarrow_{\text{a.s.}} P$.*

Proof. The main burden of proof is addressed by Theorem 4.4 in establishing the strong law of large numbers for the pseudo-empirical measure, and which is originally given in WVK's Proposition 7. Indeed, in Simar's proof ([27], Section 3.3, pp. 1203–1204), we only use the fact that $G_{n,k} \to_{\text{a.s.}} \lambda_k$ for every $k \in \mathbb{N}_0$. The rest of the proof carries over, and the current theorem follows. □

It is worth noting that the consistency of the maximum likelihood estimator does not even require that condition (i) in the Definition 4.1 of the rare-events source to hold, since Theorem 4.4 in fact holds without that condition. In that sense, it is very general. However, when every neighborhood of 0 or ∞ has positive probability under P, it limits the types of functions that we can allow in the canonical problems, including sequence probabilities and entropies as discussed in WVK. When P is not compactly supported, it is also difficult to establish the rates of convergence.

4.5.2 Minimum Distance Estimator

We now define a minimum distance estimator for our setting. The reason that we suggest this alternate construction of $\tilde{P}_n$ is that it is useful to quantify the convergence rate to P, and the minimum distance estimator provides such a rate. However, it does so with the further assumption that P has a finite support, whose size is bounded by a known number m.

Also note that the definition of the estimator circumvents questions of existence by allowing for a margin of ϵ from the infimum, and does not necessarily call for uniqueness.

Definition 4.8. *For a probability measure Q on $\mathbb{R}^+$, let $\pi(Q)$ denote the Poisson Q-mixture. Then, given the pseudo-empirical measure $G_{n,\cdot}$, a* minimum distance estimator *with precision ϵ is any probability measure $\tilde{P}_n^{\mathrm{MD},m,\epsilon}$ on $\mathbb{R}^+$ that satisfies*

$$\begin{aligned}\sup_k & \left| F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon})) - F(k; G_{n,\cdot}) \right| \\ & \leq \inf_Q \sup_k |F(k; \pi(Q)) - F(k; G_{n,\cdot})| + \epsilon,\end{aligned}$$

where the infimum is taken on probability measures supported on at most m points, on $\mathbb{R}^+$.

We now provide the main consistency and rate results associated with such estimators.

Theorem 4.9. *Let $r > 0$ be such that $n^r d_{\mathrm{W}}(P_n, P) \to 0$, and assume that it is known that P is supported on at most m points. Let $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}$ be a sequence of minimum distance estimators chosen such that $\epsilon_n < n^{-\min\{r,1/2\}}$. Then as $n \to \infty$, we have that for any $\delta > 0$:*

$$n^{\min\{r/2,1/4\}-\delta} d_{\mathrm{W}} \left(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}, P \right) \to_{\mathrm{a.s.}} 0. \tag{4.24}$$

Remark: Since d_{W} induces the weak convergence topology, it also follows that $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n} \Rightarrow_{\mathrm{a.s.}} P$.

Proof. To derive rate results in [6], Chen establishes a bound on the Wasserstein distance between mixing distributions, using the Kolmogorov-Smirnov distance between the c.d.f.s of the resulting mixtures. For this, he first introduces a notion of strong identifiability (Definition 2, p. 225), and shows that Poisson mixtures satisfy it (Section 4, p. 228). He then shows (in Lemma 2, p. 225) that if we have strongly identifiable mixtures and if two mixing distributions have a support of at most m points within a fixed compact set, such as C, then we can find a constant M (which depends non-constructively on m and C), such that for any two such mixing distributions Q_1 and Q_2, we have:

$$d_{\mathrm{W}} (Q_1, Q_2)^2 \leq M \sup_k |F(k; \pi(Q_1)) - F(k; \pi(Q_2))| \tag{4.25}$$

The main burden of proof therefore falls on our Theorem 4.6 in establishing a Kolmogorov-Smirnov-type convergence for the pseudo-empirical measure. The argument we present next is based on Chen's proof (Theorem 2, p. 226). We have:

$$\begin{aligned}\sup_k & \left| F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k; G_{n,\cdot}) \right| \\ & \leq \sup_k \left| F(k; \pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k; \lambda) \right| \\ & \quad + \sup_k |F(k; \lambda) - F(k; G_{n,\cdot})|\end{aligned}$$

$$\leq \quad 2\sup_k |F(k;\lambda) - F(k;G_{n,\cdot})| + \epsilon_n,$$

where the final inequality is due to the definition of $\tilde{P}_n^{\mathrm{MD},m,\epsilon_n}$. By Theorem 4.6, and by our choice of ϵ_n, it follows that for all $\delta > 0$, we have:

$$n^{\min\{r,1/2\}-2\delta} \sup_k \left| F(k;\pi(\tilde{P}_n^{\mathrm{MD},m,\epsilon_n})) - F(k;G_{n,\cdot}) \right| \to_{\text{a.s.}} 0. \tag{4.26}$$

By combining (4.25) and (4.26), the theorem follows . □

Note that Chen's result can be used to show more. In particular, if we think of the true mixing distribution as residing in some neighborhood of a fixed distribution, then the convergence holds uniformly over that neighborhood. This may be interpreted as a form of robustness, but we do not dwell on it here.

■ 4.6 Applications

To solve canonical problems in the setting of Lemma 4.2, when an a priori bound on C is known or when f is bounded on $\mathbb{R}^+$, it suffices to construct a sequence of probability measures $\tilde{P}_n$ that weakly converges to P for almost every sample sequence. Since Theorem 4.8 provides such a sequence, we need not go further than that.

However, to work within the more general setting of Lemma 4.3, where no knowledge of C is assumed and f can be any locally Lipschitz function, we can use the result of Theorem 4.9. In this section, we start by illustrating this for some of the quantities considered by WVK. We then suggest two new applications: alphabet size and support interval estimation. We conclude by remarking on some algorithmic considerations.

■ 4.6.1 Estimating Entropies and Probabilities

First consider the entropy of the source $H(p_n)$, and the associated problem, in normalized form, of estimating $Y_n^H \equiv H(p_n) - \log n$. One can then write:

$$Y_n^H = -\int_C \log x \mathrm{d}P_n(x),$$

and therefore, by comparing to equation (4.1) with $f_n(x) = f(x) = -log(x)$, we have a canonical estimation problem, and since Y_n^H is deterministic, it is also strong. If we have a bound on C, we can use Lemma 4.2. Otherwise, note that on intervals of the form $[D^{-1}, D]$, $\log x$ is D-Lipshitz. Therefore if for some $s > 0$, $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$, as given by Theorem 4.9 for example, then we can apply Lemma 4.3 using $D_n = n^s$. If s exists but is unknown, we can still apply Lemma 4.3 using any sequence that is $o(n^s)$, such as $D_n = e^{\log^\epsilon n}$, for some $\epsilon > 0$. The consistent estimator becomes:

$$\hat{Y}_n^H \equiv -\int_{\mathbb{R}^+} \log_{D_n} x \mathrm{d}\tilde{P}_n(x). \tag{4.27}$$

Next consider the probability of the sequence $p_n(X_{n,1}, \cdots, X_{n,n})$, and the associated normalized problem of estimating $Y_n^p \equiv \frac{1}{n}\log p_n(X_{n,1}, \cdots, X_{n,n}) + \log n$. We have (WVK, Lemma 5):

$$\begin{aligned} \mathbf{E}[Y_n^p] &= \mathbf{E}[\log p_n(X_n)] + \log n \\ &= \int_C \log x \mathrm{d}P_n(x), \end{aligned}$$

and therefore we also have a canonical estimation problem. Using McDiarmid's theorem, one can also show that (WVK, Lemma 6) $|\mathbf{E}[Y_n^p] - Y_n^p| \to_{\text{a.s.}} 0$, and therefore we once again have a strong canonical estimation problem, and we can construct a consistent estimator as in the case of entropy. Referring to equation (4.27), we have $\hat{Y}_n^p \equiv -\hat{Y}_n^H$.

4.6.2 Estimating the Alphabet Size

Consider the size of the alphabet $|A_n|$. Since the model describes large, asymptotically infinite, alphabets, we look at the normalized problem of estimating $Y_n^A = |A_n|/n$. We have (cf. [2]):

$$\begin{aligned} Y_n^A &= \frac{1}{n}\sum_{a\in A} 1 = \sum_{a\in A} \frac{p_n(a)}{np_n(a)} \\ &= \int_C \frac{1}{x}\mathrm{d}P_n(x). \end{aligned}$$

Once again, having a deterministic sequence of the form of (4.1) with $f_n(x) = f(x) = 1/x$, it follows that $\{Y_n^A\}_{n\in\mathbb{N}}$ is a strong canonical problem. If we have a bound on C, we can use Lemma 4.2. Otherwise, note that on intervals of the form $[D^{-1}, D]$, $1/x$ is D^2-Lipshitz. Therefore if for some $s > 0$, $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$, as given by Theorem 4.9 for example, then we can apply Lemma 4.3 using $D_n = n^{s/2}$. As in Section 4.6.1, if s exists but is unknown, we can still apply Lemma 4.3 using any sequence that is $o(n^s)$, such as $D_n = e^{\log^\epsilon n}$, for some $\epsilon > 0$. The consistent estimator becomes:

$$\hat{Y}_n^A \equiv \int_{\mathbb{R}^+} x_{D_n}^{-1} \mathrm{d}\tilde{P}_n(x). \tag{4.28}$$

4.6.3 Estimating the Support Interval

As discussed in Section 4.2.2, estimating the support interval is not a canonical problem per se. However, we show here that we can extend the framework in a straightforward fashion to provide consistent estimators of both $\check{c}$ and $\hat{c}$.

Lemma 4.10. *Let $\tilde{P}_n \Rightarrow_{\text{a.s.}} P$ such that for some $s > 0$, we have $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$. This is particularly true under the conditions of Theorem 4.9. Given $q \neq 0$ and $D \geq 1$, let x_D^q denote the D-tapered version of x^q.*

If $q_n = \log n/\log\log n$ and $D_n = n^{s/(2q_n)}$, then we have: as $n \to \infty$,

$$\begin{aligned} \left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \mathrm{d}\tilde{P}_n(x)\right)^{1/q_n} &\quad \to_{\text{a.s.}} \quad \check{c} \\ \textit{and} \quad \left(\int_{\mathbb{R}^+} x_{D_n}^{q_n} \mathrm{d}\tilde{P}_n(x)\right)^{1/q_n} &\quad \to_{\text{a.s.}} \quad \hat{c}. \end{aligned}$$

Proof. For conciseness, let us drop the argument of the probability measures, and write $\mathrm{d}P$ for $\mathrm{d}P(x)$. We provide the proof only for $\check{c}$, since the argument is analogous for $\hat{c}$. Recall that $\check{c}$ is the essential infimum of a random variable $Z \sim P$. Therefore, for any $D \geq (\check{c}^{-1} \vee \hat{c})$, we have:

$$\left(\int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}P\right)^{1/q} \to \check{c} \qquad \text{as } q \to \infty. \tag{4.29}$$

In the absence of a rate of convergence, we cannot simply plug in $\tilde{P}_n$. But since we know that $n^s d_{\mathrm{W}}(\tilde{P}_n, P) \to_{\text{a.s.}} 0$, we can use the dual representation of the Wasserstein distance and the fact that for every q and D the function $\frac{1}{q} D^{-1-q} x_D^{-q}$ is Lipschitz(1) over $\mathbb{R}^+$ to state: as $n \to \infty$,

$$n^s \sup_{q,D} \frac{D^{-1-q}}{q} \left| \int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}P \right| \to_{\text{a.s.}} 0. \tag{4.30}$$

We now want to relate this to the difference of the q^{th} roots. Note that each of the integrals in (4.30) is bounded from below by D^{-q}. Using this and the fact that for any a and $b > 0$ we have $\left|a^{1/q} - b^{1/q}\right| \leq \frac{1}{q}(a \wedge b)^{\frac{1}{q}-1} |a - b|$, we can write:

$$\begin{aligned} &\left| \left(\int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}\tilde{P}_n\right)^{1/q} - \left(\int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}P\right)^{1/q} \right| \\ &\qquad \leq \quad D^{2q} \cdot \ \frac{D^{-1-q}}{q} \left| \int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_D^{-q} \mathrm{d}P \right|. \end{aligned}$$

The choices $q_n = \log n / \log\log n$ and $D_n = n^{s/(2q_n)}$, allow us to have $D_n^{2q_n} = n^s$, and yet guarantee that as $n \to \infty$ both q_n and $D_n \to \infty$. With this, we can use the convergence of equation (4.30), to state: as $n \to \infty$,

$$\begin{aligned} &\left| \left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \mathrm{d}\tilde{P}_n\right)^{1/q_n} - \left(\int_{\mathbb{R}^+} x_{D_n}^{-q_n} \mathrm{d}P\right)^{1/q_n} \right| \\ &\qquad \leq n^s \frac{D_n^{-1-q_n}}{q_n} \left| \int_{\mathbb{R}^+} x_{D_n}^{-q_n} \mathrm{d}\tilde{P}_n - \int_{\mathbb{R}^+} x_{D_n}^{-q_n} \mathrm{d}P \right| \to_{\text{a.s.}} 0. \end{aligned} \tag{4.31}$$

We then combine (4.29) and (4.31) to complete the proof. □

Remarks. Note the following:

(i) Other scaling schemes can be devised for q_n and D_n, as long as they both grow to ∞ as $n \to \infty$, yet $D_n^{2q_n}$ remains at most $\mathcal{O}(n^s)$.

(ii) If a bound $[D_{\min}, D_{\max}] \supset C$ is already known, then we can taper x^q accordingly, without growing D_n. In this case, we can also speed up the rate of convergence by choosing $q_n = \frac{s}{2} \log n / \log \frac{D_{\max}}{D_{\min}}$.

(iii) If only an upper bound or only a lower bound is known, we can taper x^q accordingly, and only grow/shrink the missing bound. In this case we leave $q_n = \log n / \log \log n$ as in the Lemma.

(iv) In the Lemma and the alternatives in these remarks, if s is unknown we can replace it wherever it appears (together with constant factors) with a suitably decaying term, that guarantees the behavior of remark (i). For example, in the Lemma, we can choose $D_n = n^{1/(q_n \sqrt{\log \log n})}$, since then $D_n^{2q_n}$ becomes $o(n^s)$ for any s, and the proof applies.

4.6.4 Algorithmic Considerations

One of the appealing properties of the maximum likelihood estimator is that, by a result of Simar in [27], it is supported on finitely many points. Simar also suggests a particular algorithm for obtaining the $\tilde{P}_n^{\mathrm{MLE}}$, the convergence of which was later established in [5], with further improvements. One can also solve for the MLE using the EM algorithm, as reviewed in [25]. Penalized variants are also suggested, such as in [20]. The literature on the non-parametric maximum likelihood estimator for mixtures is indeed very rich. As for the minimum distance estimator, in [6] Chen suggests variants of the work in [10], where they use algorithms based on linear programming.

Chapter 5

Conclusion

■ 5.1 Summary

In this thesis, we considered the problem of estimating the total probability of symbols that appear very rarely in successive independent samples from a discrete distribution. We closely examined the conditions under which one can meaningfully make inferences about these rare probabilities. We first gave evidence that the bias and additive concentration properties of the classical Good-Turing estimator of the missing mass are distinctly non-trivial only when in a heavy-tailed regime. For this we proposed the notion of accrual rate, which describes how probability accrues up from the rarest symbols. This explains the success of Good-Turing estimation in natural language modeling, where heavy tails are the norm.

We then focused on consistent estimation of rare probabilities. We showed that, here too, heavy-tailed distributions play an important role. For this, we used the notion of regular variation, which is a restricted but more precise version of accrual rate. As an illustration to what happens in the absence of regularly varying heavy tails, we showed that even for a geometric distribution Good-Turing is not consistent. On the other hand, under heavy-tailed regular variation, we showed that Good-Turing estimation is always consistent. We established this by showing multiplicative concentration for the rare probabilities, also extending existing additive concentration results in the process. Lastly, within this regime, we constructed a family of consistent estimators that address some of the shortcomings of the Good-Turing estimator. Since the core mechanism of this approach is to estimate the regular variation index, this brings rare probability estimation in this discrete setting closer to the continuous setting of tail estimation and extreme value theory. Cross-fertilization between these frameworks is now possible, in particular by using semi-parametric approaches to index estimation.

Lastly, we studied an alternative model where instead of a fixed distribution, one samples from a distribution whose support increases with the number of samples. This captures situations of large alphabets, where the underlying sample space is finite but comparable in size to the number of samples. Every event in this model is rare, and one may call it a ‘rare events’ regime. In this context, we considered a large class of canonical estimation problems, including entropy, sample size, and probability range estimation. We presented an abstract solution methodology for consistent estimation, based on the construction of a sequence of random measures that almost surely converge weakly to a distribution that characterizes the rare events regime. We then cast the Good-Turing estimator as a pseudo-empirical measure from a Poisson mixture, and used mixture density estimation to give two concrete constructions of the desired sequence of measures, thus establishing consistent estimators for the quantities of interest.

■ 5.2 Future Work

■ 5.2.1 Functional Convergences

In the context of the infinite occupancy model as a framework for modeling discrete rare events, there are several directions, which can result in radically better inference methods. As we have seen, since regular variation is the natural framework, estimating its index, α, is central. We have suggested some simple consistent estimators $\hat{\alpha}$ for the index. However if one wants to use $\hat{\alpha}$ in estimators for quantities beyond total probabilities, such as entropies as we did in the scaling context, or in statistical decision rules, one needs to characterize the convergence rates of the strong laws. Concentration, large deviation, and Berry-Esseen bounds for the CLTs are pertinent in this regard. We expect such work to capitalize on the modern treatment of second order regular variation conditions, [1].

But a more powerful approach to performing inference in this context would be to characterize convergence via a functional CLT. In particular, it is desirable to extend the horizon of r as $n \to \infty$ at the appropriate growth speed, e.g. $R_n \to \infty$, and give uniform convergence rates for the vector $(K_{n,r})_{r=1,\cdots,R_n}$. This can result in an elegant parallel to empirical process theory, and can enable semiparametric inference to obtain estimators, for α and other quantities, that take advantage of more of the available information in the sample. Such an approach further tightens the parallel with inference in extreme value theory, [1], as nonparametric tail-index estimators rely on extending-horizon, intermediate order statistics.

■ 5.2.2 Dependent Data

One could also consider the Good-Turing estimation problem in the presence of Markov observations, that is the problem of estimation of the stationary probability of discrete unobserved or rarely observed states. A primary motivation to study this extension is to gauge the risk of a dynamic system evolving into a hitherto unseen state. Despite the fact that many applied areas where the Good-Turing estimator is used have intrinsically dependent observations, as in the case of speech and language modeling, this problem has not had any direct treatment. For example, in n-gram models, one reverts back to the conditional independence structure to perform inference in the classical Good-Turing setting, coupled with smoothing techniques that layer out n-gram models of varying orders. One can show that naive extensions of i.i.d. estimators to Markov chains lose many of the original performance guarantees, even the property of asymptotic unbiasedness. However, when certain, such as Doeblin-type, conditions are imposed on the chain, some of these properties may be recovered. Since many well-behaved models do satisfy these assumptions, such results form the basis of a rigorous generalization.

■ 5.2.3 General Rare Event Models

We have seen in this thesis that in the absence of structural restrictions, rare event models may be ill-posed for inference. In particular, in the absence of regularly varying heavy tails, we have found that even the simple example of a geometric distribution may lead to inconsistent estimates. This is in the same spirit of the importance of regular variation in estimating tail probabilities in extreme value theory. On the other hand, we have seen that other structures, such as the scaling regime of [31], offer an alternative to fixed-distribution sampling, and that not only can one perform rare probability estimation in such contexts, but also solve a larger class of inference problems. Yet another approach to modeling rare events is to capture them in terms of requirements on data for a given estimation task, which we may call the sample-complexity perspective. This is done for example in the recent work of [30]. Therefore, an exciting open problem is to relate fixed-distribution sampling, scaling methods, and sample-complexity approaches. The goal of making this correspondence precise is to identify the universal structures that are both necessary and sufficient for inference about rare events.

“ *Intentio hominis totum facit possibile.* ”

Bibliography

[1] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications.* Wiley, 2004.

[2] S. Bhat and R. Sproat. Knowing the unseen: estimating vocabulary size over unseen samples. In *Proceedings of the ACL*, pages 109–117, Suntec, Singapore, August 2009.

[3] P. Billingsley. *Probability and Measure.* Wiley, NY, third edition, 1995.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation.* Cambridge University Press, Cambridge, 1987.

[5] D. Böhning. Convergence of Simar's algorithm for finding the maximum likelihood estimator of a compound Poisson process. *Annals of Statistics*, 10(3):1006–1008, 1982.

[6] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.

[7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical report, Center for Research in Computing Technology, Harvard University, Cambridge, Massachusetts, August 1998. TR-10-98.

[8] H. Chernoff. A measure of the asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[9] A. Cohen and H. B. Sackrowitz. Admissibility of estimators of the probability of unobserved outcomes. *Annals of the Institute of Statistical Mathematics*, 42(4): 623–636, 1990.

[10] J. J. Deely and R. L. Kruse. Construction of sequences estimating the mixing distribution. *Annals of Mathematical Statistics*, 39(1):286–288, 1968.

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

[12] W. W. Esty. A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics*, 11(3):905–912, 1983.

[13] W. Feller. On a general class of "contagious" distributions. *Annals of Mathematical Statistics*, 14(4):389–400, 1943.

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

[15] A. Gandolfi and C. C. A. Sastri. Nonparametric estimations about species not observed in a random sample. *Milan Journal of Mathematics*, 72(1):81–105, 2004.

[16] A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

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

[18] J. Karamata. Sur un mode de croissance régulière. Théorèmes fondamenteaux. *Bulletin de la Société Mathématique de France*, 61:55–62, 1933.

[19] S. Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17(4):373–401, 1967.

[20] B. G. Leroux. Consistent estimation of a mixing distribution. *Annals of Statistics*, 20(3):1350–1360, 1992.

[21] 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.

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

[23] M. I. Ohannessian and M. A. Dahleh. Distribution-dependent performance of the Good-Turing estimator for the missing mass. In *19th International Symposium on Mathematical Theory of Networks and Systems, MTNS*, 2010.

[24] M. I. Ohannessian, V. Y. F. Tan, and M. A. Dahleh. Canonical estimation in a rare-events regime. In *49th Allerton Conference on Communication, Control, and Computing, Allerton*, pages 679–682, 2011.

[25] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[26] H. E. Robbins. Estimating the total probability of the unobserved outcomes of an experiment. *Annals of Mathematical Statistics*, 39(1):256–257, 1968.

[27] L. Simar. Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics*, 4(6):1200–1209, 1976.

[28] M. J. Steele. Le Cam's inequality and Poisson approximations. *American Mathematical Monthly*, 101(1):48–54, 1994.

[29] H. Teicher. Identifiability of mixtures. *Annals of Mathematical Statistics*, 32(1): 244–248, 1961.

[30] G. Valiant and P. Valiant. Estimating the unseen: an $n/log(n)$-sample estimator for entropy and support, shown optimal via new CLTs. In *43rd ACM Symposium on Theory of Computing, STOC*, pages 1840–1847, 2011.

[31] A. B. Wagner, P. Viswanath, and S. R. Kulkarni. Probability estimation in the rare-events regime. *IEEE Transactions on Information Theory*, 57(6):3207–3229, 2011.

[32] C. L. Wood and M. M. Altavela. Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions. *Biometrika*, 65(1):235–239, 1978.

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