JMLR: Workshop and Conference Proceedings vol 23 (2012) 21.1–21.24 25th Annual Conference on Learning Theory

Rare Probability Estimation under Regularly Varying Heavy Tails

Mesrob I. Ohannessian MESROB@MIT.EDU
Munther A. Dahleh DAHLEH@MIT.EDU
LIDS, MIT, 32 Vassar Street, Cambridge, MA 02139

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

This paper studies the problem of estimating the probability of symbols that have occurred very rarely, in samples drawn independently from an unknown, possibly infinite, discrete distribution. In particular, we study the multiplicative consistency of estimators, defined as the ratio of the estimate to the true quantity converging to one. We first show that the classical Good-Turing estimator is not universally consistent in this sense, despite enjoying favorable additive properties. 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 consistent, address some of the shortcomings of the Good-Turing estimator. For example, they perform smoothing implicitly and have the absolute discounting structure of many heuristic algorithms. This also establishes a discrete parallel to extreme value theory, and many of the techniques therein can be adapted to the framework that we set forth.

Keywords: Rare events, probability estimation, Good-Turing, consistency, concentration

1. Introduction

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 volumes, 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?

Motivated particularly by problems of computational language modeling, we are interested in the following archetypal problem. Let $X_1, \cdots, X_n$ be an observation sequence of random variables drawn independently (i.i.d.) 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 $\mathbb{P}$. In language modeling, sub-words, words, and syntactic structures are but a few of the wide array of possible characterizations of natural language. Although an i.i.d. model may seem too simple to address the complexity of this domain, it remains a core construction upon which more sophistication can be built. For example, n-gram models combine correlation among various hierarchies with a basic learning process that is based on conditional independence within each hierarchy.

We are interested in using the observations to perform probability estimation. For events occurring frequently, this is a task handled easily by the maximum likelihood estimator, i.e. the empirical probability. However, this works questionably, if at all, for infrequent events. In particular, it is desirable not to assign zero probability to symbols that are never seen in the training data, since that would sabotage any technique that uses the learned model to evaluate the likelihood of test instances which happen to contain a new symbol. Therefore, our focus is on obtaining estimators for qualitatively rare events. One concrete class of such events are the subsets $B_{n,r}$ of symbols 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 which do not appear in the observation. Define the *rare probabilities* as:

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

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 (1953), who gives due credit to Turing. Their solution to this estimation problem, the *Good-Turing estimator*, is:

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

where $K_{n,r} := |B_{n,r}|$ is the number of distinct symbols appearing exactly r times in the sample, i.e. the number of boxes containing exactly r balls. The study of $K_{n,r}$ for every r, and of the total number of distinct symbols in the sample $K_n := \sum_{r\geq 1} K_{n,r}$, is known as the *occupancy problem*. Why have we not used the obvious estimator, the empirical probability of $B_{n,r}$, which would be $\frac{rK_{n,r}}{n}$ in contrast to (1)? 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 (1) guarantees a bias of no more than $1/n$ universally, i.e. regardless of the underlying distribution. This is not true for the empirical estimator.

Many other statistical properties of this estimator, and in particular for the missing mass, have been studied beyond the bias results, such as its asymptotic normality, Esty (1983), its admissibility with respect to mean squared error only with finite support and inadmissibility otherwise, and being a minimum variance unbiased estimator of $\mathbf{E}[M_{n-1,r}]$ (both by Cohen and Sackrowitz (1990)). More recently McAllester and Schapire (2000) also showed that $G_{n,0}$ concentrates above $M_{n,0}$. More properties and alternative algorithms may be found in the survey of Gandolfi and Sastri (2004). Most of these results put no assumption on the underlying distribution, therefore giving great generality. Without underlying assumptions, however, worst case distributions can severely hamper what one can say about our ability to estimate rare probabilities. In particular, most convergence results, including those just listed, are often in *additive* form, that is one is concerned with characterizing the behavior of the difference between the estimated and true probability values. For a fixed discrete distribution, rare probabilities all decay to zero, and one would expect a more meaningful characterization to be in *multiplicative* form. What we mean by this is that the natural mode of convergence is for ratios to converge to one, probabilistically. So, in this paper, we call an estimator $\hat{M}_{n,r}$ of $M_{n,r}$ *consistent*, if $\hat{M}_{n,r}/M_{n,r} \to 1$, where the convergence can be either in probability or almost surely, over the observation sequence.

We first show that, without restricting the class of distributions, the Good-Turing estimator is not consistent. We then use Karamata's theory of regular variation to give a general description of heavy-tailed discrete distributions, and show that this is a sufficient condition for the consistency

of the Good-Turing estimator. We do so by extending the McAllester and Ortiz (2003) additive concentration results for the missing mass to all of the rare probabilities, and then using the regular variation property to show multiplicative concentration. Additionally, 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 (Beirlant et al. (2004)), which has been successfully employed to estimate rare probabilities in continuous settings. With the insight provided by this paper, many of the techniques therein can be adapted to our discrete setting.

Rare probability estimation, especially in the context of language modeling, has been extensively investigated. Simple approaches, such as Laplace or add-one estimators and their variants have been compared to Good-Turing estimators, as in Gale and Sampson (1995), highlighting the latter's superiority albeit at the expense of volatility and necessity to smooth out. Some of the most successful approaches to n-gram modeling and learning have been the algorithm proposed by Kneser and Ney (1995) and its variants, as studied in the influential survey of Chen and Goodman (1998). These are inspired by Good-Turing, but incorporate it in a general hierarchical smoothing scheme, using back-off or interpolation between n-gram levels. We do not address this hierarchical smoothing here, but shed light on a smoothing technique used within each level, called "absolute discounting". This appears naturally in the estimators that we propose, where we furthermore identify the "discount" precisely as the regular variation index. This correspondence between Kneser-Ney algorithms and heavy tails has also been pointed out on the Bayesian front by Teh (2006). Extending early work by MacKay and Peto (1995), who only focused on light-tailed priors, Teh uses the two-parameter Poisson-Dirichlet process, proposed by Pitman and Yor (1997), as prior. Instances of this random measure have indeed regularly varying heavy tails, almost surely (Gnedin et al. (2007)). In this paper, however, we do not model or parameterize the entire distribution. We focus rather on describing the tail alone, which is sufficient for both characterizing the behavior of the rare probabilities and consistently estimating them. Lastly, in light of the duality between compression and probability estimation, the problem of rare probability estimation also appears in the information theory community, especially when handling very large alphabets. In particular, in their pioneering work, Orlitsky et al. (2004) show that universal compression of sequences is possible over arbitrarily large alphabets, provided one drops the identity of symbols, while preserving their order. In this context, the Good-Turing estimator is shown to behave better than simpler approaches, but not as well as what one could optimally achieve.

The rest of this paper is organized as follows. In Section 2, we give the basic definition of consistency and show the failure of Good-Turing to be consistent for geometric distributions. In Section 3, we define regularly varying distributions with heavy tails. The rest of the needed background material is given in Appendix A as an exposition to the exponential moment method and the property of negative association. Using these tools, in Section 4.1 we extend the additive concentration results of McAllester, Schapire and Ortiz – McAllester and Schapire (2000); McAllester and Ortiz (2003) – to all rare probabilities.

Employing the same methodology, but adding regular variation, we derive multiplicative concentration and strong laws under heavy tails in Section 4.2. Lastly, in Section 5, we use the strong laws to show the consistency of Good-Turing estimation under heavy-tailed regular variation, and to construct a family of new consistent estimators. All detailed proofs can be found in Appendix B, organized by section. We end with a summary in Section 6.

NOTATION

Throughout the paper, 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 with random quantities.

2. The Good-Turing Estimator is not Universally Consistent

Let us first formalize the multiplicative notion of consistency, which we use throughout the paper, and which naturally conforms to the asymptotics of rare probabilities.

Definition 1 (Consistency) *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. We also write the latter as $\hat{M}_{n,r}/M_{n,r} \to_{\text{a.s.}} 1$ or $\hat{M}_{n,r} \sim_{\text{a.s.}} M_{n,r}$.*

We first show that consistency is not trivial, as even in the case of fairly well-behaved distributions, the Good-Turing estimator does not result in a consistent estimator for the missing mass.

Proposition 2 *For a geometric distribution $p_j = (1-q)q^j$ for $j \in \mathbb{N}_0$, with small enough $q \in (0,1)$, there exists a positive $\eta > 0$, and a subsequence n_i such that for i large enough we have that $G_{n_i,0}/M_{n_i,0} = 0$ with probability no less than η. In particular, it follows that the Good-Turing estimator of the missing mass is not consistent.*

This motivates us to ask what are sufficient conditions to obtain consistent estimation. In particular, the problem seems to be that with a light-tailed distribution like the geometric, there are not enough samples to learn the rare probabilities well enough for consistency. With this insight, we move next to show that regularly varying heavy-tailed distributions are a natural situation in which one has enough samples and, most importantly, consistency.

3. Regularly Varying Heavy-Tailed Distributions

We now characterize heavy-tailed discrete distributions, by using Karamata's theory of regular variation, which was developed originally in Karamata (1933), with the standard reference being Bingham et al. (1987). The application we have here is based on the early work of Karlin (1967), which was recently given an excellent exposition by Gnedin et al. (2007). We follow the notational convention of the latter.

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

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

where δ_x is a Dirac mass at x.

Using ν, we define the following function, which was used originally by Karlin (1967) 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:

$$\boldsymbol{\nu}(x) := \nu[x,1].$$

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

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

Definition 3 (Regular Variation) *Following Karlin (1967), we say that* $\mathbb{P}$ *is* regularly varying *with* regular variation index $\alpha \in (0, 1)$*, if the following holds:*

$$\boldsymbol{\nu}(x) \sim x^{-\alpha} \ell(1/x), \quad \text{as} \quad x \downarrow 0, \tag{2}$$

where $\ell(t)$ *is a slowly varying function, i.e. for all* $c > 0$*,* $\ell(ct)/\ell(t) \to 1$ *as* $t \to \infty$*.*

The following portmanteau theorem (which compiles results found in Gnedin et al. (2007)) is a very useful collection of conditions that are equivalent to regular variation as given by Definition 3, in addition to facts that follow from regular variation.

Theorem 4 *Equation (2) is equivalent to any (and therefore all) of the following:*

- *Probability accrual:*

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

- *Expected number of distinct observed symbols:*

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

- *Expected number of symbols observed exactly once:*

$$\mathbf{E}[K_{n,1}] \sim \alpha\Gamma(1-\alpha) n^\alpha \ell(n), \quad \text{as} \quad n \to \infty, \tag{3}$$

- *Number of distinct observed symbols:*

$$K_n \sim_{\text{a.s.}} \Gamma(1-\alpha) n^\alpha \ell(n), \quad \text{as} \quad n \to \infty, \tag{4}$$

- *Number of symbols observed exactly once:*

$$K_{n,1} \sim_{\text{a.s.}} \alpha\Gamma(1-\alpha) n^\alpha \ell(n), \quad \text{as} \quad n \to \infty, \tag{5}$$

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

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

$$\mathbf{E}[K_{n,r}] \sim \frac{\alpha\Gamma(r-\alpha)}{r!} n^\alpha \ell(n), \quad \text{as} \quad n \to \infty, \tag{6}$$

$$K_{n,r} \sim_{\text{a.s.}} \frac{\alpha\Gamma(r-\alpha)}{r!} n^\alpha \ell(n), \quad \text{as} \quad n \to \infty. \tag{7}$$

Theorem 4 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), (5), (6), and (7), 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 the paper.

Lastly we note that, with additional care, it is possible to also incorporate the cases $\alpha = 0$ (slow variation, light tail) and $\alpha = 1$ (rapid variation, very heavy tail), as is done for example in Gnedin et al. (2007). The example of the geometric distribution illustrates why care is needed. Indeed, the geometric satisfies regular variation with $\alpha = 0$, yet the decay of rare probabilities, in expectation, can have oscillatory behavior, as hinted at by the proof of Proposition 2.

4. Concentration Results for Rare Probabilities

4.1. Additive Concentration for General Distributions

We now follow the methodology of McAllester and Ortiz (2003) in order to extend their additive concentration results for the missing mass $M_{n,0}$, to all of $M_{n,r}$. 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 5 *Consider an arbitrary $\mathbb{P}$. Then, for every $r = 0, 1, 2, \cdots$, there exist absolute constants $a_r, b_r, \epsilon_r > 0$ such that for every $n > 2r$, and for all $0 < \epsilon < \epsilon_r$, we have*

$$\mathbb{P}\left\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\right\} \leq a_r e^{-b_r \epsilon^2 n}. \tag{8}$$

Proof sketch For the proof of this theorem, we cannot directly parallel the proofs of McAllester and Ortiz (2003) for the additive concentration of the missing mass, because unlike $M_{n,0}$, $M_{n,r}$ cannot be expressed as a sum of negatively associated random variables. Instead, we work with a quantity that can be expressed as such, namely the total probability of all symbols appearing no more than r times:

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

Indeed, we can express it as follows $M_{n,0\to r} = \sum_j p_j Z_{n,j,r}$, where $Z_{n,j,r} = \mathbf{1}\{C_{n,j} \leq r\}$, where $C_{n,j} := \sum_{i=1}^{n} \mathbf{1}\{X_i = j\}$ designates the count of symbol j. Thus $Z_{n,j,r}$'s are indicator random variables associated with each symbol j, 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 17, then by Lemma 16 so are $\{Z_{n,j,r}\}_{j\in\mathbb{N}}$.

To establish additive concentration for $M_{n,0\to r}$, we use the exponential moment method and the Gibbs variance lemma and negative association, as outlined in Appendix A, in a close parallel to McAllester and Ortiz (2003). The main technical challenge is to properly bound the Gibbs variance for the upper deviation. To complete the proof, we show that it's sufficient to establish additive concentration for all $M_{n,0\to r}$, in order to have it for $M_{n,r}$. ■

4.2. Multiplicative Concentration and Strong Laws under Heavy Tails

Our objective 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 (1967) (Theorem 9) and (with an explicit proof) by Gnedin et al. (2007) (Proposition 2.5). Beyond being a generalization for all r, 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 6 *Assume $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3. 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:*

$$\mathbb{P}\left\{\left|\frac{M_{n,r}}{\mathbf{E}[M_{n,r}]} - 1\right| > \delta\right\} \leq a_r e^{-b_r \delta^2 n^\alpha \ell(n)}. \tag{9}$$

Proof sketch We cannot deduce the multiplicative concentration of Theorem 6 directly from the additive concentration of Theorem 5, 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 detailed proof. ■

The strong laws for the rare probabilities are easily established using the multiplicative concentration of Theorem 6.

Proposition 7 *If $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3, then for every $r = 0, 1, 2, \cdots$, we have*

$$\frac{M_{n,r}}{\mathbf{E}[M_{n,r}]} \to_{\text{a.s.}} 1, \tag{10}$$

and the asymptotic expression:

$$M_{n,r} \sim_{\text{a.s.}} \mathbf{E}[M_{n,r}] \sim \frac{\alpha\Gamma(r + 1 - \alpha)}{r!} n^{\alpha-1}\ell(n). \tag{11}$$

5. Consistent Probability Estimation

5.1. Consistency of Good-Turing Estimation

As a first application of the strong laws of Proposition 7, in conjunction with the strong laws for $K_{n,r}$, we prove the strong consistency of the Good-Turing estimator in this regime.

Proposition 8 *If $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3, then the Good-Turing estimators are strongly consistent for all $r = 0, 1, \cdots$:*

$$\frac{G_{n,r}}{M_{n,r}} \to_{\text{a.s.}} 1. \tag{12}$$

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, Beirlant et al. (2004), 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 (4) and (5), 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:

$$\hat{\alpha} := \frac{K_{n,1}}{K_n} \to_{\text{a.s.}} \alpha. \tag{13}$$

Note that this is by no means the only approach to estimating the index. For example, other asymptotic expressions that appear in Theorem 4 may be harnessed. Moreover, one may devise methods that are inspired from techniques in extreme value theory Beirlant et al. (2004), 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 (13) in each block, call it $\hat{\alpha}_m$, $m = 1, \cdots, M$, then average out:

$$\hat{\alpha} = \frac{1}{M} \sum_m \hat{\alpha}_m. \tag{14}$$

With proper choices of M and N, this empirically shows much less volatility than a straight application of (13) (i.e. $M = 1$, $N = n$). In Figure 1, we qualitatively illustrate the improved volatility of a Gumbelian estimator with an example. 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 (14).

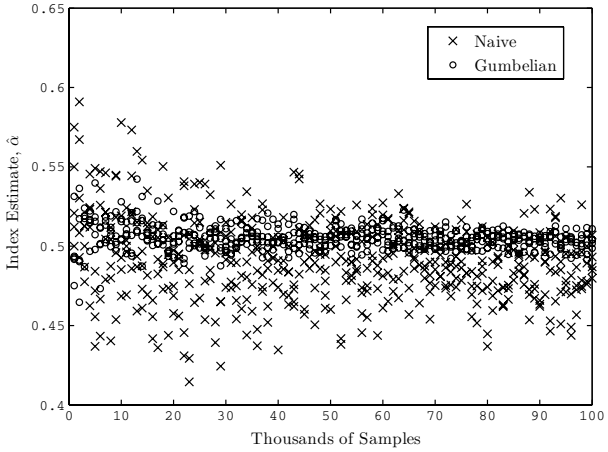


Figure 1: Qualitative comparison of naive and Gumbelian index estimators.

TWO PROBABILITY ESTIMATOR CONSTRUCTIONS

We have shown that the Good-Turing estimator is consistent in the regularly varying heavy-tailed setting. But could one do better by working within this framework explicitly? We now 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. As suggested in the preamble of this section, our constructions are in two stages. We first assume that we have chosen a consistent estimator $\hat{\alpha} \to_{\text{a.s.}} \alpha$. This can be given by (13), or potentially more powerful estimators. We then use the estimated index to construct consistent estimators for rare probabilities.

Proposition 9 *Consider the following family of estimators, for $r = 1, 2, \cdots$:*

$$\hat{M}^{(1)}_{n,r} := (r - \hat{\alpha})\frac{K_{n,r}}{n}, \tag{15a}$$

and for $r = 0$:

$$\hat{M}^{(1)}_{n,0} := 1 - \sum_{r \geq 1} \hat{M}^{(1)}_{n,r} = \hat{\alpha}\frac{K_n}{n}. \tag{15b}$$

If $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3. Then $\hat{M}^{(1)}_{n,r}$ are strongly consistent for all $r = 0, 1, \cdots$, that is $\hat{M}^{(1)}_{n,r}/M_{n,r} \to_{\text{a.s.}} 1$.

One motivation for introducing the $\hat{M}^{(1)}_{n,r}$ family is because it has the "absolute discounting" form that is a component of many language learning heuristics, and especially the Kneser-Ney line of state-of-the-art algorithms, proposed originally by Kneser and Ney (1995) and extended by Chen and Goodman (1998) and others. What we mean by absolute discounting is that, effectively, the multiplicity of symbols in the empirical distribution is corrected by a constant: $\frac{rK_{n,r}}{n}$ is replaced by $\frac{(r-\hat{\alpha})K_{n,r}}{n}$, as though a symbol that appeared r times did in fact only appear $r - \hat{\alpha}$ times. Most interestingly, we have systematically established the nature of the discount as the regular variation index.

It is worth mentioning that this structure 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 of interest. Good-Turing is coarse in this sense, and Good (1953) originally suggests various ways to "smooth" this behavior out. Here, on the contrary, the estimator evaluates to 0 if and only if there are no symbols in the occupancy level itself. We can thus think of $\hat{M}^{(1)}_{n,r}$ as having smoothing built-in.

Instead of using individual occupancy numbers $K_{n,r}$, we can perform additional smoothing by using K_n, the number of distinct observed symbols. Since K_n has less variability, e.g. it is non-decreasing with sample size, the resulting estimator inherits that robustness.

Proposition 10 *Consider the following family of estimators, for $r = 0, 1, \cdots$:*

$$\hat{M}^{(2)}_{n,r} := \frac{\hat{\alpha}\Gamma(r + 1 - \hat{\alpha})}{r!\Gamma(1 - \hat{\alpha})}\frac{K_n}{n} \equiv \binom{r - \hat{\alpha}}{r}\hat{\alpha}\frac{K_n}{n}. \tag{16}$$

If $\mathbb{P}$ is regularly varying with index $\alpha \in (0, 1)$, as in Definition 3. Then $\hat{M}^{(2)}_{n,r}$ are strongly consistent for all $r = 0, 1, \cdots$, that is $\hat{M}^{(2)}_{n,r}/M_{n,r} \to_{\text{a.s.}} 1$.

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 (13), we also have:

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

GOOD-TURING AS A TWO-STAGE ESTIMATOR

If we use the naive index estimator $\hat{\alpha}$ as in (13), then we have:

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

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 (13), and then use it to obtain a probability estimator, as in (15) or (16). 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 (14), 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 2(*a*), 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.

In Figure 2(*b*) 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.

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:

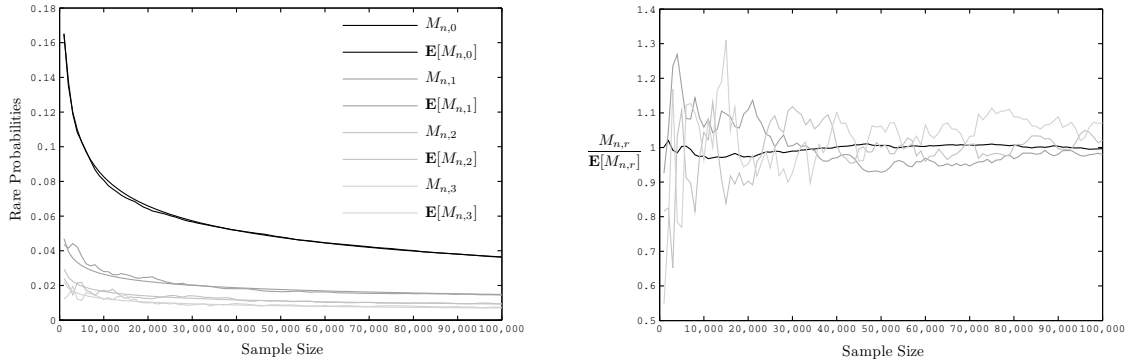
- We use the naive estimator for α, as given by Equation (13), in order to show that the benefit of the new estimators comes from their structure too, and not only because from better index estimation.

- We purposefully use fewer samples than in Figure 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 distinctive case, $r = 2$. The typical behavior for larger r is comparable. We show the raw behavior of the estimators in Figure 3(*a*). To make the comparison crisper, we also show the behavior of the ratios of each estimator to $M_{n,r}$ in Figure 3(*b*). For reference, this figure also shows the 1-line and the ratio of the mean itself, i.e. $\mathbf{E}[M_{n,r}]/M_{n,r}$.

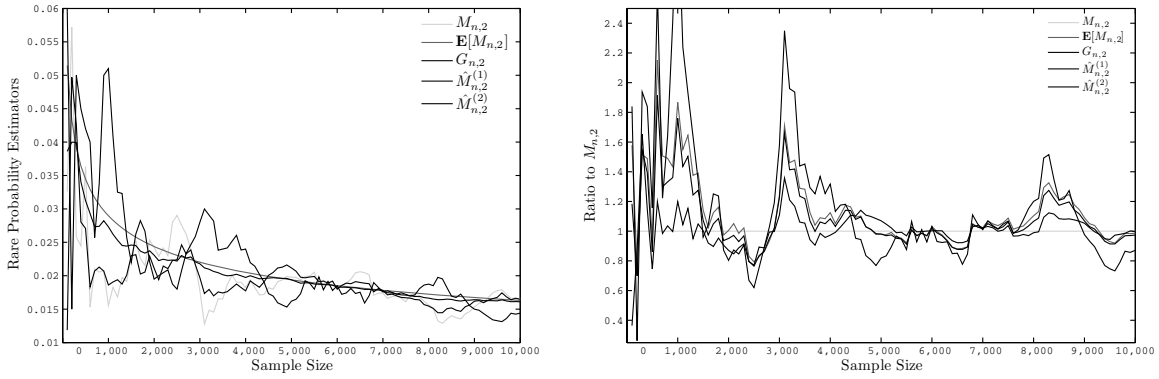
We end with a few qualitative comments. First, it is apparent that $\hat{M}^{(1)}_{n,r}$ outperforms all estimators when it comes to tracking $M_{n,r}$ closely. It is followed by $\hat{M}^{(2)}_{n,r}$ in performance, while $\hat{G}^{(1)}_{n,r}$

is consistently more volatile than both of the new estimators. Also note that $\hat{M}_{n,r}^{(2)}$ is the smoothest estimator. However, it tracks the expectation $\mathbf{E}[M_{n,r}]$ rather 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 $\mathbf{E}[M_{n,r}]$.



(*a*) (*b*)

Figure 2: (*a*) Decay behavior of $M_{n,r}$ and $\mathbf{E}[M_{n,r}]$. (*b*) Multiplicative concentration of $M_{n,r}$ around $\mathbf{E}[M_{n,r}]$ and strong law behavior, where $M_{n,r}/\mathbf{E}[M_{n,r}]$ approaches 1.



(*a*) (*b*)

Figure 3: (*a*) Behavior of $G_{n,2}$, $\hat{M}_{n,2}^{(1)}$, and $\hat{M}_{n,2}^{(2)}$, alongside $M_{n,2}$ and $\mathbf{E}[M_{n,2}]$. (*b*) Ratios $G_{n,2}/M_{n,2}$, $\hat{M}_{n,2}^{(1)}/M_{n,2}$, and $\hat{M}_{n,2}^{(2)}/M_{n,2}$ alongside $\mathbf{E}[M_{n,2}]/M_{n,2}$.

6. Summary

In this paper, we studied the problem of rare probability estimation, from the perspective of the (multiplicative) consistency of the estimator: requiring that the ratio of the estimate to the true

quantity converges 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. To characterize these, we used Karamata's theory of regular variation Karamata (1933), following closely the early development of Karlin (1967) in the context of infinite urn schemes. We then used the McAllester and Ortiz (2003) method 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 used the newly established strong laws, in addition to 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, such as in the algorithm proposed by Kneser and Ney (1995) and extended by Chen and Goodman (1998) and others. 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 (Beirlant et al. (2004)). This correspondence opens the door for modern semiparametric methods to be applied in the present framework.

Heavy tails are a very good model for natural language, as observed early on by Zipf (1949). As such, it is satisfying that we have shown here that this is a property that is *sufficient* for consistently estimating rare probabilities. The core multiplicative concentrations have room to generalize to heavy tails that are potentially not regularly varying, as long as the mean and variance growths balance out to yield a proper unbounded exponent. Naturally, to completely describe when rare probability estimation is possible in a meaningful manner, one ought to establish *necessary* conditions as well.

Acknowledgments

This work was supported in part by NSF grant 6922470 and ONR grant 6918937.

References

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

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

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.

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.

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.

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

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.

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

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.

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.

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

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

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

R. Kneser and H. Ney. Improved smoothing for m-gram language modeling. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 679–682, 1995.

D. MacKay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):289–307, 1995.

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.

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

A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. on Information Theory*, 50(7):1469–1481, 2004.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 985–992, 2006.

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

Appendix A. Preliminaries

A.1. Exponential Moment Method and the Gibbs Variance Lemma

We adhere closely to the exposition of McAllester and Ortiz (2003). The exponential moment method takes its name from Markov's inequality applied to an exponentiated random variable. It is embodied in the following statement, which traces back to Chernoff (1952).

Theorem 11 *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:*

$$S(W, w) = \sup_{\beta \in R} w\beta - \log Z(W, \beta), \quad \text{with} \quad Z(W, \beta) = \mathbf{E}[e^{\beta W}].$$

Then, the lower and upper deviation probabilities can be bounded by:

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

We now give some background on Gibbs measures, which have distinct roots in statistical mechanics, but are also an integral part of the exponential moment method. For a given W, let (β_-, β_+) be the largest open interval over which $Z(W, \beta)$ is finite. In this paper this interval will always be the entire real line. Denote the law of W by μ, then with each $\beta \in (\beta_-, \beta_+)$, we can associate a new probability measure, the *Gibbs measure*:

$$\mu_\beta(\mathrm{d}w) = \frac{e^{\beta w}}{Z(W, \beta)} \mu(\mathrm{d}w).$$

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 (2003) distill a particular result out of Chernoff's original work, and dub it the Gibbs variance lemma.

Lemma 12 (Gibbs Variance) *Let W be an arbitrary finite mean real-valued random variable. Then for any $w \in (w_-, w_+)$ and $\beta \in (\beta_-, \beta_+)$, we have:*

$$\begin{aligned} S(W, w) &= w\beta(w) - \log Z(W, \beta(w)) \\ &= D\left(\mu_{\beta(w)} \| \mu\right) \\ &= \int_{\mathbf{E}[W]}^{w} \int_{\mathbf{E}[W]}^{v} \frac{1}{\sigma^2(W, \beta(u))} \mathrm{d}u\mathrm{d}v, \end{aligned} \tag{17}$$

$$\log(Z(W, \beta)) = \mathbf{E}[W]\beta + \int_0^\beta \int_0^\alpha \sigma^2(W, \gamma)\mathrm{d}\gamma\mathrm{d}\alpha. \tag{18}$$

The importance of this lemma is that it showcases how we can establish concentration by controlling the variance $\sigma^2(W, \beta)$ in (17) and (18). The following two lemmas (reproducing Lemmas 9 and 11 from McAllester and Ortiz (2003), with the exception of part *(ii)* below) are established by doing precisely that.

Lemma 13 *Let W be an arbitrary finite mean real-valued random variable.*

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

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

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

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

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

Lemma 14 *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:*

$$S(W, \mathbf{E}[W] - \epsilon) \ge \frac{\epsilon^2}{2\sum_j b_j^2 q_j}. \tag{19}$$

A.2. 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 Dubhasi and Ranjan (1998), and are also outlined in McAllester and Ortiz (2003). 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 15 (Negative Association) *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 coordinate-wise non-increasing or coordinate-wise non-decreasing, we have:*

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

This lemma constructs new negatively associated random variables from existing ones.

Lemma 16 *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 17 *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, for the purpose of the exponential moment method, sums of negatively associated random variables can be treated like a sum of independent random variables with the same marginals. More precisely, the exponent of the negatively associated sum dominates that of the independent sum, and thus bounding the latter from below, bounds the former also.

Lemma 18 *Say $W = \sum_{j=1}^{k} W_j$, where $W_1, \cdots, W_k$ are negatively associated. Let $\tilde{W} = \sum_{j=1}^{k} \tilde{W}_j$, where $\tilde{W}_1, \cdots, \tilde{W}_k$ are* independent *real-valued random variables such that for each j the law of $\tilde{W}_j$ is the same as the (marginal) law of W_j. Then for all w, we have:*

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

It is worth noting that this approach is not unlike the Poissonization technique used by Karlin (1967), Gnedin et al. (2007), 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 18 does precisely that.

Appendix B. Proofs

B.1. Proofs of Section 2

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 first show the following precise behavior for the counts of symbols seen exactly once.

Lemma 19 *For the subsequence $n_i = \lfloor c/p_i \rfloor = \lfloor cq^{-i}/(1-q) \rfloor$, with $c > 0$, we have:*

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

where

$$h(c,q) := \sum_{m=-\infty}^{\infty} cq^m e^{-cq^m}. \tag{20}$$

Proof In general, by Poissonization (e.g. Gnedin et al. (2007), Lemma 1) or using the dominated convergence theorem with $\left(1 - \frac{s}{n}\right)^n \uparrow e^{-s}$, one has that as $n \to \infty$:

$$\left| \sum_{j=0}^{\infty} np_j(1-p_j)^n - \sum_{j=0}^{\infty} np_j e^{-np_j} \right| \to 0. \tag{21}$$

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, $n_i = \lfloor c/p_i \rfloor$. For the moment, assume that there exists a rather convenient sequence $j_i \to \infty$ that grows slow enough such that $i - j_i \to \infty$ and

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

This gives us enough approximation leeway to show first that we can replace n_i by c/p_i in Equation (21), without altering the limit:

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

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:

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

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. ■

Proof [Proposition 2]

For any real-valued non-negative random variable W, Markov's inequality yields:

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

Recall h from Equation (20). Assume for the moment 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 $\eta \in (0, 1 - h(c,q))$. Consider the subsequence n_i obtained in Lemma 19 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 $i > i_0$:

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

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 no

less than $\eta >$ we have $G_{n_i,0}/M_{n_i,0} = 0$. Therefore, for all geometric distributions with $q < q_0$, $G_{n,0}/M_{n,0} \nrightarrow 1$ in probability, let alone almost surely.

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

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

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. ■

B.2. Proofs of Section 4

The following lemma allows us to move from $M_{n,0\to r}$ to $M_{n,r}$.

Lemma 20 *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*

$$\mathbb{P}\left\{|M_{n,0\to r} - \mathbf{E}[M_{n,0\to r}]| > \epsilon\right\} \le \tilde{a}_r e^{-\tilde{b}_r \epsilon^2 n}. \tag{22}$$

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 (8) holds.

Proof Define the events: $A = \{\mathbf{E}[M_{n,0\to r}] - \epsilon/2 \le M_{n,0\to r} \le \mathbf{E}[M_{n,0\to r}] + \epsilon/2\}$, and $B = \{\mathbf{E}[M_{n,0\to r-1}] - \epsilon/2 \le M_{n,0\to r-1} \le \mathbf{E}[M_{n,0\to r-1}] + \epsilon/2\}$. Then $A \cap B \subset \{\mathbf{E}[M_{n,r}] - \epsilon \le M_{n,r} \le \mathbf{E}[M_{n,r}] + \epsilon\}$, 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 $n > 2r$, $0 < \epsilon < \tilde{\epsilon}_r \wedge \tilde{\epsilon}_{r-1}$:

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

This establishes Equation 8, 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 [Theorem 5]

By Lemma 20, 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 18 that in order to establish concentration for $M_{n,0\to r}$ it suffices to show concentration for the quantity:

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

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 13, with $W = \tilde{M}_{n,0\to r}$. To obtain the lower exponent, it is easiest to use Lemma 14, with $W_j = \tilde{Z}_{n,j,r}$ and $b_j = q_j$. We have:

$$\begin{aligned}
\sum_j p_j^2 q_j &= \sum_{k=0}^{r} \sum_j p_j \binom{n}{k} p_j^{k+1} (1-p_j)^{n-k} \\
&= \sum_{k=0}^{r} \frac{\binom{n}{k}}{\binom{n+1}{k+1}} \underbrace{\sum_j p_j \binom{n+1}{k+1} p_j^{k+1} (1-p_j)^{(n+1)-(k+1)}}_{\mathbf{E}[M_{n+1,k+1}]} \\
&\leq \sum_{k=0}^{r} \frac{r+1}{n+1} \mathbf{E}[M_{n+1,k+1}] = \frac{r+1}{n+1} \mathbf{E}[M_{n+1,1\to r+1}] \leq \frac{r+1}{n+1}. \qquad (23)
\end{aligned}$$

Adapting this bound to Equation (19), 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 13. 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 on $\beta \leq n - r$, and write:

$$\begin{aligned}
\sigma^2(W, \beta) &\leq \sum_j p_j^2 q_j (1-p_j)^{-\beta} \\
&= \sum_{k=0}^{r} \sum_j p_j \binom{n}{k} p_j^{k+1} (1-p_j)^{n-\beta-k} \\
&= \sum_{k=0}^{r} \frac{\binom{n}{k}}{\binom{n-\beta+1}{k+1}} \underbrace{\sum_j p_j \binom{n-\beta+1}{k+1} p_j^{k+1} (1-p_j)^{(n-\beta+1)-(k+1)}}_{:= \zeta_{n,k+1}(\beta)} \\
&= \sum_{k=0}^{r} \frac{k+1}{n-\beta+1} \frac{\binom{n}{k}}{\binom{n-\beta}{k}} \zeta_{n,k+1}(\beta). \qquad (24)
\end{aligned}$$

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 $\beta \leq n - r$, 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 $0 \leq \zeta_{n,k}(\beta) \leq 1$ for every k, and $\sum_k \zeta_{n,k}(\beta) \leq 1$. Therefore, if we bound every summand, we can use the largest bound to bound the entire sum. We have:

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

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

$$\sigma^2(W,\beta) \leq \frac{r+1}{n-\beta+1} e^r \left(1-\beta/n\right)^{-r} = \frac{(r+1)e^r}{n} \left(1-\beta/n\right)^{-(r+1)}. \tag{25}$$

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 $x \leq 1/(m+1)$ implies $(1-x)^{-m} \leq e$, it follows from Equation (25) that for all $0 < \beta < \overline{\beta}$ we indeed have $\sigma^2(W,\beta) \leq \overline{\sigma}^2$. Therefore, part *(i)* of Lemma 13 applies, and we deduce that for all $0 < \epsilon \leq 1 < \frac{r+1}{r+2} e^{r+1} \equiv \overline{\beta}\overline{\sigma}^2$ we have:

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

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 (22) holds, and by Lemma 20, so does the additive concentration for $M_{n,r}$ as given by Equation (8). It is worth noting that, as remarked by McAllester and Ortiz (2003) for the missing mass, Bernstein inequalities can recover the (simpler) lower deviation result of rare probabilities, but the upper deviation appears to require the additional malleability of the Gibbs variance lemma. ■

Proof [Theorem 6]

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 18 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 13, with $W = \tilde{M}_{n,0\to r}$. For the lower exponent, we use the specialized Lemma 14 instead, with $W_j = \tilde{Z}_{n,j,r}$ and $b_j = q_j$. Replicating (23), we have:

$$\sum_j p_j^2 q_j \leq \frac{r+1}{n+1} \sum_{k=0}^{r} \mathbf{E}[M_{n+1,k+1}].$$

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:

$$\begin{aligned}\frac{r+1}{n+1}\sum_{k=0}^{r}\mathbf{E}[M_{n+1,k+1}] &= \frac{r+1}{n+1}\sum_{k=0}^{r}\frac{k+2}{n+2}\mathbf{E}[K_{n+2,k+2}] \\ &\leq (1+\eta)\frac{r+1}{n+1}\sum_{k=0}^{r}\frac{k+2}{n+2}\frac{\alpha\Gamma(k+2-\alpha)}{(k+2)!}n^{\alpha}\ell(n) \\ &= (1+\eta)(r+1)\sum_{k=0}^{r}\frac{\alpha\Gamma(k+2-\alpha)}{(k+1)!}n^{-(2-\alpha)}\ell(n).\end{aligned}$$

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 (19), 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 13. 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 replicate (24) and (25) to bound the Gibbs variance:

$$\sigma^2(W,\beta) \leq \sum_{k=0}^{r}\frac{k+1}{n-\beta+1}\frac{\binom{n}{k}}{\binom{n-\beta}{k}}\zeta_{n,k+1}(\beta) \leq \frac{(r+1)e^r}{n}\left(1-\beta/n\right)^{-(r+1)}\sum_{s=1}^{r+1}\zeta_{n,s}(\beta).$$

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:

$$\sigma^2(W,\beta) \leq \frac{(r+1)e^{r+1}}{n}\sum_{s=0}^{r+1}\zeta_{n,s}(\beta). \tag{26}$$

Unlike Theorem 5, we preserve here 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:

$$\begin{aligned}\sum_{s=0}^{r+1}\zeta_{n,s}(\beta) &\leq \sum_{s=0}^{r+1}\zeta_{n,s}(n/2) \equiv \sum_{s=0}^{r+1}\mathbf{E}[M_{n/2+1,s}] = \sum_{s=0}^{r+1}\frac{s+1}{n/2+2}\mathbf{E}[K_{n/2+2,s+1}] \\ &\leq (1+\eta)\sum_{s=0}^{r+1}\frac{s+1}{n/2+2}\frac{\alpha\Gamma(s+1-\alpha)}{(s+1)!}(n/2+2)^{\alpha}\ell(n) \\ &= (1+\eta)\sum_{s=0}^{r+1}2^{1-\alpha}\frac{\alpha\Gamma(s+1-\alpha)}{s!}n^{-(1-\alpha)}\ell(n). \end{aligned} \tag{27}$$

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 (26) and (27), 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 13 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:

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

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

By using the same development as the proof of Lemma 20, 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:

$$\mathbb{P}\left\{|M_{n,r} - \mathbf{E}[M_{n,r}]| > \epsilon\right\} \leq a_r e^{-\check{b}_r\epsilon^2 n^{2-\alpha}\frac{1}{(1+\eta)\ell(n)}}. \tag{28}$$

Now, observe that:

$$\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}] \sim \frac{\alpha\Gamma(r+1-\alpha)}{r!}n^{\alpha-1}\ell(n). \tag{29}$$

Let's define $m_r := \frac{\alpha\Gamma(r+1-\alpha)}{r!}$ for convenience. It follows from Equation (29) that there exists $n_r(\eta) > n_{r,2}(\eta)$ 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 $b_r = \check{b}_r m_r^2/(1+\eta)^3$, and $\delta_r = d_r/m_r$. Then for every $n > n_r(\eta)$ and for all $\delta < \delta_r$:

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

Therefore Equation (28) applies, and we get:

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

We end by noting that for fixed $\eta > 0$ and r, 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 size n_r depends on the particular convergence rates in the regular variation characterization, and to describe it explicitly requires more distribution specific knowledge than simply having α. ■

B.3. Proofs of Section 5

Proof [Proposition 7]

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 (9) is summable. Therefore, we can apply the Borel-Cantelli lemma in the usual way, to obtain the almost sure convergence of Equation (10). As for equation (11), it follows from this strong law and from Equation (6), using the fact that $\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}]$, as in Equation (29) above. ■

Proof [Proposition 8]

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 (6) and (7), we have that

$$\frac{G_{n,r}}{\mathbf{E}[G_{n,r}]} \to_{\text{a.s.}} 1. \tag{30}$$

On the other hand, note that by equation (6), 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

$$\frac{\mathbf{E}[G_{n,r}]}{\mathbf{E}[M_{n,r}]} \to 1. \tag{31}$$

Combining the convergences (10), (30), and (31), we obtain (12). ■

Proof [Proposition 9] Since $\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}]$, it follows from Equation (6) and the strong law for $M_{n,r}$ given by (10) 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 $\hat{\alpha} \rightarrow_{\text{a.s.}} \alpha$, 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 (7) 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 $r = 0$:

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

where we use Equation (4) and $\hat{\alpha} \rightarrow_{\text{a.s.}} \alpha$. ■

Proof [Proposition 10] 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 Proposition 9:

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

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:

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

using once more Equation (4). ■