Supplemental Material: Doubly-Competitive Distribution Estimation

Yi Hao
yih179@eng.ucsd.edu

Alon Orlitsky
alon@ucsd.edu

Abstract

Distribution estimation is a statistical-learning cornerstone. Its classical *min-max* formulation minimizes the estimation error for the worst distribution, hence underperforms for practical distributions that, like power-law, are often rather simple. Modern research has therefore focused on two frameworks: *structural* estimation that improves learning accuracy by assuming a simple structure of the underlying distribution; and *competitive*, or *instance-optimal*, estimation that achieves the performance of a genie-aided estimator up to a small excess error that vanishes as the sample size grows, regardless of the distribution. This paper combines and strengthens the two frameworks. It designs a single estimator whose excess error vanishes both at a universal rate as the sample size grows, as well as when the (unknown) distribution gets simpler. We show that the resulting algorithm significantly improves the performance guarantees for numerous competitive- and structural-estimation results. The algorithm runs in near-linear time and is robust to model mis-specification and domain-symbol permutations.

1 Outline

We organize the supplemental material as follows. In Section 2, we present experimental results demonstrating the competitiveness of our estimator. In Section 3, we prove Theorem 1 in the main paper and establish the estimator's optimality. In Section 4 and 5, we provide the proofs of Corollary 9 (log-concave distributions) and Corollary 11 (enveloped power-law distributions) in the main paper.

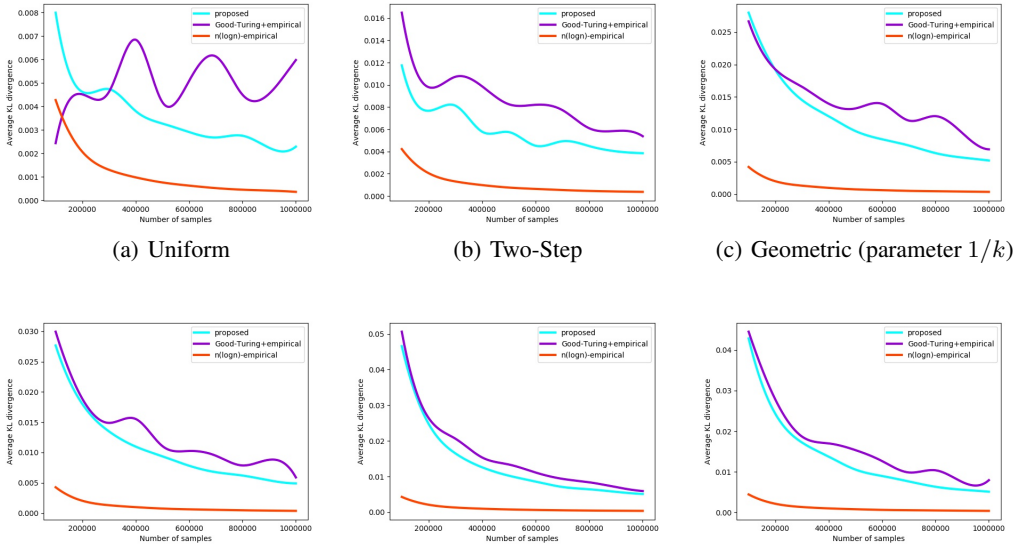
2 Experiments

Experimental plots and relevant details are shown below.

Estimators We consider three estimators: the proposed estimator with sample size n, the improved Good-Turing estimator [1] with the same sample size, and the empirical estimator with a larger $n \log n$ sample size. As shown in [1], the improved Good-Turing estimator considerably outperforms other estimators such as the Laplace estimator (add-1 estimator), the Krichevsky-Trofimov estimator [2], and the Braess-Sauer estimator [3]. Hence we do not include the latter estimators here.

Hyper-Parameters Our algorithm employs three hyper-parameters: c_1 is inversely related to the variance of the probability estimates and is best chosen above 1, c_2 controls the boundary between frequent and infrequent multiplicities and is best chosen below 1, and c_3 is proportional to the threshold separating small and large probabilities and is best chosen be around 1. In the experiments, we simply set $c_1 = 2$, $c_2 = 0.5$, and $c_3 = 1$.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



(a) Uniform (b) Two-Step (c) Geometric (parameter $1/k$)

(d) Zipf with parameter 0.5 (e) Uniform prior (Dirichlet 1) (f) Dirichlet 2 prior

Figure 1: Experimental results for support $k = 10,000$, number of samples n ranging from $10k$ to $100k$, averaged over 30 independent trials.

Distributions We choose alphabet size $k = 10,000$ and consider six different distributions over $[k]$: a uniform distribution of support size k; a two-step distribution with half the symbols having probability $1/(2k)$, and the other half having probability $3/(2k)$; a geometric distribution with parameter $g = 1/k$, i.e., $p_i = (1-g)^{i-1}g$, truncated at $i = k$ and renormalized; a Zipf distribution with parameter 0.5, i.e., $p_i \propto i^{-0.5}$, truncated at $i = k$ and renormalized; a distribution generated by the uniform prior on Δ_k; and a distribution generated by a Dirichlet-2 prior.

Experimental settings For each distribution we repeated the experiments 30 times and show the average KL-divergence between the underlying distribution and the distribution estimates. The relative performance of the three estimators is consistent over a wide range of sample sizes. To better differentiate the performance of the three estimators, we limit the dynamic range of the error by showing the results for sample sizes n ranging from $10 \cdot k$ to $100 \cdot k$.

Code The code is available at `https://github.com/ucsdyi/Competitive`.

Conclusions As can be observed in all six plots, the proposed estimator outperforms the improved Good-Turing estimator. Because of the estimator construction, outlined in Section 5 of the main paper, the improvement is most pronounced when $n \geq k$.

3 Proof of Theorem 1

In this section we prove Theorem 1 in the main paper.

Proof sketch From the discussion in Section 5 of the main paper, we need to estimate only M_μ. Relations such as $\mathbb{E}[M_{\mu-1}] = (\mu/n)\mathbb{E}[\Phi_\mu]$ suggest constructing estimators for $\mathbb{E}[\Phi_\mu]$. By the identity $\mathbb{E}[\Phi_\mu] = \sum_{x\in[k]} \mathbb{E}[\mathbb{1}_x^\mu]$, we can further reduce the problem to estimating $\mathbb{E}[\mathbb{1}_x^\mu]$. We then approximate $\mathbb{E}[\mathbb{1}_x^\mu]$ by scaled versions of $\mathbb{1}_x^{\mu'}$ where μ' is close to μ. This simple approach yields unbiased estimators $E_{x,\mu}^{\mu'}$ with sub-optimal variances. An important observation is that $\mathbb{1}_x^\mu \cdot \mathbb{1}_x^{\mu'} = 0$ for all $\mu \neq \mu'$, making it possible to construct a new estimator $E_{x,\mu}$ with near-optimal variance by averaging a sequence of these unbiased estimators. Note that $E_{x,\mu}$ is still unbiased. Summing the estimators over $[k]$, we estimate $\mathbb{E}[\Phi_\mu]$ by $E_\mu := \sum_{x\in[k]} E_{x,\mu}$. As shown in [6], a genie that knows both $\mathbb{E}[\Phi_{\mu+1}]$ and $\mathbb{E}[\Phi_\mu]$ could accurately estimate M_μ by $(\Phi_\mu(\mu+1)/n)(\mathbb{E}[\Phi_{\mu+1}]/\mathbb{E}[\Phi_\mu])$.

Hence, to approximate the genie's performance, we leverage the estimator for $\mathbb{E}[\Phi_\mu]$ and use $\hat{O}_\mu := (\Phi_\mu(\mu+1)/n)(E_{\mu+1}/E_\mu)$. Note that this estimator is the ratio of two estimators and hence not easy to analyze. To simplify the analysis, we modify E_μ slightly so that it has a structure similar to that of $E_{\mu+1}$. Then we prove that for relatively large, and frequent multiplicities, namely $\mu = \Omega(\log n)$ and $\Phi_\mu = \Omega(\log^2 n)$, the proposed estimator almost achieves the performance of the genie. As illustrated in Section 5 of the main paper, for other multiplicities, analysis shows that Good-Turing and empirical estimators are already near-optimal. Combined, these estimates form our final estimator for the vector M, and establish the guarantees stated in Theorem 1.

The Expected Total Probability Mass

To simplify our analysis, we adopt the standard "Poisson sampling" technique [4]. Instead of having a sample sequence of fixed length n, we make the sample size a Poisson random variable N with mean value n. Let p be an arbitrary distribution over $[k]$, and X^N be a length-$\mathrm{Poi}(n)$ sample sequence from p. Let N_x denote the number of times symbol x appearing in X^N, and let Φ_μ denote the number of symbols appearing μ times. For simplicity, denote $\mathbb{1}_x^\mu := \mathbb{1}_{N_x=\mu}$. Then, the total probability mass of the symbols that appear μ times is

$$M_\mu = \sum_{x\in[k]} p_x \mathbb{1}_x^\mu.$$

By the argument in Section 5 of the main paper, it suffices to design an estimator for M_μ.

The expectation of M_μ is

$$\mathbb{E}[M_\mu] = \mathbb{E}\left[\sum_{x\in[k]} p_x \mathbb{1}_x^\mu\right] = \sum_{x\in[k]} p_x e^{-np_x}\frac{(np_x)^\mu}{\mu!} = \frac{\mu+1}{n}\sum_{x\in[k]}\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right] = \frac{\mu+1}{n}\mathbb{E}\left[\Phi_{\mu+1}\right].$$

Furthermore, as shown in [6], a genie that knows both $\mathbb{E}[\Phi_{\mu+1}]$ and $\mathbb{E}[\Phi_\mu]$ could estimate M_μ really well using the estimator

$$O_\mu := \Phi_\mu \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}.$$

Both observations suggest that we should find a good estimator for $\mathbb{E}[\Phi_{\mu+1}]$.

Estimating an Indicator Variable

The above derivation shows that $\mathbb{E}\left[\Phi_{\mu+1}\right] = \sum_{x\in[k]}\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right]$. Symmetry further reduces the problem to estimating a single term $\mathbb{E}\left[\mathbb{1}_x^{\mu+1}\right]$. For notational convenience, we change $(\mu+1)$ to μ.

For any two natural numbers μ and μ', let $a_\mu^{\mu'} := \mu'!/\mu!$. Direct computation yields

$$\mathbb{E}[\mathbb{1}_x^\mu] = \mathbb{E}[\mathbb{1}_x^{\mu'}] a_\mu^{\mu'} (np_x)^{\mu-\mu'}.$$

To further simplify our derivations, let us assume that another two independent length-$\mathrm{Poi}(n)$ sample sequences from p are given, say $X^{N'}$ and $X^{N''}$ where $N' \sim \mathrm{Poi}(n)$ and $N'' \sim \mathrm{Poi}(n)$. Denote by N_x' and N_x'' the number of times symbol x appearing, and Φ_μ' and Φ_μ'' the number of symbols appearing μ times, in $X^{N'}$ and $X^{N''}$, respectively. This is equivalent to the commonly-used "sample splitting" technique [5], namely, we split the given sample sequence into three independent subsequences of roughly the same length. It is not hard to see that even without these additional sample sequences, performing sample splitting shall change the right-hand side of Theorem 1 by at most a multiplicative factor of three, hence does not affect the statement of the theorem. By the last identity and properties of Poisson random variables, for $\mu \geq \mu'$, the following estimator is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$,

$$E_{x,\mu}^{\mu'} := \mathbb{1}_x^{\mu'} a_\mu^{\mu'} (N_x')^{\underline{\mu-\mu'}},$$

where $A^{\underline{B}}$ is the falling factorial of A of order B.

Let c_1 be a positive absolute constant. In the subsequent proofs, we will assume that c_1 is sufficiently small and lies in $(0,1)$ to avoid large constants in the expressions. For $c_1 > 1$, the proof of Theorem 1

still follows from the remaining arguments. Other related constants have also been chosen to simplify the proofs and expressions. For example, we set $c_3 = 100$ to eliminate some edge cases.

While the bias of $E_{x,\mu}^{\mu'}$ in estimating $\mathbb{E}[\mathbb{1}_x^{\mu}]$ is zero, the variance of $E_{x,\mu}^{\mu'}$ satisfies

$$\mathrm{Var}(E_{x,\mu}^{\mu'}) \le \mathbb{E}[(E_{x,\mu}^{\mu'})^2] \le \left(a_\mu^{\mu'}\right)^2 \mathbb{E}[\mathbb{1}_x^{\mu'}] \cdot \mathbb{E}\left[\left((N'_x)^{\underline{\mu-\mu'}}\right)^2\right].$$

The quantity on the right-hand side is the product of three terms. We bound the first term using the following lemma.

Lemma 1. *For sufficiently large n and any two natural numbers μ, μ' such that $n \log n > \mu > 100 \log n$ and*

$$\mu - c_1\sqrt{\frac{\mu}{\log n}} \le \mu' \le \mu - 1,$$

we have

$$(a_\mu^{\mu'})^2 \le \frac{4}{(\mu^{\mu-\mu'})^2}.$$

Proof. The quantity of interest satisfies

$$\begin{aligned}
a_\mu^{\mu'} &= \frac{\mu'!}{\mu!} \\
&= \frac{1}{\mu^{\mu-\mu'}} \cdot \frac{\mu^{\mu-\mu'}}{\prod_{j=0}^{\mu-\mu'-1}(\mu-j)} \\
&\le \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{\mu}{\mu - c_1\sqrt{\frac{\mu}{\log n}}}\right)^{c_1\sqrt{\frac{\mu}{\log n}}} \\
&= \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{1}{1 - c_1\sqrt{\frac{1}{\mu\log n}}}\right)^{c_1\sqrt{\frac{\mu}{\log n}}} \\
&\le \frac{1}{\mu^{\mu-\mu'}} \cdot \left(\frac{1}{\left(1-\frac{1}{2}\right)^2}\right)^{\frac{c_1^2}{\log n}} \\
&\le \frac{2}{\mu^{\mu-\mu'}}.
\end{aligned}$$

□

Replacing the first term by the upper bound in the lemma implies

$$\frac{\mathrm{Var}(E_{x,\mu}^{\mu'})}{4} \le \mathbb{E}\left[\left(\frac{(N'_x)^{\underline{\mu-\mu'}}}{\mu^{\mu-\mu'}}\right)^2\right] \mathbb{E}[\mathbb{1}_x^{\mu'}].$$

It suffices to bound the last quantity. To proceed, we need the following concentration inequalities for Poisson random variables . Note that these inequities hold for any Poisson random variables, and simply follow from the well-known Chernoff bound [4].

Lemma 2. *For $X \sim \mathrm{Poi}(M)$ and any $\lambda > 0$,*

$$\mathbb{P}(X \le (1-\lambda)M) \le \left(\frac{e^{-\lambda}}{(1-\lambda)^{(1-\lambda)}}\right)^M \le e^{-\frac{\lambda^2 M}{2}},$$

and

$$\mathbb{P}(X \ge (1+\lambda)M) \le \left(\frac{e^{\lambda}}{(1+\lambda)^{(1+\lambda)}}\right)^M \le e^{-\frac{\min\{\lambda^2,\lambda\}M}{3}}.$$

Let c be a sufficiently large absolute constant. As a corollary of the lemma above, for any natural number $\mu > 100 \log n$ and j such that $\sqrt{\mu/(c \log n)} > j \geq 1$,

$$\Pr\left(N'_x > \mu + jc\sqrt{\mu \log n}\right) \Pr(\mathbb{1}_x^\mu = 1) \leq e^{-\Theta(j\sqrt{c}\log n)},$$

and for any $i \geq 1$ and natural number μ,

$$\Pr\left(N'_x > \mu + i\mu\right) \Pr(\mathbb{1}_x^\mu = 1) \leq e^{-\Theta(i\mu)},$$

Intuitively, Poisson random variables are highly concentrated around their mean values. Hence, for a Poisson random variable X and natural numbers a, b such that $a \gg b$, we should expect the product $\Pr(X \geq a) \cdot \Pr(X \leq b)$ to be small. We are ready to bound the quantity of interest.

Lemma 3. *For sufficiently large n and any two natural numbers μ, μ' such that $n \log n > \mu > 100 \log n$ and*

$$\mu - c_1\sqrt{\frac{\mu}{\log n}} \leq \mu' \leq \mu - 1,$$

we have

$$\mathbb{E}\left[\left(\frac{(N'_x)^{\mu-\mu'}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \leq e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).$$

Proof. The proof follows from the two concentration inequalities above. Note that for $\mu' \leq \mu - 1$, those inequalities still hold if we replace $\Pr(\mathbb{1}_x^\mu = 1)$ by $\Pr(\mathbb{1}_x^{\mu'} = 1)$.

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{(N'_x)^{\mu-\mu'}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \\
&\leq \left(1 + c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \sum_{j=1}^{\sqrt{\mu/(c\log n)}}\left(1 + (j+1)c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')} e^{-\Theta(j\sqrt{c}\log n)} \\
&+ \sum_{i=\sqrt{c}}^{\infty}(1 + (i+1))^{2(\mu-\mu')}e^{-\Theta(i\mu)} \\
&\leq e^{2c}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \sum_{j=1}^{\sqrt{\mu/(c\log n)}} e^{2(j+1)c}e^{-\Theta(j\sqrt{c}\log n)} + \sum_{i=\sqrt{c}}^{\infty} e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)} \\
&\leq e^{2c}\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}} + \sum_{i=\sqrt{c}}^{\infty} e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)} \\
&= e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).
\end{aligned}$$

□

Ignoring the $1/n^{\Theta(\sqrt{c})}$ term, the proof actually shows that $E_{x,\mu}^{\mu'}$ is at most a constant multiple of $\sqrt{\mathbb{E}[\mathbb{1}_x^{\mu'}]}$, with high probability.

Under Poisson sampling, the multiplicity N_x is also a Poisson random variable with mean np_x. Note that $\mathbb{E}[\mathbb{1}_x^{\mu'}] = e^{-np_x}(np_x)^{\mu'}/\mu'! \leq (np_x)e^{-np_x}(np_x)^{\mu'-1}/(\mu'-1)! = \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$. This observation together with an argument analogous to that above yields

Lemma 4. *Under the same conditions as in Lemma 3,*

$$\mathbb{E}\left[\left(\frac{(N'_x)^{\mu-\mu'}}{\mu^{\mu-\mu'}}\right)^2\right]\mathbb{E}[\mathbb{1}_x^{\mu'}] \leq e^{2c}\left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right).$$

Since $\mathbb{E}[\mathbb{1}_x^{\mu'}] = \Pr(N_x = \mu')$, and $E_{x,\mu}^{\mu'} = \mathcal{O}(\sqrt{\mathbb{E}[\mathbb{1}_x^{\mu'}]})$ with high probability, there exists an absolute constant c' satisfying

$$\Pr(E_{x,\mu}^{\mu'} \geq c') \leq \frac{p_x}{n^{\Theta(c)}}.$$

An Estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$

While $E_{x,\mu}^{\mu'}$ is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$, in the last section we showed that it can have a constant variance. To reduce the estimation variance, we estimate $\mathbb{E}[\mathbb{1}_x^\mu]$ by the following estimator

$$E_{x,\mu} := \frac{1}{c_1\sqrt{\mu/\log n}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E_{x,\mu}^{\mu'}.$$

The estimator simply averages a sequence of $E_{x,\mu}^{\mu'}$'s and remains as an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$. An important observation is that $E_{x,\mu}$ is the sum of $E_{x,\mu}^{\mu'} = \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N_x')^{\underline{\mu-\mu'}}$, and only one of these terms can be non-zero, as $\mathbb{1}_x^{\mu'} \cdot \mathbb{1}_x^\mu = 0$ for all $\mu \neq \mu'$. Therefore, the inequality $\Pr(E_{x,\mu}^{\mu'} \geq c') \leq p_x/n^{\Theta(c)}$ immediately translates to

$$\Pr\left(E_{x,\mu} > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \leq \frac{p_x}{n^{\Theta(c)}}.$$

We have designed $E_{x,\mu}$ in a way such that its variance would be small. Specifically,

Lemma 5. *Under the same conditions as in Lemma 3,*

$$\mathrm{Var}(E_{x,\mu}) \leq \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right).$$

Proof. The variance of $E_{x,\mu}$ satisfies

$$\begin{aligned}
\mathrm{Var}(E_{x,\mu}) &\overset{(a)}{=} \left(\frac{1}{c_1\sqrt{\mu/\log n}}\right)^2 \mathrm{Var}\left(\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N_x')^{\underline{\mu-\mu'}}\right) \\
&\overset{(b)}{\leq} \Theta\left(\frac{\log n}{\mu}\right) \mathbb{E}\left(\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N_x')^{\underline{\mu-\mu'}}\right)^2 \\
&\overset{(c)}{=} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{E}\left[\left(\mathbb{1}_x^{\mu'} a_\mu^{\mu'}(N_x')^{\underline{\mu-\mu'}}\right)^2\right] \\
&\overset{(d)}{\leq} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right),
\end{aligned}$$

where (a) follows from $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, (b) follows from $\mathrm{Var}(X) \leq \mathbb{E}X^2$, (c) follows from $\mathbb{1}_x^{\mu'} \cdot \mathbb{1}_x^\mu = 0$ for all $\mu \neq \mu'$, and (d) follows from Lemma 4. □

Estimating $\mathbb{E}[\Phi_\mu]$

The last section shows that $E_{x,\mu}$ is a well-behaved estimator for $\mathbb{E}[\mathbb{1}_x^\mu]$. Following the identity $\mathbb{E}[\Phi_\mu] = \sum_{x\in[k]} \mathbb{E}[\mathbb{1}_x^\mu]$, we naturally estimate $\mathbb{E}[\Phi_\mu]$ by

$$E_\mu := \sum_{x\in[k]} E_{x,\mu}.$$

By construction, E_μ is an unbiased estimator for $\mathbb{E}[\Phi_\mu]$. Due to Poisson sampling, all the multiplicities N_x are independent. Following lemma 20, the variance of E_μ admits

$$\begin{aligned}\mathrm{Var}(E_\mu) &= \sum_{x\in[k]} \mathrm{Var}(E_{x,\mu}) \\ &\le \sum_{x\in[k]} \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{p_x}{n^{\Theta(\sqrt{c})}}\right) \\ &= \Theta\left(\frac{\log n}{\mu}\right) \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left(\mathbb{E}[\Phi_{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}\right).\end{aligned}$$

Furthermore, by a non-asymptotic version of the Stirling's formula,

$$\mathbb{E}[\mathbb{1}_x^\mu] = e^{-np_x}\frac{(np_x)^\mu}{\mu!} \le e^{-\mu}\frac{\mu^\mu}{\mu!} \le e^{-\mu}\mu^\mu \cdot (2\pi)^{-1/2}\frac{e^\mu}{\mu^{\mu+1/2}} = \frac{1}{\sqrt{2\pi\mu}}.$$

Combining this with the following inequality mentioned in the last section,

$$\Pr\left(E_{x,\mu} > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \frac{p_x}{n^{\Theta(c)}},$$

we immediately get

$$\Pr\left(|E_{x,\mu} - \mathbb{E}[\mathbb{1}_x^\mu]| > \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \Pr\left(E_{x,\mu} > -\frac{1}{\sqrt{2\pi\mu}} + \frac{c'}{c_1\sqrt{\mu/\log n}}\right) \le \frac{p_x}{n^{\Theta(c)}},$$

where we have increased the value of c' by 1.

We are ready to characterize the tail probability of E_μ, for which we use the following variation [6] of the well-known Bernstein inequality.

Lemma 6. *Let $Y_1, \dots, Y_m$ be m independent variables such that with probability $\ge 1-\varepsilon_i$, $|Y_i - \mathbb{E}[Y_i]| < M$, then for any $\delta \in (0,1)$,*

$$\Pr\left(\left|\sum_i Y_i - \mathbb{E}\left[\sum_i Y_i\right]\right| > \sqrt{2\sum_i \mathrm{Var}(Y_i)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}\right) \le 2\delta + \sum_i \varepsilon_i.$$

Set $\delta = n^{-10}$, $m = k$, and $Y_x = E_{x,\mu}$ for all $x \in [k]$, and choose $M = c'/c_1\sqrt{\mu/\log n}$ and $\varepsilon_x = p_x/n^{\Theta(c)}$ for all $x \in [k]$. For a sufficiently large absolute constant c_4, the concentration inequality above combines all the previous results and yields

$$\Pr\left(|E_\mu - \mathbb{E}[\Phi_\mu]| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{E}[\Phi_{\mu'}]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

Next we derive a similar inequality for which $\mathbb{E}[\Phi_{\mu'}]$'s in the inner sum are replaced with $\mathbb{E}[\Phi_\mu]$.

To do this, we utilize the following lemma [6], which shows that $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu-1}]$ are often close to each other. Note that we have made the constants explicit.

Lemma 7. *For $\mu \ge 100\log n$,*

$$|\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]| \le 5\sqrt{\frac{\log n}{\mu}}\mathbb{E}[\Phi_{\mu-1}] + \frac{3}{n^2},$$

and for $\mu \ge 1$,

$$\mathbb{E}[\Phi_\mu] \le \mathcal{O}\left((\log n)\mathbb{E}[\Phi_{\mu-1}] + \frac{1}{n}\right).$$

By the above lemma, for $n \log n > \mu \geq 100 \log n$,

$$\mathbb{E}[\Phi_{\mu-1}] + \mu \frac{3}{n^2} \leq \left(1 + 10\sqrt{\frac{\log n}{\mu}}\right)\left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right).$$

This recursive inequality implies that for sufficiently small constant c_1 and any μ' satisfying $\mu - c_1\sqrt{\mu/\log n} \leq \mu' \leq \mu - 1$,

$$\begin{aligned}
\mathbb{E}[\Phi_{\mu'}] + \mu' \frac{3}{n^2} &\leq \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right) \prod_{i=\mu'+1}^{\mu} \left(1 + 10\sqrt{\frac{\log n}{i}}\right) \\
&\leq \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right) \left(1 + 10\sqrt{\frac{\log n}{\mu - c_1\sqrt{\mu/\log n}}}\right)^{c_1\sqrt{\mu/\log n}} \\
&\leq \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right) \left(1 + \sqrt{\frac{121\log n}{\mu}}\right)^{11c_1\sqrt{\frac{\mu}{121\log n}}} \\
&\leq \left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right) e^{11c_1} \\
&\leq 2\left(\mathbb{E}[\Phi_\mu] + (\mu+1)\frac{3}{n^2}\right),
\end{aligned}$$

where we have used the fact that $(1+1/x)^x < e$ for $x > 0$. Consequently, under the same conditions,

$$\mathbb{E}[\Phi_{\mu'}] \leq 2\mathbb{E}[\Phi_\mu] + 2(\mu+1)\frac{3}{n^2} \leq 2\mathbb{E}[\Phi_\mu] + \frac{7\log n}{n}.$$

Hence for sufficiently small c_1,

$$\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{E}[\Phi_{\mu'}] \leq 2c_1\sqrt{\mu/\log n}\left(\mathbb{E}[\Phi_\mu] + \frac{7\log n}{n}\right) \leq 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu] + \frac{7\log n}{\sqrt{n}}.$$

This together with the previous tail bound yields

Lemma 8. *For $n \log n > \mu \geq 100 \log n$,*

$$\Pr\left(|E_\mu - \mathbb{E}[\Phi_\mu]| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

An Alternative Estimator for $\mathbb{E}[\Phi_{\mu-1}]$

Under the proper conditions mentioned previously, $E_{x,\mu-1}$ is not only unbiased in estimating $\mathbb{E}[\mathbb{1}_x^{\mu-1}]$, but also has small variance. However, our latter analysis calls for bounding the difference between $E_{x,\mu-1}$ and $E_{x,\mu}$, and it is inconvenient to use $E_{x,\mu-1}$ since it may have fewer terms than $E_{x,\mu}$. Hence to simplify our derivations, we construct the following estimator for $\mathbb{E}[\mathbb{1}_x^{\mu-1}]$,

$$E'_{x,\mu-1} := \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E_{x,(\mu-1)}^{\mu'},$$

and consequently estimate $\mathbb{E}[\Phi_{\mu-1}]$ by

$$E'_{\mu-1} := \sum_{x\in[k]} E'_{x,\mu-1}.$$

By an argument that is almost the same as that in the last few sections,

Lemma 9. *For $n \log n > \mu \geq 100 \log n$,*

$$\Pr\left(\left|E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]\right| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_{\mu-1}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

The Difference between Two Estimators

In this section, we consider

$$E_\mu^{(1)} := E_\mu - E'_{\mu-1} = \sum_{x\in[k]} (E_{x,\mu} - E'_{x,\mu-1}),$$

the difference between the two estimators E_μ and $E'_{\mu-1}$. We show that, $E_\mu^{(1)}$, as an unbiased estimator for $\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]$, highly concentrates around its mean. In the subsequent sections, we leverage this property to design an accurate estimator for the total probability M_μ.

Similar to the previous derivations, we start by considering a single term

$$E_{x,\mu}^{(1)} := E_{x,\mu} - E'_{x,\mu-1}.$$

We can bound the absolute value of $E_{x,\mu}^{(1)}$ as follows.

$$\begin{aligned}
|E_{x,\mu}^{(1)}| &= |E_{x,\mu} - E'_{x,\mu-1}| \\
&= \left| \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E_{x,\mu}^{\mu'} - \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E_{x,(\mu-1)}^{\mu'} \right| \\
&\le \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left| E_{x,\mu}^{\mu'} - E_{x,\mu}^{\mu-1} \right| \\
&= \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \left| \mathbb{1}_x^{\mu'} a_\mu^{\mu'} (N'_x)^{\underline{\mu-\mu'}} - \mathbb{1}_x^{\mu'} a_{\mu-1}^{\mu'} (N'_x)^{\underline{(\mu-1)-\mu'}} \right| \\
&= \frac{1}{c_1}\sqrt{\frac{\log n}{\mu}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} \mathbb{1}_x^{\mu'} a_\mu^{\mu'} (N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right|.
\end{aligned}$$

The above inequality together with $\mathrm{Var}(E_{x,\mu}^{(1)}) \le \mathbb{E}(E_{x,\mu}^{(1)})^2$ implies

$$\mathrm{Var}(E_{x,\mu}^{(1)}) \le \frac{1}{c_1^2}\frac{\log n}{\mu} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} (a_\mu^{\mu'})^2 \mathbb{E}[\mathbb{1}_x^{\mu'}] \mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right| \Big)^2,$$

where we have used $\mathbb{1}_x^{\mu'} \cdot \mathbb{1}_x^{\mu} = 0$ for all $\mu \neq \mu'$. Note that the bound on the right-hand side is a sum of three-term products. Assume that $n \gg 1$ and $n \log n > \mu > 100 \log n$, and consider one of these products that corresponds to an arbitrary μ' satisfying $\mu - c_1(\mu/\log n) \le \mu' \le \mu - 1$. Lemma 1 bounds its first term as $(a_\mu^{\mu'})^2 \le 4/(\mu^{\mu-\mu'})^2$. Replacing the first term with this bound, the following lemma further upper bounds the resulting quantity.

Lemma 10. *Under the same conditions as in Lemma 3,*

$$\frac{\mathbb{E}[\mathbb{1}_x^{\mu'}]}{(\mu^{\mu-\mu'})^2} \mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}} \left| (N'_x - \mu) - (\mu - \mu') + 1 \right| \Big)^2 \le \Theta\left(\frac{\log n}{\mu}\right) \mathbb{E}[\mathbb{1}_x^{\mu'}] + \frac{1}{n^{\Theta(\sqrt{c})}}.$$

Proof. Since $(N'_x - \mu)$ can be negative, we need the concentration inequality

$$\Pr\left(N'_x - \mu < -c\sqrt{\mu \log n} \right) \Pr(\mathbb{1}_x^{\mu} = 1) \le e^{-\Theta(\sqrt{c}\log n)},$$

which follows from Lemma 2. Similar to the proof of Lemma 3, we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_x^{\mu'}]\mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}}|(N'_x-\mu)-(\mu-\mu')+1|\Big)^2\frac{1}{(\mu^{\mu-\mu'})^2}\\
&\le \mathbb{E}[\mathbb{1}_x^{\mu'}]\left(\frac{2c\sqrt{\mu\log n}}{\mu}\right)^2\left(1+c\sqrt{\frac{\log n}{\mu}}\right)^{2((\mu-1)-\mu')}+\left(\frac{2\mu-1}{u}\right)^2 e^{-\Theta(\sqrt{c}\log n)}\\
&+\sum_{j=1}^{\sqrt{\mu/(c\log n)}}\left(1+(j+1)c\sqrt{\frac{\log n}{\mu}}\right)^{2(\mu-\mu')}e^{-\Theta(j\sqrt{c}\log n)}+\sum_{i=\sqrt{c}}^{\infty}(1+(i+1))^{2(\mu-\mu')}e^{-\Theta(i\mu)}.
\end{aligned}$$

Since $(1+1/x)^x < e$ for $x > 0$, the first term on the right-hand side can be bounded by $4(c^2 \cdot e^{2c})\mathbb{E}[\mathbb{1}_x^{\mu'}](\log n)/\mu$. The sum of the remaining three terms is at most

$$4e^{-\Theta(\sqrt{c}\log n)}+\sum_{j=1}^{\sqrt{\mu/(c\log n)}}e^{2(j+1)c}e^{-\Theta(j\sqrt{c}\log n)}+\sum_{i=\sqrt{c}}^{\infty}e^{\Theta(\sqrt{\mu}\log i)}e^{-\Theta(i\mu)}\le\frac{1}{n^{\Theta(\sqrt{c})}}.$$

Consolidating these bounds yields the desired result. □

By $\mathbb{E}[\mathbb{1}_x^{\mu'}] \le \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$, an analogous argument yields

Lemma 11. *Under the same conditions as in Lemma 3,*

$$\frac{\mathbb{E}[\mathbb{1}_x^{\mu'}]}{(\mu^{\mu-\mu'})^2}\mathbb{E}\Big((N'_x)^{\underline{(\mu-1)-\mu'}}|(N'_x-\mu)-(\mu-\mu')+1|\Big)^2\le\Theta\left(\frac{\log n}{\mu}\right)\mathbb{E}[\mathbb{1}_x^{\mu'}]+\frac{p_x}{n^{\Theta(\sqrt{c})}}.$$

There is always a unique μ' such that $\mathbb{1}_x^{\mu'} = 1$. The proof of Lemma 10 together with $\mathbb{E}[\mathbb{1}_x^{\mu'}] \le \mathbb{E}[\mathbb{1}_x^{\mu'-1}](np_x)$ also shows that for a sufficiently large absolute constant c'',

$$\Pr\left(|E_{x,\mu}^{(1)}|>\frac{c''\log n}{\mu}\right)\le\frac{p_x}{n^{\Theta(c)}}.$$

Furthermore, the expectation of $E_{x,\mu}^{(1)}$ satisfies

Lemma 12. *For any natural number μ such that $n\log n > \mu \ge 100\log n$,*

$$|\mathbb{E}[E_{x,\mu}^{(1)}]|\le 1/\mu.$$

Proof. Recall that $E_{x,\mu}^{(1)}$ is an unbiased estimator for $\mathbb{E}[\mathbb{1}_x^{\mu}] - \mathbb{E}[\mathbb{1}_x^{\mu-1}]$. Therefore,

$$\begin{aligned}
|\mathbb{E}[E_{x,\mu}^{(1)}]| &= |\mathbb{E}[\mathbb{1}_x^{\mu}]-\mathbb{E}[\mathbb{1}_x^{\mu-1}]|\\
&=\left|e^{-np_x}\frac{(np_x)^{\mu}}{\mu!}-e^{-np_x}\frac{(np_x)^{\mu-1}}{(\mu-1)!}\right|\\
&=\left|e^{-np_x}\frac{(np_x)^{\mu-1}(np_x-\mu)}{\mu!}\right|.
\end{aligned}$$

In general, consider the function $g_\mu(y) := e^{-y}y^{\mu-1}(y-\mu)/\mu!$ for $y \ge 0$. The first-order derivative of $g_\mu(y)$ with respect to y is

$$g'_\mu(y)=-\frac{1}{\mu!}e^{-y}y^{-2+\mu}(\mu^2+y^2-\mu(1+2y))$$

which has two roots, $y_1 := \mu - \sqrt{\mu}$ and $y_2 := \mu + \sqrt{\mu}$. Since both $g_\mu(0)$ and $\lim_{y\to\infty} g(y)$ equal to zero, the maximum of $|g_\mu(y)|$ for $y \ge 0$ is $\max\{|g(y_1)|, |g(y_2)|\}$. By a non-asymptotic version of

the Stirling's formula,

$$\begin{aligned}
|g(y_1)| &= e^{-\mu+\sqrt{\mu}}\frac{(\mu-\sqrt{\mu})^{\mu-1}\sqrt{\mu}}{\mu!} \\
&\le e^{-\mu+\sqrt{\mu}}(\mu-\sqrt{\mu})^{\mu-1}\sqrt{\mu}\frac{e^{\mu}}{\sqrt{2\pi}\mu^{\mu+\frac{1}{2}}} \\
&= e^{\sqrt{\mu}}\left(\frac{\mu-\sqrt{\mu}}{\mu}\right)^{\mu-1}\frac{1}{\sqrt{2\pi}\mu} \\
&= e^{\sqrt{\mu}}\left(1-\frac{1}{\sqrt{\mu}}\right)^{\sqrt{\mu}(\sqrt{\mu}-(1/\sqrt{\mu}))}\frac{1}{\sqrt{2\pi}\mu} \\
&\le e^{\sqrt{\mu}}e^{-(\sqrt{\mu}-(1/\sqrt{\mu}))}\frac{1}{\sqrt{2\pi}\mu} \\
&= \frac{e^{1/\sqrt{\mu}}}{\sqrt{2\pi}\mu} \le \frac{1}{\mu}.
\end{aligned}$$

Similarly, we can also show that $|g(y_1)| \le 1/\mu$. □

Increase the value of c'' by 1. The above lemma implies

$$\Pr\left(|E^{(1)}_{x,\mu} - \mathbb{E}[E^{(1)}_{x,\mu}]| > \frac{c''\log n}{\mu}\right) \le \frac{p_x}{n^{\Theta(c)}}.$$

Turning back to $E^{(1)}_{\mu}$ and using Lemma 11, we can bound the variance of $E^{(1)}_{\mu}$ as

$$\mathrm{Var}(E^{(1)}_{\mu}) \le \sum_{x\in[k]}\mathrm{Var}(E^{(1)}_{x,\mu}) \le \Theta\left(\frac{\log^2 n}{\mu^2}\right)\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\left(\mathbb{E}[\Phi_{\mu'}] + \frac{1}{n^{\Theta((c/k)\wedge k)}}\right).$$

Let c'_4 be a sufficiently large absolute constant. By the Bernstein-inequality variation in Lemma 6,

$$\Pr\left(\left|E^{(1)}_{\mu} - \mathbb{E}[E^{(1)}_{\mu}]\right| > c'_4\frac{\log^2 n}{\mu}\sqrt{1+\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\mathbb{E}[\Phi_{\mu'}]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

Furthermore, by Lemma 7, for sufficiently small constant c_1 and any μ' satisfying $\mu - c_1\sqrt{\mu/\log n} \le \mu' \le \mu - 1$,

$$\sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1}\mathbb{E}[\Phi_{\mu'}] \le 2c_1\sqrt{\mu/\log n}\left(\mathbb{E}[\Phi_{\mu}] + \frac{7\log n}{n}\right) \le 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_{\mu}] + \frac{7\log n}{\sqrt{n}}.$$

Combined, the two inequalities above yield

Lemma 13. *For $n\log n > \mu \ge 100\log n$,*

$$\Pr\left(\left|E^{(1)}_{\mu} - \mathbb{E}[E^{(1)}_{\mu}]\right| > c'_4\frac{\log^2 n}{\mu}\sqrt{2+2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_{\mu}]}\right) \le \Theta\left(\frac{1}{n^{10}}\right).$$

Estimating the Total Probability Mass

A genie estimator that knows both $\mathbb{E}[\Phi_{\mu}]$ and $\mathbb{E}[\Phi_{\mu-1}]$ could accurately estimate $M_{\mu-1}$ by

$$O_{\mu-1} := \Phi_{\mu-1}\frac{\mu}{n}\frac{\mathbb{E}[\Phi_{\mu}]}{\mathbb{E}[\Phi_{\mu-1}]}.$$

and achieve the following guarantee [6] for a sufficiently large constant c''_4.

Lemma 14. *For μ satisfying $n\log n > \mu \geq 100\log n$ and $\mathbb{E}[\Phi_{\mu-1}] \geq 1$,*

$$\Pr\left(|M_{\mu-1} - O_{\mu-1}| \geq c_4'' \frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) \leq \mathcal{O}\left(\frac{1}{n^{10}}\right).$$

Replace $\mathbb{E}[\Phi_\mu]/\mathbb{E}[\Phi_{\mu-1}]$ with $E_\mu/E'_{\mu-1}$. Our estimator is simply

$$\hat{O}_{\mu-1} := \Phi_{\mu-1}\frac{\mu}{n}\frac{E_\mu}{E'_{\mu-1}}.$$

Note that we use $E'_{\mu-1}$ instead of $E_{\mu-1}$ just to simplify the proofs. Clearly, our objective is to characterize the estimation error $|M_{\mu-1} - \hat{O}_{\mu-1}|$. By the triangle inequality and the above lemma, it suffices to bound $\left|O_{\mu-1} - \hat{O}_{\mu-1}\right|$. To do this, we use the following interesting result.

Lemma 15. *If $b > 0$, $b + \Delta b > 0$, and $|\Delta b| \leq 0.9b$,*

$$\left|\frac{a+\Delta a}{b+\Delta b} - \frac{a}{b}\right| \leq \mathcal{O}\left(\frac{|\Delta b||a| + |\Delta a||b|}{b^2}\right).$$

The above lemma appears in [6] and follows by simple algebra. Set $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, $b = \mathbb{E}[\Phi_{\mu-1}]$, $\Delta a = E_\mu - E'_{\mu-1} - \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, and $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$. Note that $a = \mathbb{E}[E_\mu^{(1)}]$ and $\Delta a = E_\mu^{(1)} - \mathbb{E}[E_\mu^{(1)}]$. Assuming that $n\log n > \mu \geq 100\log n$, we analyze each term below.

For $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, by Lemma 7,

$$|a| \leq 5\sqrt{\frac{\log n}{\mu}}\mathbb{E}[\Phi_{\mu-1}] + \frac{3}{n^2}.$$

For $\Delta a = E_\mu^{(1)} - \mathbb{E}[E_\mu^{(1)}]$, as shown in Lemma 13,

$$\Pr\left(|\Delta a| > c_4'\frac{\log^2 n}{\mu}\sqrt{2 + 2c_1\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

For $b = \mathbb{E}[\Phi_{\mu-1}]$, Lemma 7 implies a lower bound

$$b \geq \left(1 + 5\sqrt{\frac{\log n}{\mu}}\right)^{-1}\left(\mathbb{E}[\Phi_\mu] - \frac{3}{n^2}\right) \geq \frac{2}{3}\mathbb{E}[\Phi_\mu] - \frac{2}{n^2},$$

as well as an upper bound

$$b \leq \left(1 + 10\sqrt{\frac{\log n}{\mu}}\right)\mathbb{E}[\Phi_\mu] + \frac{3}{n^2} \leq 2\mathbb{E}[\Phi_\mu] + \frac{3}{n^2}.$$

For $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$, Lemma 9 states that

$$\Pr\left(|\Delta b| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{1}{c_1^2} + 2c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_{\mu-1}]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

Our bound on $|b|$ further implies

$$\Pr\left(|\Delta b| > c_4\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\frac{2}{c_1^2} + 6c_1\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_\mu]}\right) \leq \Theta\left(\frac{1}{n^{10}}\right).$$

Here, we can choose a sufficiently large constant c_5 so that, if $n \gg 1$, $\mu > 100\log n$, and $\mathbb{E}[\Phi_{\mu-1}] > c_5(\log^2 n)/10$, then $|\Delta b| < 0.9b$ with probability at least $1 - \Theta(n^{-10})$. Also note that $\Phi_\mu =$

$\sum_{x\in[k]} \mathbb{1}_x^\mu$. In Lemma 6, set $\delta = n^{-10}$, $m = k$, and $Y_x = \mathbb{1}_x^\mu$ for all $x \in [k]$, $M = 1$, and choose $\varepsilon_x = 0$ for all $x \in [k]$. Then,

$$\Pr\left(|\Phi_\mu - \mathbb{E}[\Phi_\mu]| > \sqrt{20\sum_{x\in[k]} \mathrm{Var}(\mathbb{1}_x^\mu)\log n} + \frac{20}{3}\log n\right) \le \frac{2}{n^{10}}.$$

Together with $\mathrm{Var}(\mathbb{1}_x^\mu) \le \mathbb{E}(\mathbb{1}_x^\mu)^2 = \mathbb{E}[\mathbb{1}_x^\mu]$, the above inequality implies

Lemma 16. *For sufficiently large n, and μ satisfying $\mu > 100\log n$ and $\mathbb{E}[\Phi_{\mu-1}] > c_5(\log^2 n)/10$,*

$$\Pr\left(|\Phi_\mu - \mathbb{E}[\Phi_\mu]| > \sqrt{20\mathbb{E}[\Phi_\mu]\log n} + \frac{20}{3}\log n\right) \le \frac{2}{n^{10}}.$$

For our purpose, it suffices to apply the estimator $\hat{O}_{\mu-1}$ to indices μ satisfying $\mu > 100\log n$ and $\mathbb{E}[\Phi_\mu] \ge 0.5c_5\log^2 n$. While not knowing p, we can use the independent sample sequence $X^{N''}$ to ensure that with high probability, $\mathbb{E}[\Phi_\mu] \ge 0.5c_5\log^2 n$. More concretely, we only apply $\hat{O}_{\mu-1}$ to indices μ satisfying $\Phi''_{\mu-1} > c_5\log^2 n$. By construction, $\mathbb{E}[\Phi_\mu] = \mathbb{E}\left[\Phi''_\mu\right]$. Then for sufficiently large c_5 and n, and $\mathbb{E}[\Phi_{\mu-1}] < 0.5c_5\log^2 n$, Lemma 16 implies

$$\Pr\left(\Phi''_{\mu-1} > c_5\log^2 n\right) \le \Pr\left(\left|\Phi''_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]\right| > \sqrt{20\mathbb{E}[\Phi_{\mu-1}]\log n} + \frac{20}{3}\log n\right) \le \frac{2}{n^{10}}.$$

Hence for μ satisfying the conditions mentioned previously, we can assume that $\mathbb{E}[\Phi_{\mu-1}] \ge 0.5c_5\log^2 n$. Under this assumption, Lemma 7 implies that $\mathbb{E}[\Phi_\mu] \ge \mathbb{E}[\Phi_{\mu-1}]/3 \ge c_5(\log^2 n)/6$. By the same reasoning, $\mathbb{E}[\Phi_\mu]/18 \le \mathbb{E}[\Phi_{\mu-1}]/6 \le \Phi_{\mu-1} \le 6\mathbb{E}[\Phi_{\mu-1}] \le 18\mathbb{E}[\Phi_\mu]$ with probability at least $1 - \Theta(n^{-10})$. In other words, we can also assume that $\Phi_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_\mu])$.

Recall that $a = \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, $b = \mathbb{E}[\Phi_{\mu-1}]$, $\Delta a = E_\mu - E'_{\mu-1} - \mathbb{E}[\Phi_\mu - \Phi_{\mu-1}]$, and $\Delta b = E'_{\mu-1} - \mathbb{E}[\Phi_{\mu-1}]$. The union bound together with Lemma 15 combines all the results in this section and yields that with probability at least $1 - \Theta(n^{-10})$,

$$\begin{aligned}
\left|O_{\mu-1} - \hat{O}_{\mu-1}\right| &= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{E_\mu}{E_{\mu-1}} - \frac{\mathbb{E}[\Phi_\mu]}{\mathbb{E}[\Phi_{\mu-1}]}\right| \\
&= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{E_\mu - E_{\mu-1}}{E_{\mu-1}} - \frac{\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu-1}]}{\mathbb{E}[\Phi_{\mu-1}]}\right| \\
&= \Phi_{\mu-1}\frac{\mu}{n}\left|\frac{a+\Delta a}{b+\Delta b} - \frac{a}{b}\right| \\
&\le \mathcal{O}\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{|\Delta b||a| + |\Delta a||b|}{b^2}\right) \\
&\le \mathcal{O}\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{\left(\frac{\log^{\frac{3}{2}} n}{\sqrt{\mu}}\sqrt{\sqrt{\frac{\mu-1}{\log n}}\mathbb{E}[\Phi_\mu]}\right)\sqrt{\frac{\log n}{\mu+1}}\mathbb{E}[\Phi_{\mu-1}]}{(\mathbb{E}[\Phi_\mu])^2}\right) \\
&\quad + \mathcal{O}\left(\Phi_{\mu-1}\frac{\mu}{n}\frac{\left(\frac{\log^2 n}{\mu}\sqrt{\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right)\mathbb{E}[\Phi_\mu]}{(\mathbb{E}[\Phi_\mu])^2}\right) \\
&= \mathcal{O}\left(\frac{\log^2 n}{n}\sqrt{\sqrt{\frac{\mu}{\log n}}\mathbb{E}[\Phi_\mu]}\right).
\end{aligned}$$

Again, we can make c_5 sufficiently large so that with probability at least $1 - \mathcal{O}(n^{-10})$, the upper bound is at most $0.9O_{\mu-1}$ and $O_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]\mu/n)$. Combined, the upper bound of $0.9O_{\mu-1}$, the identity $\Phi_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_\mu])$, and Lemma 14 imply

Lemma 17. *For μ satisfying $n\log n > \mu \geq 100\log n$ and $\Phi''_\mu > c_5\log^2 n$,*

$$\Pr\left(|M_{\mu-1} - \hat{O}_{\mu-1}| \geq 2c''_4\frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) \leq \mathcal{O}\left(\frac{1}{n^{10}}\right).$$

Therefore, with probability at least $1-\mathcal{O}(n^{-10})$, we have both $\hat{O}_{\mu-1} = \Theta(\mathbb{E}[\Phi_{\mu-1}]\mu/n)$ and

$$|M_{\mu-1} - \hat{O}_{\mu-1}| \leq \mathcal{O}\left(\frac{\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n}{n}\right) = \mathcal{O}\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^2 n}{n}\right).$$

Furthermore, if these two claims hold,

$$\frac{(M_{\mu-1} - \hat{O}_{\mu-1})^2}{\hat{O}_{\mu-1}} \leq \mathcal{O}\left(\frac{\left((\sqrt{\mathbb{E}[\Phi_{\mu-1}](\mu-1)}\log^2 n)/n\right)^2}{\mathbb{E}[\Phi_{\mu-1}]\mu/n}\right) \leq \mathcal{O}\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

Finally, we note that these results hold with high probability, i.e., $1-\mathcal{O}(n^{-10})$, instead of surely. To make sure that the KL-divergence between the underlying truth and our estimates is not infinity, we modify our estimator slightly and denote

$$\hat{O}'_{\mu-1} := \min\{\max\{1/n, \hat{O}_{\mu-1}\}, \log^2 n\}.$$

We use $\hat{O}'_{\mu-1}$ to estimate $M_{\mu-1}$ iff μ satisfies $n\log n > \mu \geq 100\log n$ and $\Phi''_\mu > 2c_5\log^2 n$. Note that this estimator also admits the above inequalities, since with probability at least $1-\mathcal{O}(n^{-10})$, the value of the original estimator satisfies $\hat{O}_{\mu-1} = \Theta(\mu\Phi_{\mu-1}/n) \leq \mathcal{O}(N/n) = \mathcal{O}(\log n) < \log^2 n$ and $\hat{O}_{\mu-1} = \Theta(\mu\Phi_{\mu-1}/n) \geq \Omega((\log^3 n)/n) > 1/n$, implying that $\hat{O}'_{\mu-1} = \hat{O}_{\mu-1}$.

The Good-Turing Estimator

The Good-Turing estimator estimates $M_{\mu-1}$ by

$$\hat{G}_{\mu-1} := \frac{\mu}{n}\Phi_\mu.$$

Let c'_5 be a sufficiently large absolute constant. The following lemma [7] characterizes the performance of $\hat{G}_{\mu-1}$ in estimating $M_{\mu-1}$.

Lemma 18. *For μ satisfying $\mathbb{E}[\Phi_{\mu-1}] \geq 1$ and $\delta \in (0,1)$,*

$$\Pr\left(|M_{\mu-1} - \hat{G}_{\mu-1}| > c'_5\sqrt{\mathbb{E}[\Phi_\mu]+1}\frac{\mu\log^2\frac{n}{\delta}}{n}\right) \leq \delta.$$

For indices μ satisfying $2 \leq u \leq 100\log n$ and $\Phi''_{\mu-1} > 2c_5(\log^2 n)$, we simply use the following variant of the Good-Turing estimator,

$$\hat{G}'_{\mu-1} := \max\left\{\frac{1}{n}, \hat{G}_{\mu-1}\right\}.$$

Given $\Phi''_{\mu-1} > 2c_5(\log^2 n)$, by derivations in the last section, we can assume that $\Phi_{\mu-1} = \Theta(\Phi_\mu) = \Theta(\mathbb{E}[\Phi_{\mu-1}]) = \Theta(\mathbb{E}[\Phi_\mu]) \geq \log^2 n$, and with probability at least $1-\mathcal{O}(n^{-10})$, we would be correct. Choose $\delta = n^{-10}$ in Lemma 18. Then,

$$\Pr\left(|M_{\mu-1} - \hat{G}_{\mu-1}| > 15^2 c'_5\sqrt{\mathbb{E}[\Phi_\mu]}\frac{\mu\log^2 n}{n}\right) \leq \frac{1}{n^{10}}.$$

Additionally, note that $\mu \leq 100\log n$. Hence with probability at least $1-\mathcal{O}(n^{-10})$,

$$|M_{\mu-1} - \hat{G}_{\mu-1}| \leq \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_\mu]}\frac{\mu\log^2 n}{n}\right) \leq \mathcal{O}\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^{5/2} n}{n}\right),$$

and

$$\frac{(M_{\mu-1}-\hat{G}_{\mu-1})^2}{\hat{G}_{\mu-1}} \le \mathcal{O}\left(\frac{\left(\sqrt{\mathbb{E}[\Phi_\mu]}\frac{\mu\log^2 n}{n}\right)^2}{\frac{\mu}{n}\mathbb{E}[\Phi_\mu]}\right) = \mathcal{O}\left(\frac{\mu\log^4 n}{n}\right) = \mathcal{O}\left(\frac{\log^5 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

The estimator $G'_{\mu-1}$ also admits these inequalities since with probability at least $1-\mathcal{O}(n^{-10})$, we have $G'_{\mu-1} = \mu\Phi_\mu/n \ge 2(\log^3 n)/n > 1/n$, implying $G'_\mu = G_\mu$.

An Estimator for M_0

For $\mu = 1$, regardless of the value of $\Phi''_{\mu-1}$, we estimate the total probability $M_{\mu-1} = M_0$, by the estimator $\hat{G}'_0 = \max\{1,\Phi_1\}/n$. We divide our analysis into two cases according to $\mathbb{E}[\Phi_0]$.

Case 1: If $\mathbb{E}[\Phi_0] \ge 1$, then by Lemma 18, with probability at least $1-\mathcal{O}(n^{-10})$,

$$|M_0 - \hat{G}'_0| \le |M_0 - \hat{G}_0| + \frac{1}{n} \le \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_1]+1}\frac{\log^2 n}{n}\right)$$

If $\mathbb{E}[\Phi_1] \ge c_5\log^2 n$, then by Lemma 16 and arguments in the last section, with probability at least $1-\mathcal{O}(n^{-10})$, we have $\mathbb{E}[\Phi_1] = \Theta(\Phi_1) \ge \Omega(\log^2 n)$. This together with the above inequality further implies $\hat{G}'_0 = \Phi_1/n$ and

$$|M_0 - \hat{G}'_0| \le \mathcal{O}\left(\sqrt{\Phi_1}\frac{\log^2 n}{n}\right).$$

Therefore, with probability at least $1-\mathcal{O}(n^{-10})$, we have $\Phi_1 > 0$ and

$$\frac{(M_0-\hat{G}'_0)^2}{\hat{G}'_0} \le \mathcal{O}\left(\frac{\left(\sqrt{\Phi_1}\frac{\log^2 n}{n}\right)^2}{\Phi_1/n}\right) \le \mathcal{O}\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_1>0}\right).$$

If $\mathbb{E}[\Phi_1] < c_5\log^2 n$, then by the first inequality, with probability at least $1-\mathcal{O}(n^{-10})$,

$$|M_0 - \hat{G}'_0| \le \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_1]+1}\frac{\log^2 n}{n}\right) \le \mathcal{O}\left(\frac{\log^3 n}{n}\right),$$

which further implies

$$\frac{(M_0-\hat{G}'_0)^2}{\hat{G}'_0} \le \mathcal{O}\left(\frac{\left(\frac{\log^3 n}{n}\right)^2}{1/n}\right) \le \mathcal{O}\left(\frac{\log^6 n}{n}\right).$$

Case 2: If $\mathbb{E}[\Phi_0] \le 1$, then by Lemma 7,

$$\mathbb{E}[\Phi_1] \le \mathcal{O}\left((\log n)\mathbb{E}[\Phi_0] + \frac{1}{n}\right) \le \mathcal{O}(\log n).$$

Furthermore, by Lemma 16, with probability at least $1-\mathcal{O}(n^{-10})$,

$$\Phi_0 \le \mathcal{O}(\log n).$$

For $\delta \in (0,1)$ and symbols x satisfying $p_x \ge \log(n/\delta)/n$, we have $\Pr(\mathbb{1}_x^0 = 1) = e^{-np_x} \le \delta/n$. Note that the number of such symbols is at most n. Hence by the union bound,

$$\Pr\left(\exists x\in[k] \text{ s.t. } p_x > \frac{\log(n/\delta)}{n}, \mathbb{1}_x^0 = 1\right) \le n\cdot\frac{\delta}{n} = \delta.$$

Setting $\delta = n^{-10}$ in the above inequality yields

$$\Pr\left(\forall x\in[k] \text{ s.t. } p_x > \frac{11\log(n)}{n}, \mathbb{1}_x^0 = 0\right) \ge 1-n^{-10}.$$

Therefore if we further have $\Phi_0 \leq \mathcal{O}(\log n)$,

$$M_0 = \sum_{x\in[k]} \mathbb{1}_x^0 \cdot p_x \leq \mathcal{O}(\Phi_0) \cdot \frac{11\log(n)}{n} = \mathcal{O}\left(\frac{\log^2 n}{n}\right).$$

In addition, since $\mathbb{E}[\Phi_1] \leq \mathcal{O}(\log n)$, Lemma 16 implies that with probability at least $1 - \mathcal{O}(n^{-10})$,

$$\Phi_1 \leq \mathcal{O}(\log n).$$

Consolidating these results shows that with probability at least $1 - \mathcal{O}(n^{-10})$,

$$|M_0 - \hat{G}_0'| \leq \mathcal{O}\left(\frac{\log^2 n}{n} + \frac{\log n}{n}\right) = \mathcal{O}\left(\frac{\log^2 n}{n}\right).$$

and

$$\frac{(M_0 - \hat{G}_0')^2}{\hat{G}_0'} \leq \mathcal{O}\left(\frac{\left(\frac{\log^2 n}{n}\right)^2}{1/n}\right) \leq \mathcal{O}\left(\frac{\log^4 n}{n}\right).$$

Summary of case 1 and 2: With probability at least $1 - \mathcal{O}(n^{-10})$,

$$|M_0 - \hat{G}_0'| = \mathcal{O}\left(\frac{(\sqrt{\Phi_1}+1)\log^3 n}{n}\right)$$

and

$$\frac{(M_0 - \hat{G}_0')^2}{\hat{G}_0'} \leq \mathcal{O}\left(\frac{\log^6 n}{n}\right).$$

The Empirical Estimator

For $\Phi''_{\mu-1} \leq 2c_5(\log^2 n)$ and $\mu \geq 2$, we use the empirical estimator,

$$\hat{\phi}_{\mu-1} := \frac{\mu-1}{n}\Phi_{\mu-1}.$$

By Lemma 16, since $\Phi''_{\mu-1} \leq 2c_5(\log^2 n)$, we can assume that $\mathbb{E}[\Phi_{\mu-1}] \leq \mathcal{O}(\log^2 n)$ and $\Phi_{\mu-1} \leq \mathcal{O}(\log^2 n)$, and be correct with probability at least $1 - \mathcal{O}(n^{-10})$.

The following lemma in [7] characterizes the performance of $\hat{\phi}_{\mu-1}$ in estimating $M_{\mu-1}$.

Lemma 19. *For $\mu \geq 2$ and $\delta \in (0,1)$,*

$$\Pr\left(|M_{\mu-1} - \hat{\phi}_{\mu-1}| \leq \mathcal{O}\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log\frac{n}{\delta}}{n}\right)\right) \geq 1-\delta.$$

Setting $\delta = n^{-10}$ in the lemma implies that with probability at least $1 - n^{-10}$,

$$|M_{\mu-1} - \hat{\phi}_{\mu-1}| \leq \mathcal{O}\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log n}{n}\right) \leq \mathcal{O}\left(\frac{\sqrt{\Phi_{\mu-1}(\mu-1)}\log^2 n}{n}\right).$$

Assume that all the inequalities above hold. Then,

$$\frac{(M_{\mu-1} - \hat{\phi}_{\mu-1})^2}{\hat{\phi}_{\mu-1}} \leq \mathcal{O}\left(\frac{\left(\Phi_{\mu-1}\frac{\sqrt{\mu}\log n}{n}\right)^2}{\frac{\mu-1}{n}\Phi_{\mu-1}}\right) = \mathcal{O}\left(\frac{\Phi_{\mu-1}\log^2 n}{n}\right) = \mathcal{O}\left(\frac{\log^4 n}{n}\mathbb{1}_{\Phi_{\mu-1}>0}\right).$$

As a final remark, we can choose $c_2 = 2c_5$.

Final Estimator

In case our estimates sum to 1, we can simply estimate each M_μ by

$$\hat{M}_\mu := \begin{cases} \hat{G}'_\mu & \text{if } \mu = 0, \\ \hat{\phi}_\mu & \text{if } \mu \geq 1 \text{ and } \Phi_\mu \leq c_2(\log^2 n), \\ \hat{O}'_\mu & \text{if } \mu > c_3 \log n \text{ and } \Phi_\mu > c_2(\log^2 n), \\ \hat{G}'_\mu & \text{if } c_3 \log n \geq \mu \geq 1 \text{ and } \Phi_\mu > c_2(\log^2 n), \end{cases}$$

Otherwise, we normalize these probability estimates by their sum,

$$T := \sum_{\mu \geq 0} \hat{M}_\mu,$$

and approximate each M_μ by $\hat{M}^*_\mu := \hat{M}_\mu / T$.

First we show that T is often close to 1. By Lemma 2, under Poisson sampling,

$$\Pr\left(1 \leq \sum_{\mu \geq 1} \Phi_\mu \mu = \mathrm{Poi}(n) \leq n \log n\right) \geq 1 - \mathcal{O}(e^{-n}).$$

By the union bound and results in the previous sections, with probability at least $1 - \mathcal{O}(n^{-8})$,

$$|M_\mu - \hat{M}_\mu| \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_\mu \mu}}{n}\right), \forall \mu \geq 1,$$

$$\frac{(M_\mu - \hat{M}_\mu)^2}{\hat{M}_\mu} \leq \tilde{\mathcal{O}}\left(\frac{\mathbb{1}_{\Phi_\mu > 0}}{n}\right), \forall \mu \geq 1,$$

$$|M_0 - \hat{M}_0| \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_1} + 1}{n}\right),$$

and

$$\frac{(M_0 - \hat{M}_0)^2}{\hat{M}_0} \leq \tilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

These inequalities further imply that with probability at least $1 - \mathcal{O}(n^{-8})$,

$$\begin{aligned} |T - 1| &\leq |\hat{M}_0 - M_0| + \sum_{\mu \geq 1} |\hat{M}_\mu - M_\mu| \\ &\leq \tilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_1} + 1}{n}\right) + \sum_{\mu \geq 1} \tilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_\mu \mu}}{n}\right) \\ &= \sum_{\mu \geq 0} \tilde{\mathcal{O}}\left(\frac{\sqrt{\Phi_\mu \mu}}{n}\right) \\ &\leq \tilde{\mathcal{O}}\left(\sqrt{\frac{\sum_{\mu \geq 1} \mathbb{1}_{\Phi_\mu > 0}}{n}}\right), \end{aligned}$$

where the second inequality follows from $\sum_{\mu \geq 1} \mu \Phi_\mu < n \log n$ and the Cauchy-Schwarz inequality.

To characterize the performance of estimator $\hat{M}^* := \{\hat{M}^*_\mu\}_{\mu\geq 0}$, we bound the KL-divergence by the χ-squared distance. By the above inequalities, with probability at least $1 - \mathcal{O}(n^{-8})$,

$$\begin{aligned}
\sum_{\mu\geq 0} M_u \log \frac{M_\mu}{\hat{M}^*_\mu} &\leq \sum_{\mu\geq 0} \frac{(M_\mu - \hat{M}^*_\mu)^2}{\hat{M}^*_\mu} \\
&\leq 2(T-1)^2 + \sum_{\mu\geq 0} 2T \frac{(M_\mu - \hat{M}_\mu)^2}{\hat{M}_\mu} \\
&\leq \tilde{\mathcal{O}}\left(\frac{\sum_{\mu\geq 1} \mathbb{1}_{\Phi_\mu>0}}{n}\right) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right) + \sum_{\mu\geq 1} \tilde{\mathcal{O}}\left(\frac{\mathbb{1}_{\Phi_\mu>0}}{n}\right) \\
&= \tilde{\mathcal{O}}\left(\frac{\sum_{\mu\geq 1} \mathbb{1}_{\Phi_\mu>0}}{n}\right) \\
&= \tilde{\mathcal{O}}\left(\frac{D_\Phi}{n}\right).
\end{aligned}$$

Finally, for each $x \in [k]$, define our probability estimate by

$$\hat{p}^*_x(X^n) = \frac{\hat{M}^*_{N_x}}{\Phi_{N_x}}.$$

The following identity [1] completes the proof of Theorem 1.

$$\tilde{\ell}_{X^n}(p, \hat{p}^*) = \sum_{\mu\geq 0} M_\mu \log \frac{M_\mu}{\hat{M}^*_\mu}.$$

4 Proof of Corollary 9

We begin with a lemma that partially characterizes the shape of a log-concave distribution.

Lemma 20. *[8] Let p be a log-concave distribution with mean μ_p and standard deviation σ_p. Let $\alpha, \beta \in [k]$ be integers satisfying $\alpha \leq \mu_p - \Omega(\sigma_p(1+\log(1/\varepsilon)))$ and $\beta \geq \mu_p + \Omega(\sigma_p(1+\log(1/\varepsilon)))$. Then,*

$$\sum_{x=1}^{\alpha} p_x + \sum_{x=\beta}^{k} p_x \leq 2\varepsilon.$$

In addition, for σ_p larger than an absolute constant, the maximum probability satisfies

$$\max_{x\in[k]} p_x \in [1/(8\sigma_p), 1/\sigma_p].$$

Setting $\varepsilon = 1/n^5$ in the above lemma, we obtain

$$\begin{aligned}
\Pr(D_\Phi > \Omega(\log(n^5)\sigma)) &\leq \Pr(D > \Omega(\log(n^5)\sigma)) \\
&\leq \Pr(\exists x, s.t.\ x \notin (\alpha, \beta), N_x \geq 1) \\
&\leq \sum_{x=1}^{\alpha} np_x + \sum_{x=\beta}^{k} np_x \\
&\leq 2 \cdot n^{-4}.
\end{aligned}$$

Therefore, $\mathbb{E}[D_\Phi] \leq \mathcal{O}\left(\log(n^5)\sigma\right)$. Now, we use the second part of Lemma 20 to derive a different upper bound on $\mathbb{E}[D_\Phi]$. Let $j_{\max}$ be the index such that $\max_{x\in[k]} p_x \in I_{j_{\max}}$.

$$(j_{\max} - 1)^2 \frac{\log n}{n} < \max_{x\in[k]} p_x < \frac{\log n}{\sigma}.$$

The above inequality implies $j_{\max} < \sqrt{2n/\sigma} + 1$. Using the same reasoning as in Section 4.2 in the main paper, we get

$$\mathbb{E}[D_\Phi] \leq \mathcal{O}\left((\sqrt{n/\sigma})^{\frac{2}{3}} \cdot n^{\frac{1}{3}}\right) \cdot \log n = \tilde{\mathcal{O}}\left((\sigma n)^{-\frac{1}{3}}\right).$$

Combining the above two upper bounds on $\mathbb{E}[D_\Phi]$ yields

Corollary 1. *For any distribution $p \in \mathcal{L}_k^{n,\sigma}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{\sigma}{n}\right).$$

5 Proof of Corollary 11

Consider the collection $\mathcal{P}_k^{\alpha,c} := \{p \in \Delta_k : p_x \le c \cdot x^{-\alpha}\}$ of enveloped (truncated) power-law distributions. Note that this definition generalizes power-law families, and that distributions in $\mathcal{P}_k^{\alpha,c}$ are not necessarily log-convex. Let $\beta \in (0,1)$ be a parameter to be determined, and x_0 be the threshold such that $2n(c \cdot x_0^{-\alpha}) = n^\beta$. The symbols $x \in [k]$ that are no larger than x_0 contribute at most x_0 to D_Φ. On the other hand, for any $x > x_0$, we have $\mathbb{E}[N_x] = np_x \le n(c \cdot x^{-\alpha}) < 0.5n^\beta$. Therefore, for $x > x_0$,

$$\begin{aligned}
\Pr(N_x > 2n^\beta) &\le \frac{np_x}{1-p_x} \Pr(N_x \ge 2n^\beta) \\
&\le 2np_x \Pr\left(N_x \ge \mathbb{E}[N_x] + n^\beta\right) \\
&\le 2np_x \exp\left(-n^\beta/3\right),
\end{aligned}$$

where the first inequality follows from direct comparison and the last follows from the Chernoff bound for binomial random variables. By the union bound,

$$\begin{aligned}
\Pr(\exists x \in [k] \; s.t. \; N_x > 2n^\beta) &\le \sum_{x \in [k]} \Pr(N_x > 2n^\beta) \\
&\le 2n \exp\left(-n^\beta/3\right).
\end{aligned}$$

Therefore, with probability at least $1 - 2n\exp\left(-n^\beta/3\right)$,

$$D_\Phi \le x_0 + 2n^\beta = (2c)^{\frac{1}{\alpha}} n^{\frac{1-\beta}{\alpha}} + 2n^\beta.$$

Optimizing the right-hand side by choosing $\beta = 1/(\alpha+1)$, the inequality simplifies to

$$D_\Phi \le ((2c)^{\frac{1}{\alpha}} + 2) n^{\frac{1}{\alpha+1}}.$$

Since $D_\Phi \le n$, we can convert this high-probability result into the expectation bound,

$$\mathbb{E}[D_\Phi] \le \mathcal{O}(n^{\frac{1}{\alpha+1}}).$$

Along with Corollary 6 in the main paper this implies

Corollary 2. *For any distribution $p \in \mathcal{P}_k^{\alpha,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

References

[1] A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

[2] R. Krichevsky and V. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

[3] D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[4] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

[5] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[6] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pages 764–796, 2013.

[7] E. Drukh and Y. Mansour. Concentration bounds for unigrams language model. In *International Conference on Computational Learning Theory*, pages 170–185, 2004.

[8] I. Diakonikolas, D. M. Kane, and A. Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.