JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–33

Optimal Probability Estimation with Applications to Prediction and Classification

Jayadev Acharya JACHARYA@UCSD.EDU
Ashkan Jafarpour ASHKAN@UCSD.EDU
Alon Orlitsky ALON@UCSD.EDU
Ananda Theertha Suresh ASURESH@UCSD.EDU
University of California, San Diego

Abstract

Via a unified view of probability estimation, classification, and prediction, we derive a uniformly-optimal combined-probability estimator, construct a classifier that uniformly approaches the error of the best possible label-invariant classifier, and improve existing results on pattern prediction and compression.

Keywords: Good-Turing estimators, convergence rates, linear estimators, competitive classification

1. Introduction

Probability estimation, prediction, and classification, are at the core of statistics, information theory, and machine learning. Using a unified approach, we derive several results on these three related problems.

Let M_μ denote the combined probability of all elements appearing $\mu \in \{0, \ldots, n\}$ times in n independent samples of a discrete distribution p. Building on the basic empirical estimator, the classical Good-Turing estimator in Good (1953), and their combinations, McAllester and Schapire (2000) and Drukh and Mansour (2004) derived estimators that approximate M_μ to within $\widetilde{\mathcal{O}}(n^{-0.4})$, where this and all subsequent bounds hold with probability close to one and apply uniformly to all distributions p regardless of their support size and probability values. These estimators can be extended to approximate $M^n \overset{\text{def}}{=} (M_0, \ldots, M_n)$ to within ℓ_1 distance of $\widetilde{\mathcal{O}}(n^{-1/6})$. In this paper, we:

1. Show that the above estimators perform best among all simple combinations of empirical and Good-Turing estimators in that for some distributions, any simple combination of these two estimators aimed at approximating M^n, will incur ℓ_1 distance $\widetilde{\Omega}(n^{-1/6})$.

2. Derive a linear-complexity estimator that approximates M^n to ℓ_1 distance $\widetilde{\mathcal{O}}(n^{-1/4})$ and KL-divergence $\widetilde{\mathcal{O}}(n^{-1/2})$, and prove that this performance is optimal in that any estimator for M^n must incur at least these distances for some underlying distribution.

3. Apply this estimator to sequential prediction and compression of patterns, deriving a linear-complexity algorithm with per symbol expected redundancy $\widetilde{\mathcal{O}}(n^{-1/2})$, improving the previously-known $\mathcal{O}(n^{-1/3})$ bound for any polynomial-complexity algorithm.

4. Modify the estimator to derive a linear-complexity classifier that takes two length-n training sequences, one distributed *i.i.d.* according to a distribution p and one according to q, and classifies a single test sample generated by p or q, with error at most $\widetilde{\mathcal{O}}(n^{-1/5})$ higher than that achievable by the best label-invariant classifier designed with knowledge of p and q, and show an $\widetilde{\Omega}(n^{-1/3})$ lower bound on this additional error for any classifier.

The paper is organized as follows. Sections 2, 3, and 4, address probability estimation, prediction, and classification, respectively, providing a more comprehensive background, precise definitions, and detailed results for each problem. Section 5 outlines some of the analysis involved. For space considerations, the proof are relegated to the Appendix.

2. Probability estimation

2.1. Background

A *probability distribution* over a discrete set $\mathcal{X}$ is a mapping $p : \mathcal{X} \to [0,1]$ such that $\sum_{x\in\mathcal{X}} p_x = 1$. Let $\mathcal{D}_{\mathcal{X}}$ denote the collection of all distributions over $\mathcal{X}$. We study the problem of estimating an unknown distribution $p \in \mathcal{D}_{\mathcal{X}}$ from a sample $X^n \stackrel{\text{def}}{=} X_1,\ldots,X_n$ of n random variables, each drawn independently according to p. A *probability estimator* is a mapping $q : \mathcal{X}^n \to \mathcal{D}_{\mathcal{X}}$ associating a distribution $q \stackrel{\text{def}}{=} q(x^n) \in \mathcal{D}_{\mathcal{X}}$ with every sample x^n.

For any distribution p over a finite support $\mathcal{X}$, given sufficiently many samples, many reasonable estimators, will eventually estimate p well. Take for example the *empirical-frequency-estimator* E that associates with every symbol the proportion of times it appeared in the observed sample. For example, given $\mathcal{O}\left(\frac{|\mathcal{X}|}{\delta^{2.1}}\log\frac{1}{\epsilon}\right)$ samples, a number linear in the distribution's support size $|\mathcal{X}|$, E estimates p to within ℓ_1 distance δ with probability $\geq 1-\epsilon$ (see, *e.g.,* Das, 2012). Paninski (2004) proved that no estimator can estimate all distributions over $\mathcal{X}$ using $o(|\mathcal{X}|)$ samples. Orlitsky et al. (2005) showed that not only p, but even just the probability multiset $\{p(x) : x \in \mathcal{X}\}$ cannot be uniformly ℓ_1-estimated, and Valiant and Valiant (2011) proved that estimating the probability multiset with earth-mover distance ≤ 0.25 requires $\Omega\left(\frac{|\mathcal{X}|}{\log|\mathcal{X}|}\right)$ samples.

Estimation that requires a number of samples proportional or nearly-proportional to the distribution's support size suffers several drawbacks. Some common distributions, such as Poisson and Zipf have infinite support size. Many practical problems, for example those involving natural-language processing or genomics, have very large support sizes (the sets of possible words or nucleotide locations). Additionally, in many cases the alphabet size is unknown, hence no bounds can derived on the estimation error.

For these and related reasons, a number of researchers have recently considered distribution properties that can be estimated *uniformly.* A *uniform bound* is one that applies to all distributions p regardless of the support set $\mathcal{X}$. As we saw, p itself cannot be uniformly estimated. Intuitively, the closer a property is to p, the harder it is to uniformly approximate. It is perhaps surprising therefore that a slight modification of p, that as we shall see is sufficient for many applications, can be uniformly approximated.

Good (1953) noted that reasonable estimators assign the same probability to all symbols appearing the same number of times in a sample. For example, in the sample $b, a, n, a, n, a, s,$

the same probability to b and s. The performance of such estimators is determined by the combined probability they assign to all symbols appearing any given number of times, namely by how well they estimate the *combined probability*, or *mass*,

$$M_\mu \overset{\text{def}}{=} \sum_{x:\mu_x=\mu} p_x$$

of symbols appearing $\mu \in \{0,1,\ldots,n\}$ times, where μ_x is the number of times a symbol $x \in \mathcal{X}$ appears in the sample x^n. Let $\mathbb{1}_x^\mu$ be the indicator function that is 1 iff $\mu_x = \mu$.

2.2. Previous results

Let Φ_μ denote the number of symbols appearing μ times in a sample of size n. Empirical frequency estimates M_μ by

$$E_\mu = \frac{\mu}{n} \cdot \Phi_\mu .$$

The Good-Turing estimator in Good (1953), estimates M_μ by

$$G_\mu \overset{\text{def}}{=} \frac{\mu+1}{n} \cdot \Phi_{\mu+1} . \tag{1}$$

The Good-Turing estimator is an important tool in a number of language processing applications, (*e.g.,* Chen and Goodman, 1996). However for several decades it defied rigorous analysis, partly because of the dependencies between μ_x for different x's. First theoretical results were provided by McAllester and Schapire (2000). Using McDiarmid's inequality (McDiarmid, 1989), they showed that for all $0 \le \mu \le n$, with probability $\ge 1-\delta$,

$$|G_\mu - M_\mu| = \mathcal{O}\left(\sqrt{\frac{\log(3/\delta)}{n}}\left(\mu + 1 + \log\frac{n}{\delta}\right)\right).$$

Note that this bound, like all subsequent ones in this paper, holds uniformly, namely applies to all support sets $\mathcal{X}$ and all distributions $p \in \mathcal{D}_{\mathcal{X}}$.

To express this and subsequent results more succinctly, we will use several abbreviations. $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ will be used to hide poly-logarithmic factors in n and $1/\delta$, and for a random variable X, we will use

$$X \underset{\delta}{=} \widetilde{\mathcal{O}}(\alpha) \quad \text{to abbreviate} \quad \Pr\left(X \ne \widetilde{\mathcal{O}}(\alpha)\right) \le \delta,$$

and similarly $X \underset{\delta}{=} \widetilde{\Omega}(\alpha)$ for $\Pr\left(X \ne \widetilde{\Omega}(\alpha)\right) < \delta$. For example, the above bound becomes

$$|G_\mu - M_\mu| \underset{\delta}{=} \widetilde{\mathcal{O}}\left(\frac{\mu+1}{\sqrt{n}}\right).$$

As could be expected, most applications require simultaneous approximation of M_μ over a wide range of μ's. For example, as shown in Section 4, classification requires approximating $M^n \overset{\text{def}}{=} (M_0, \ldots, M_n)$ to within a small ℓ_1 distance, while prediction requires approximation to within a small KL-Divergence.

Drukh and Mansour (2004) improved the Good-Turing bound and combined it with the empirical estimator to obtain an estimator G' with ℓ_∞ convergence,

$$||G'^n - M^n||_\infty \overset{\text{def}}{=} \max_{0\le\mu\le n} |G'_\mu - M_\mu| \underset{\delta}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{0.4}}\right),$$

where $G'^n \overset{\text{def}}{=} (G'_0, \ldots, G'_n)$. With some more work one can extend their results to the more practically useful ℓ_1 convergence,

$$||G'^n - M^n||_1 \overset{\text{def}}{=} \sum_{\mu=0}^{n} |G'_\mu - M_\mu| \underset{\delta}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/6}}\right). \tag{2}$$

Subsequently, Wagner et al. (2006) considered ℓ_1 convergence for a subclass of distributions where all symbols probabilities are proportional to $1/n$, namely for some constants c_1, c_2, all probabilities p_x are in the range $[c_1/n, c_2/n]$. Recently, Ohannessian and Dahleh (2012) showed that the Good-Turing estimator is not uniformly multiplicatively consistent over all distributions, and described a class of distributions for which it is.

2.3. New results

First we show that upper bound (2) and the one in Drukh and Mansour (2004) are tight in that no simple combination of G_μ and the empirical estimator E_μ can approximate M_μ better. A proof sketch is provided in Appendix B.11.

Lemma 1 *For every n, there is a distribution such that*

$$\sum_{\mu=0}^{n} \min\left(|E_\mu - M_\mu|, |G_\mu - M_\mu|\right) \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/6}}\right).$$

In Subsections 5.3–5.5, we construct a new estimator F'_μ and show that it estimates M_μ better than G'_μ and essentially as well as any other estimator. A closer inspection of Good and Turing's intuition in Gale and Sampson (1995) shows that the average probability of a symbol appearing μ times is

$$\frac{M_\mu}{\Phi_\mu} \approx \frac{\mu+1}{n} \cdot \frac{\mathbb{E}[\Phi_{\mu+1}]}{E[\Phi_\mu]}. \tag{3}$$

If we were given the values of the $\mathbb{E}[\Phi_\mu]$'s, we could use this equation to estimate the M_μ's. Since we are not given these values, Good-Turing (1) approximates the expectation ratio by just $\Phi_{\mu+1}/\Phi_\mu$. However, while Φ_μ and $\Phi_{\mu+1}$ are by definition unbiased estimators of their expectations $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu+1}]$ respectively, their variance is high, leading to a probability estimation G_μ that may be far from M_μ.

In Section 5.4 we smooth the estimate of $\mathbb{E}[\Phi_\mu]$ by expressing it as a linear combination of the values of $\Phi_{\mu'}$ for μ' near μ. Lemma 15 shows that an appropriate choice of the smoothing coefficients yields an estimate $\widehat{\mathbb{E}[\Phi_\mu]}$ that approximates $\mathbb{E}[\Phi_\mu]$ well.

Incorporating this estimate into Equation (3), yields a new estimator F_μ. Combining it with the empirical and Good-Turing estimators for different ranges of μ and Φ_μ, we obtain a modified estimator F'_μ that has a small KL divergence from M_μ, and hence by Pinsker's inequality, also small ℓ_1 distance uniformly over all distributions.

Theorem 2 *For every distribution and every n, $F'^n \overset{\text{def}}{=} (F'_0, \ldots, F'_n)$ satisfies,*

$$D(M^n||F'^n) \overset{\text{def}}{=} \sum_{\mu=0}^{n} M_\mu \log\left(\frac{M_\mu}{F'_\mu}\right) \underset{1/n}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/2}}\right) \text{ and } ||F'^n - M^n||_1 \underset{1/n}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/4}}\right).$$

In Section 5.7 we show that the proposed estimator is optimal. An estimator is *label-invariant*, often called *canonical*, if its estimate of M_μ remains unchanged under all permutations of the symbol labels. For example, its estimate of M_1 will be the same for the sample a, a, b, b, c as it is for u, u, v, v, w. Clearly all reasonable estimators are label-invariant.

Theorem 3 *For any label-invariant estimator $\widehat{M}$, there is a distribution such that*

$$D(M^n||\widehat{M}^n) \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right) \text{ and } ||\widehat{M}^n - M^n||_1 \underset{1/n}{=} \widetilde{\Omega}\left(\frac{1}{n^{1/4}}\right).$$

Finally we note that the estimator F'_μ can be computed in time linear in n. Also, observe that while the difference between ℓ_1 distance of $1/n^{1/6}$ and $1/n^{1/4}$ may seem small, an equivalent formulation of the results would ask for the number of samples needed to estimate within a ℓ_1 distance ϵ. Good-Turing and empirical frequency would require $(1/\epsilon)^6$ samples, while the estimator we construct needs $(1/\epsilon)^4$ samples. For $\epsilon = 1\%$, the difference between the two is a factor of 10,000.

3. Prediction

3.1. Background

Probability estimation can be naturally applied to prediction and compression. Upon observing a sequence $X^i \overset{\text{def}}{=} X_1, \ldots, X_i$ generated *i.i.d.* according to some distribution $p \in \mathcal{D}_{\mathcal{X}}$, we would like to form an estimate $q(x|x^i)$ of $p(x)$ to minimize a cumulative loss $\sum_{i=1}^{n} f_p(q(X_{i+1}|X^i), X_{i+1})$ (see for example Vovk, 1995; Merhav and Feder, 1998).

The most commonly used loss is *log-loss*, $f_p(q(x_{i+1}|x^i), x_{i+1}) = \log(q(x_{i+1}|x^i)/p(x_{i+1}))$. Its numerous applications include compression, (*e.g.,* Rissanen, 1984), MDL principle, (*e.g.,* Grünwald, 2007), and learning theory, (*e.g.,* Cesa-Bianchi and Lugosi, 1999). Its expected value is the KL-divergence between the underlying distribution p and the prediction q.

Again we consider label-invariant predictors that use only ordering and frequency of symbols, not the specific labels. Following Orlitsky et al. (2003), after observing n samples, we assign probability to each of the previously-observed symbols, and to observing a new symbol new. For example, if after three samples, the sequence observed is aba, we assign the probabilities $q(a|aba)$, $q(b|aba)$, and $q(\text{new}|\text{aba})$ that reflects the probability at which we think a symbol other than a or b will appear. These three probabilities must add to 1. Furthermore, if the sequence is bcb, then the probability we assign to b must be the same as the probability we previously assigned to a.

Equivalently, Orlitsky et al. (2003) defined the *pattern* of a sequence to be the sequence of integers, where the i^{th} new symbol appearing in the original sequence is replaced by the integer i. For example, the pattern of aba is 121. We use Ψ^n and to denote a length-n pattern, and Ψ_i to denote its ith element.

The prediction problem is now that of estimating $\Pr(\Psi_{n+1}|\Psi^n)$, where if Ψ^n consists of m distinct symbols then the distribution is over $[m+1]$, and $m+1$ reflects a new symbol. For example, after observing 121, we assign probabilities to 1, 2, and 3.

3.2. Previous results

Orlitsky et al. (2003) proved that the Good-Turing estimator achieves constant per-symbol worst-case log-loss, and constructed two sequential estimators with diminishing worst-case log-loss: a computationally efficient estimator with log-loss $\mathcal{O}(n^{-1/3})$, and a high complexity estimator with log-loss $\mathcal{O}(n^{-1/2})$. Orlitsky et al. (2004) constructed a low-complexity block estimator for patterns with worst-case per-symbol log-loss of $\mathcal{O}(n^{-1/2})$. For expected log-loss, Shamir (2004) improved this bound to $\mathcal{O}(n^{-3/5})$ and Acharya et al. (2012b) further improved it to $\widetilde{\mathcal{O}}(n^{-2/3})$, but their estimators are computationally inefficient.

3.3. New results

Using Theorem 2, we obtain a computationally efficient predictor q that achieves expected log-loss of $\widetilde{\mathcal{O}}(n^{-1/2})$. Let F'_μ be the estimator proposed in Section 5.5. Let $q(\Psi_{n+1}|\Psi^n) = \frac{F'_\mu}{\Phi_\mu}$ if Ψ_{n+1} appears μ times in Ψ^n, and F'_0, if it is Ψ_{n+1} is a new symbol. The following corollary, proved in Appendix C, bounds the predictor's performance.

Corollary 4 *For every distribution p,*

$$\mathbb{E}_p[D(p(\Psi_{n+1}|\Psi^n)||q(\Psi_{n+1}|\Psi^n)] = \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/2}}\right).$$

4. Classification

4.1. Background

Classification is one of the most studied problems in machine learning and statistics (Boucheron et al., 2005). Given two *training* sequences X^n and Y^n, drawn *i.i.d.* according to two distributions p and q respectively, we would like to associate a new *test* sequence Z^m drawn *i.i.d.* according to one of p and q with the training sequence that was generated by the same distribution.

It can be argued that natural classification algorithms are *label invariant*, namely, their decisions remain the same under all one-one symbol relabellings, (*e.g.,* Batu, 2001). For example, if given training sequences `abb` and `cbc`, and a classifier associates `b` with `abb`, then given `utt` and `gtg`, it must associate `t` with `utt`.

Our objective is to derive a *competitive* classifier whose error is close to the best possible by any label-invariant classifier, uniformly over all (p, q). Namely, a single classifier whose error probability differs from that of the best classifier for the given (p, q) by a quantity that diminishes to 0 at a rate determined by the sample size n alone, and is independent of p and q.

4.2. Previous results

A number of classifiers have been studied in the past, including the likelihood-ratio, generalized-likelihood, and Chi-Square tests. However while they perform well when the number of samples is large, none of them is uniformly competitive with all label-invariant classifiers.

When $m = \Theta(n)$, classification can be related to the problem of *closeness testing* that asks whether two sequences X^n and Y^n are generated by the same or different distributions. Over the last decade, closeness testing has been considered by a number of researchers. Batu et al. (2000) showed that testing if the distributions generating X^n and Y^n are identical or are at least δ apart in ℓ_1 distance requires $n = \widetilde{\mathcal{O}}(k^{2/3})$ samples where the constant depends on δ. Acharya et al. (2011) took a competitive view of closeness testing and derived a test whose error is $\leq \epsilon e^{\mathcal{O}(n^{2/3})}$ where ϵ is the error of the best label-invariant protocol for this problem, designed in general with knowledge of p and q.

Their result shows that if the optimal closeness test requires n samples to achieve an error $\leq \epsilon$, then the proposed test achieves the same error with $\widetilde{\mathcal{O}}(n^3)$ samples. Acharya et al. (2012a) improved it to $\widetilde{\mathcal{O}}(n^{3/2})$ and proved a lower bound of $\widetilde{\Omega}(n^{7/6})$ samples.

4.3. New results

We consider the case where $m = 1$, namely the test data is a single sample. Many machine-learning problems are defined in this regime, for example, we are given the DNA sequences of several individuals and need to decide whether or not they are susceptible to a certain disease (*e.g.*, Braga-Neto, 2009).

It may seem that when $m = 1$, the best classifier is a simple majority classifier that associates Z with the sequence X^n or Y^n where Z appears more times. Perhaps surprisingly, the next example shows that this is not the case.

Example 5 *Let $p = U[n]$ and $q = U[2n]$ be the uniform distributions over $\{1, \ldots, n\}$ and $\{1, \ldots, 2n\}$, and let the test symbol Z be generated according to $U[n]$ or $U[2n]$ with equal probability. We show that the empirical classifier, that associates Z with the sample in which it appeared more times, entails a constant additional error more than the best achievable.*

The probability that Z appears in both X^n and Y^n is a constant. And in all these cases, the optimal label-invariant test that knows p and q assigns Z to $U[n]$, namely X^n, because $p(Z) = 1/n > 1/2n = q(Z)$. However, with constant probability, Z appears more times in Y^n than in X^n, and then the empirical classifier associates Z with the wrong training sample, incurring a constant error above that of the optimal classifier.

Using probability-estimation techniques, we derive a uniformly competitive classifier. Before stating our results we formally define the quantities involved. Recall that $X^n \sim p$ and $Y^n \sim q$. A classifier S is a mapping $S : \mathcal{X}^* \times \mathcal{X}^* \times \mathcal{X} \to \{\mathtt{x}, \mathtt{y}\}$, where $S(\overline{x}, \overline{y}, z)$ indicates whether z is generated by the same distribution as $\overline{x}$ or $\overline{y}$. For simplicity we assume that $Z \sim p$ or q with equal probability, but this assumption can be easily relaxed. The error probability of a classifier S with n samples is

$$\mathcal{E}^S_{p,q}(n) = \frac{1}{2} \Pr\left(S(X^n, Y^n, Z) = \mathtt{y} | Z \sim p\right) + \frac{1}{2} \Pr\left(S(X^n, Y^n, Z) = \mathtt{x} | Z \sim q\right).$$

Let $\mathcal{S}$ be the collection of label-invariant classifiers. For every p, q, let $\mathcal{E}^{S_{p,q}}_{p,q}(n) = \min_{S \in \mathcal{S}} \mathcal{E}^{S}_{p,q}(n)$ be the lowest error achieved for (p, q) by any label-invariant classifier, where the classifier $S_{p,q}$ achieving $\mathcal{E}^{S_{p,q}}_{p,q}(n)$ is typically designed with prior knowledge of (p, q).

We construct a linear-time label-invariant classifier S whose error is close to $\mathcal{E}^{S_{p,q}}_{p,q}(n)$. We first extend the ideas developed in the previous section to pairs of sequences and develop an estimator $F'^{p}_{\mu,\mu'}$, and then use this estimator to construct a classifier whose extra error is $\widetilde{\mathcal{O}}(n^{-1/5})$.

Theorem 6 *For all (p, q), there exists a classifier S such that*

$$\mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/5}}\right).$$

In Appendix D we state the classifier that has extra error $\widetilde{\mathcal{O}}(n^{-1/5})$ and prove Theorem 6. In Appendix D.6 we also provide a non-tight lower bound for the problem and show that for any classifier S, there exist (p, q), such that $\mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}\left(n^{-1/3}\right)$.

5. Analysis of probability estimation

We now outline proofs of Lemma 1 and Theorems 2 and 3. In Section 5.1 we introduce Poisson sampling, a useful technique for removing the dependencies between multiplicities. In Section 5.2, we state some limitations of empirical and Good-Turing estimators, and use an example to motivate Lemma 1. In Section 5.3 we motivate the proposed estimator via an intermediate *genie-aided estimator*. In Section 5.5 we propose the new estimator. In Section 5.6 we sketch the proof of Theorem 2. In Section 5.7, we sketch the proof of Theorem 3 providing lower bounds on estimation.

5.1. Poisson sampling

In the standard sampling method, where a distribution is sampled n times, the multiplicities are dependent. Analysis of functions of dependent random variables requires various concentration inequalities, which often complicates the proofs. A useful approach to make them independent and hence simplify the analysis is to do Poisson sampling. The distribution is sampled a random n' times, where n' is a Poisson random variable with parameter n.

The following fact, mentioned without proof states that the multiplicities are independent under Poisson sampling.

Fact 7 (**Mitzenmacher and Upfal, 2005**) *If a distribution p is sampled i.i.d.* Poi(n) *times, then the number of times symbol x appears is an independent Poisson random variable with mean np_x, namely,* $\Pr(\mu_x = \mu) = \frac{e^{-np_x}(np_x)^{\mu}}{\mu!}$.

In Appendix B.1 we provide a simple proof for the following lemma, which shows that proving properties for Poi(n) sampling implies properties for sampling the distribution exactly n times. Hence in the rest of the paper, we prove the properties of an estimator under Poisson sampling.

Lemma 8 (**Mitzenmacher and Upfal, 2005**) *If when a distribution is sampled* $\mathrm{Poi}(n)$ *times, a certain property holds with probability* $\geq 1-\delta$*, then when the distribution is sampled exactly* n *times, the property holds with probability* $\geq 1-\delta\cdot e\sqrt{n}$*.*

To illustrate the advantages of Poisson sampling, we first show that Good-Turing estimator is unbiased under Poisson sampling. We use this fact to get a better understanding of the proposed estimator. It is proved in Appendix B.2.

Claim 9 *For every distribution* p *and every* μ*,*

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

5.2. Limitations of Good-Turing and empirical estimators

We first prove an upper bound on the estimation error of Good-Turing and empirical estimators. Proofs for variations of these lemmas are in Drukh and Mansour (2004). We give simple proofs in Appendix B.3 and B.4 using Bernstein's inequality and Chernoff bound.

Lemma 10 (Empirical estimator) *For every distribution* p *and every* $\mu\geq 1$*,*

$$|M_\mu-E_\mu|\underset{\delta}{=}\mathcal{O}\left(\Phi_\mu\frac{\sqrt{\mu+1}\log\frac{n}{\delta}}{n}\right).$$

Lemma 11 (Good-Turing estimator) *For every distribution* p *and every* μ*, if* $\mathbb{E}[\Phi_\mu]\geq 1$*, then*

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

The following example illustrates the tightness of these results.

Example 12 *Let* $U[k]$ *be the uniform distribution over* k *symbols, and let the sample size be* $n\gg k$*. The expected multiplicity of each symbol is* $\frac{n}{k}$*, and by properties of binomial distributions, the multiplicity of any symbol is* $>\frac{n}{k}+\sqrt{\frac{n}{k}}$ *with probability* ≥ 0.1*. Also, for every multiplicity* μ*,* $M_\mu=\Phi_\mu/k$*.*

- *The empirical estimate* $E_\mu=\Phi_\mu\frac{\mu}{n}$*. For* $\mu\geq\frac{n}{k}+\sqrt{\frac{n}{k}}$*, the error is* $\Phi_\mu\sqrt{\frac{1}{nk}}\approx\Phi_\mu\frac{\sqrt{\mu}}{n}$*.*
- *The Good-Turing estimate* $G_\mu=\Phi_{\mu+1}\frac{\mu+1}{n}$ *and it does not depend on* Φ_μ*. Therefore, if two sequences have same* $\Phi_{\mu+1}$*, but different* Φ_μ *then Good-Turing makes an error in at least one of the sequences. It can be shown that, the typical error is* $\sqrt{\mathbb{E}[\Phi_\mu]}\frac{1}{k}\approx\sqrt{\mathbb{E}[\Phi_\mu]}\frac{\mu}{n}$*, as the standard deviation of* Φ_μ *is* $\sqrt{\mathbb{E}[\Phi_\mu]}$*.*

The errors in the above example are very close to the upper bounds in Lemma 10 and Lemma 11. Using a finer analysis and explicitly constructing a distribution, we prove Lemma 1 in Appendix B.11.

5.3. A genie-aided estimator

To motivate the proposed estimator we first describe an intermediate genie-aided estimator. In the next section, we remove the genie assumption. Although by Claim 9 Good-Turing estimator is unbiased, it has a large variance. It does not use the fact that Φ_μ symbols appear μ times, as illustrated in Example 12.

To overcome these limitations, imagine for a short while that a genie gives us the values of $\mathbb{E}[\Phi_\mu]$ for all μ. We can then define the *genie-aided* estimator,

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

We observe few properties of $\widehat{M_\mu}$. By Claim 9 $\mathbb{E}[\widehat{M_\mu}] = \mathbb{E}[G_\mu] = \mathbb{E}[M_\mu]$, and hence $\widehat{M_\mu}$ is an unbiased estimator of M_μ. It is linear in Φ_μ and hence shields against the variance of $\Phi_{\mu+1}$. For a uniform distribution with support size k, it is easy to see that $\widehat{M_\mu} = \Phi_\mu \frac{1}{k} = M_\mu$. For a general distribution, we quantify the error of this estimator in the next lemma, whose proof is given in Appendix B.5.

Lemma 13 (Genie-aided estimator) *For every distribution p and every $\mu \geq 1$, if $\mathbb{E}(\Phi_\mu) \geq 1$, then*

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

Recall that the error of E_μ and G_μ are $\widetilde{\mathcal{O}}\left(\frac{\sqrt{\mu}\Phi_\mu}{n}\right)$ and $\widetilde{\mathcal{O}}\left(\frac{\sqrt{\mathbb{E}[\Phi_{\mu+1}]}\mu}{n}\right)$, respectively. In Appendix A we show that $\mathbb{E}[\Phi_{\mu+1}] = \widetilde{\mathcal{O}}(\mathbb{E}[\Phi_\mu])$. Hence errors of both Good-Turing and empirical estimators are linear in one of μ and Φ_μ and sub-linear in the other. By comparison, the genie-aided estimator achieves the smaller exponent of both estimators, and has smaller error than both. It is advantageous to use such an estimator when both μ and Φ_μ are $\geq$ polylog(n/δ). In the next section, we replace the genie assumption by a good estimate of $\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}$.

5.4. Estimating the ratio of expected values

We now develop estimator for the ratio $\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}$ from the observed sequence. Let $\widehat{\mathbb{E}[\Phi_{\mu+1}]}$, $\widehat{\mathbb{E}[\Phi_\mu]}$ be the estimates of $\mathbb{E}[\Phi_{\mu+1}]$ and $\mathbb{E}[\Phi_\mu]$ respectively. A natural choice for the estimator $\widehat{\mathbb{E}[\Phi_\mu]}$ is a linear estimator of the form $\sum_\mu h_\mu \Phi_\mu$. One can use tools from approximation theory such as Bernstein polynomials (Lorentz, 1986) to find such a linear approximation. However a naive application of these tools is not sufficient, and instead, we exploit properties of Poisson functionals.

If we can approximate $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu+1}]$ to a multiplicative factor of $1 \pm \delta_1$ and $1 \pm \delta_2$, respectively, then a naive combination of the two yields an approximation of the ratio to a multiplicative factor of $1 \pm (|\delta_1| + |\delta_2|)$. However, as is evident from the proofs in Appendix B.7, if we choose different estimators for the numerator and the denominator, we can estimate the ratio accurately. Therefore, the estimates of $\mathbb{E}[\Phi_\mu]$, while calculating M_μ

and $M_{\mu-1}$, are different. For ease of notation we use $\widehat{\mathbb{E}[\Phi_\mu]}$ for both the cases. The usage becomes clear from the context.

We estimate $\mathbb{E}[\Phi_{\mu_0}]$ as a linear combination $\sum_{i=-r}^{r} \gamma_r(i)\Phi_{\mu_0+i}$ of the $2r+1$ nearest Φ_μ's. The coefficients $\gamma_r(i)$ are chosen to minimize to estimator's variance and bias. We show that if $\max_i |\gamma_r(i)|$ is small, then the variance is small, and that for a low bias the coefficients $\gamma_r(i)$ need to be symmetric, namely $\gamma_r(-i) = \gamma_r(i)$, and the following function should be small when $x \sim 1$,

$$B_r(x) \overset{\text{def}}{=} \gamma_r(0) + \sum_{i=1}^{r} \gamma_r(i)\left(x^i + x^{-i}\right) - 1.$$

To satisfy these requirements, we choose the coefficients according to the polynomial

$$\gamma_r(i) = \frac{r^2 - r\alpha_r|i| - \beta_r i^2}{r^2 + 2\sum_{j=1}^{r}(r^2 - r\alpha_r|j| - \beta_r j^2)},$$

where α_r and β_r are chosen so that $\sum_{i=1}^{r} \gamma_r(i)i^2 = 0$ and $\gamma_r(r) = 0$.

The next lemma bounds $B_r(x)$ for the estimator with co-efficients γ_r and is used to prove that the bias of the proposed estimator is small. It is proved in Appendix B.6.

Lemma 14 *If* $r|(x-1)| \le \min(1, x)$, *then*

$$|B_r(x)| = \mathcal{O}(r(x-1))^4.$$

The estimators for $\mathbb{E}[\Phi_{\mu_0}]$ and $\mathbb{E}[\Phi_{\mu_0+1}]$ are as follows. Let $r_{\mu_0} = \left\lfloor \frac{\sqrt{\mu_0}}{\log n(\Phi_{\mu_0}\sqrt{\mu_0})^{1/11}} \right\rfloor$. Let $\mathcal{S}_r^{\mu_0} = \{\mu \mid |\mu - \mu_0| \le r\}$. Then,

$$\widehat{\mathbb{E}[\Phi_{\mu_0+1}]} = \sum_{\mu \in \mathcal{S}_{r_{\mu_0}}^{\mu_0}} \gamma_{r_{\mu_0}}(|\mu_0 + 1 - \mu|)\frac{\mu_0 a_\mu^{\mu_0}}{\mu_0+1}\Phi_\mu,$$

$$\widehat{\mathbb{E}[\Phi_{\mu_0}]} = \sum_{\mu \in \mathcal{S}_{r_{\mu_0}}^{\mu_0+1}} \gamma_{r_{\mu_0}}(|\mu_0 - \mu|)a_\mu^{\mu_0}\Phi_\mu.$$

where, $a_\mu^{\mu_0} = \frac{\mu!}{\mu_0!}\mu_0^{\mu_0-\mu}$ and is used for simplifying the analysis. Note that $\widehat{\mathbb{E}[\Phi_\mu]}$ used to calculate M_μ and $M_{\mu-1}$ are different. r_{μ_0} is chosen to minimize the bias variance trade-off. The following lemma quantifies the quality of approximation of the ratio of $\mathbb{E}[\Phi_{\mu_0+1}]$ and $\mathbb{E}[\Phi_{\mu_0}]$. The proof is involved and uses Lemma 14. It is given in Appendix B.7.

Lemma 15 *For every distribution* p, *if* $\mu_0 \ge \log^2 n$ *and* $\frac{1}{\log n}\left(\frac{\mu_0}{\log^2 n}\right)^5 \ge E[\Phi_{\mu_0}] \ge \log^2 \frac{n}{\delta}$, *then*

$$\left|\frac{\widehat{\mathbb{E}[\Phi_{\mu_0+1}]}}{\widehat{\mathbb{E}[\Phi_{\mu_0}]}} - \frac{\mathbb{E}[\Phi_{\mu_0+1}]}{\mathbb{E}[\Phi_{\mu_0}]}\right| \underset{\delta}{=} \mathcal{O}\left(\frac{\log^2 \frac{n}{\delta}}{\sqrt{\mu_0}(\mathbb{E}[\Phi_{\mu_0}]\sqrt{\mu_0})^{4/11}}\right),$$

and if $E[\Phi_{\mu_0}] > \frac{1}{\log n}\left(\frac{\mu_0}{\log^2 n}\right)^5$ *then,*

$$\left|\frac{\widehat{\mathbb{E}[\Phi_{\mu_0+1}]}}{\widehat{\mathbb{E}[\Phi_{\mu_0}]}} - \frac{\mathbb{E}[\Phi_{\mu_0+1}]}{\mathbb{E}[\Phi_{\mu_0}]}\right| \underset{\delta}{=} \mathcal{O}\left(\frac{\log^2 \frac{n}{\delta}}{\sqrt{\mathbb{E}[\Phi_{\mu_0}]}}\right).$$

5.5. Proposed estimator

Substituting the estimators for $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu+1}]$ in the genie-aided estimator we get the proposed estimator as

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

As mentioned before, for small values of Φ_μ, empirical estimator performs well, and for small values of μ Good-Turing performs well. Therefore, we propose the following (unnormalized) estimator that uses estimator F_μ for μ and $\Phi_\mu \geq \text{polylog}(n)$.

$$F'^{\text{un}}_\mu = \begin{cases} \max\left(G_0, \frac{1}{n}\right) & \text{if } \mu = 0, \\ E_\mu & \text{if } \Phi_\mu \leq \log^2 n, \\ \max\left(G_\mu, \frac{1}{n}\right) & \text{if } \mu \leq \log^2 n \text{ and } \Phi_\mu > \log^2 n, \\ \min\left(\max\left(F_\mu, \frac{1}{n^3}\right), 1\right) & \text{otherwise.} \end{cases}$$

Letting $N \stackrel{\text{def}}{=} \sum_{\mu=0}^{n} F'^{\text{un}}_\mu$, the normalized estimator is then $F'_\mu \stackrel{\text{def}}{=} \frac{1}{N} F'^{\text{un}}_\mu$. Note that the Good-Turing and F_μ may assign 0 probability to M_μ even though $\Phi_\mu \neq 0$. To avoid infinite log loss and KL Divergence between the distribution and the estimate, both estimators are slightly modified by taking $\max\left(G_\mu, \frac{1}{n}\right)$ instead of G_μ and $\min\left(\max\left(F_\mu, \frac{1}{n^3}\right), 1\right)$ instead of F_μ so as not to assign 0 or ∞ probability mass to any multiplicity. Such modifications are common in prediction and compression, (*e.g.*, Krichevsky and Trofimov, 1981).

5.6. Proof sketch of Theorem 2

To prove Theorem 2, we will analyze the unnormalized estimator F'^{un}_μ and prove that $|N-1| \underset{10n^{-2}}{=} \widetilde{\mathcal{O}}(n^{-1/4})$ and use that to prove the desired result for the normalized estimator F'_μ. We first show that the estimation error for every multiplicity is small. The proof is in Appendix B.8.

Lemma 16 *For every distribution p, $|M_0 - F'^{\text{un}}_0| \underset{4n^{-3}}{=} \mathcal{O}\left(\frac{\log^2 n}{\sqrt{n}}\right)$, and for all $\mu \geq 1$,*

$$|M_\mu - F'^{\text{un}}_\mu| \underset{4n^{-3}}{=} \mathcal{O}\left(\frac{\min(\sqrt{\Phi_\mu}(\mu+1), \Phi_\mu^{7/11}\sqrt{\mu+1})}{n \log^{-3} n}\right).$$

The error probability in the above equation is $4n^{-3}$ can be generalized to any $\text{poly}(1/n)$. We have chosen the above error to achieve the over all error in Theorem 2 to be n^{-1}. Note that the error of F'_μ is smaller than both Good-Turing and empirical estimators up to $\text{polylog}(n)$ factors. Using Lemma 16, we show that $N \approx 1$ in the following lemma. It is proved in Appendix B.9.

Lemma 17 *For every distribution p,*

$$|N-1| \underset{10n^{-2}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/4}}\right).$$

Using the bounds on $N-1$ in Lemma 17 and bounds on $|M_\mu - F'^{\text{un}}_\mu|$ in Lemma 16 and maximizing the KL divergence, we prove Theorem 2 in Appendix B.10.

5.7. Lower bounds on estimation

We now lower bound the rate of convergence. We construct an explicit distribution such that with probability $\geq 1 - n^{-1}$ the total variation distance is $\widetilde{\Omega}(n^{-1/4})$. By Pinsker's inequality, this implies that the KL divergence is $\widetilde{\Omega}(n^{-1/2})$. Note that since distance is $\widetilde{\Omega}(n^{-1/4})$ with probability close to 1, the expected distance is also $\widetilde{\Omega}(n^{-1/4})$.

Let p be a distribution with $n_i \stackrel{\text{def}}{=} \sqrt{\frac{\pi}{2}} i \log^{1.5} n$ symbols with probability $p_i \stackrel{\text{def}}{=} \frac{\lfloor i^2 \log^3 n \rfloor}{n}$, and n_i symbols with probability $p_i + \frac{i}{n}$, for $c_1 \frac{n^{1/4}}{\log^{9/8} n} \leq i \leq c_2 \frac{n^{1/4}}{\log^{9/8} n}$. c_1 and c_2 are constants such that the sum of probabilities is 1. We sketch the proof and leave the details to the full version of the paper.

Proof [sketch of Theorem 3] The distribution p has the following properties.

- Let $\mathcal{R} = \cup_i \{np_i, np_i + 1 \ldots np_i + i\}$ for $c_1 \frac{n^{1/4}}{\log^{9/8} n} \leq i \leq c_2 \frac{n^{1/4}}{\log^{9/8} n}$. For every $\mu \in \mathcal{R}$, $\Pr(\varPhi_\mu = 1) \geq 1/3$.
- If $\varPhi_\mu = 1$, then the symbol that has appeared μ times has probability p_i or $p_i + \frac{i}{n}$ with almost equal probability.
- Label-invariant estimators cannot distinguish between the two cases, and hence incur an error of $\widetilde{\Omega}(i/n) = \widetilde{\Omega}(n^{-3/4})$ for a constant fraction of multiplicities $\mu \in \mathcal{R}$.

The total number of multiplicities in $\mathcal{R}$ is $n^{1/4} \cdot n^{1/4} = n^{1/2}$. Multiplying by the error for each multiplicity yields the bound $\widetilde{\Omega}(n^{-1/4})$. ■

References

J. Acharya, H. Das, H. Mohimani, A. Orlitsky, and Shengjun Pan. Exact calculation of pattern probabilities. In *ISIT*, pages 1498 –1502, June 2010.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *JMLR - Proceedings Track*, 19:47–68, 2011.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A.T. Suresh. Competitive classification and closeness testing. *JMLR - Proceedings Track*, 23:22.1–22.18, 2012a.

J. Acharya, H. Das, and A. Orlitsky. Tight bounds on profile redundancy and distinguishability. In *NIPS*, 2012b.

T. Batu. *Testing properties of distributions.* PhD thesis, Cornell University, 2001.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *FOCS*, pages 259–269, 2000.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Ulisses Braga-Neto. Classification and error estimation for discrete data. *Pattern Recognition*, 10(7):446–462, 2009.

N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. COLT, pages 12–18, New York, NY, USA, 1999. ACM.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 310–318. Association for Computational Linguistics, 1996.

H. Das. *Competitive Tests and Estimators for Properties of Distributions.* PhD thesis, UCSD, 2012.

E. Drukh and Y. Mansour. Concentration bounds for unigrams language model. In *COLT*, 2004.

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

I. J. Good. The population frequencies of species and the estimation of population parameters. 40(3-4):237–264, 1953.

P. D. Grünwald. *The Minimum Description Length Principle.* The MIT Press, 2007.

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

L. LeCam. *Asymptotic methods in statistical decision theory.* Springer series in statistics. Springer, New York, 1986.

G.G. Lorentz. *Bernstein polynomials.* Chelsea Publishing Company, Incorporated, 1986.

D. A. McAllester and R. E. Schapire. On the convergence rate of good-turing estimators. In *COLT*, 2000.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005.

M. I. Ohannessian and M A. Dahleh. Rare probability estimation under regularly varying heavy tails. *JMLR- Proceedings Track*, 23:21.1–21.24, 2012.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Always good turing: Asymptotically optimal probability estimation. In *FOCS*, 2003.

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

A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang. Convergence of profile based estimators. In *ISIT 2005*, pages 1843 –1847, September 2005.

L. Paninski. Variational minimax estimation of discrete distributions under kl loss. In *NIPS*, 2004.

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10), 2008.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, July 1984.

G. Shamir. A new upper bound on the redundancy of unknown alphabets. In *Proceedings of The Annual Conference on Information Sciences and Systems, Princeton, New-Jersey*, 2004.

G. Valiant and P. Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. STOC. ACM, 2011.

V. G. Vovk. A game of prediction with expert advice. COLT, pages 51–60, New York, NY, USA, 1995. ACM.

A. B. Wagner, P. Viswanath, and S. R. Kulkarni. Strong consistency of the good-turing estimator. In *ISIT*, 2006.

Appendix A. Useful facts

A.1. Concentration inequalities

The following two popular concentration inequalities are stated for completeness.

Fact 18 (Chernoff bound) *If $X \sim \text{Poi}(\lambda)$, then for $x \geq \lambda$,*

$$\Pr(X \geq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2x}\right),$$

and for $x < \lambda$,

$$\Pr(X \leq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2\lambda}\right).$$

Fact 19 (Variation of Bernstein's Inequality) *Let $X_1, X_2, \ldots X_n$ be n independent zero mean random variables such that with probability $\geq 1 - \epsilon_i$, $|X_i| < M$, then*

$$\Pr(|\sum_i X_i| \geq t) \leq 2\exp\left(-\frac{t^2}{\sum_i \mathbb{E}[X_i^2] + Mt/3}\right) + \sum_{i=1}^n \epsilon_i.$$

If $t = \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}$, then

$$\Pr\left(\left|\sum_i X_i\right| \geq \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}\right) \leq 2\delta + \sum_{i=1}^n \epsilon_i.$$

To prove the concentration of estimators, we bound the variance and show that with high probability the absolute value of each X_i is bounded by M and use Bernstein's inequality with $t = \sqrt{2\left(\sum_i \mathbb{E}[X_i^2]\right)\log\frac{1}{\delta}} + \frac{2}{3}M\log\frac{1}{\delta}$.

A.2. Bounds on linear estimators

In this section, we prove error bounds for linear estimators that are used to simplify other proofs in the paper. We first show that the difference of expected values of consecutive Φ_μ's is bounded.

Claim 20 *For every distribution p and every μ,*

$$|\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu+1}]| = \mathcal{O}\left(\mathbb{E}[\Phi_\mu]\max\left(\frac{\log n}{\mu+1}, \sqrt{\frac{\log n}{\mu+1}}\right)\right) + \frac{1}{n}.$$

Proof We consider the two cases $\mu + 1 \geq \log n$ and $\mu + 1 < \log n$ separately. Consider the case when $\mu + 1 \geq \log n$. We first show that

$$\left|\mathbb{E}[\mathbb{1}_x^\mu] - \mathbb{E}[\mathbb{1}_x^{\mu+1}]\right| = e^{-np_x}\frac{(np_x)^\mu}{\mu!}\left|1 - \frac{np_x}{\mu+1}\right| \leq 5e^{-np_x}\frac{(np_x)^\mu}{\mu!}\sqrt{\frac{\log n}{\mu+1}} + \frac{2}{n^3}. \tag{4}$$

The first equality follows by substituting $\mathbb{E}[\mathbb{1}_x^\mu] = e^{-np_x}(np_x)^\mu/\mu!$. For the inequality, note that if $|np_x - \mu - 1|^2 \leq 25(\mu+1)\log n$, then the inequality follows. If not, then by the Chernoff bound $\mathbb{E}[\mathbb{1}_x^\mu] = \Pr(\mu_x = \mu) \leq n^{-3}$ and hence $\left|\mathbb{E}[\mathbb{1}_x^\mu] - \mathbb{E}[\mathbb{1}_x^{\mu+1}]\right| \leq \mathbb{E}[\mathbb{1}_x^\mu] + \mathbb{E}[\mathbb{1}_x^{\mu+1}] \leq 2/n^3$.

By definition, $\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu+1}] = \sum_x \mathbb{E}[\mathbb{1}_x^\mu] - \mathbb{E}[\mathbb{1}_x^{\mu+1}]$. Substituting,

$$\begin{aligned}
|\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu+1}]| &\leq \sum_x \left|\mathbb{E}[\mathbb{1}_x^\mu] - \mathbb{E}[\mathbb{1}_x^{\mu+1}]\right| \\
&\overset{(a)}{=} \sum_x e^{-np_x}\frac{(np_x)^\mu}{\mu!}\left|1 - \frac{np_x}{\mu+1}\right| \\
&= \sum_{x:np_x\leq 1} e^{-np_x}\frac{(np_x)^\mu}{\mu!}\left|1 - \frac{np_x}{\mu+1}\right| + \sum_{x:np_x>1} e^{-np_x}\frac{(np_x)^\mu}{\mu!}\left|1 - \frac{np_x}{\mu+1}\right| \\
&\overset{(b)}{\leq} \sum_{x:np_x\leq 1}\frac{np_x}{\mu!} + \sum_{x:np_x>1} 5e^{-np_x}\frac{(np_x)^\mu}{\mu!}\sqrt{\frac{\log n}{\mu+1}} + \frac{2}{n^3} \\
&\leq \frac{1}{n^2} + \mathcal{O}\left(\mathbb{E}[\Phi_\mu]\sqrt{\frac{\log n}{\mu+1}}\right) + \frac{2n}{n^3} \leq \mathcal{O}\left(\mathbb{E}[\Phi_\mu]\sqrt{\frac{\log n}{\mu+1}}\right) + \frac{1}{n}.
\end{aligned}$$

where (a) follows from the fact that $\mathbb{E}[\mathbb{1}_x^\mu] = e^{-np_x}(np_x)^\mu/\mu!$. (b) follows from the fact that $np_x \leq 1$ in the first summation and Equation (4). The proof for the case $\mu + 1 < \log n$ is similar and hence omitted. ■

The next claim bounds the variance of any linear estimator in terms of its coefficients.

Claim 21 *For every distribution p,*

$$\mathrm{Var}\left(\sum_x \sum_\mu \mathbb{1}_x^\mu f(x,\mu)\right) \le \sum_x \sum_\mu \mathbb{E}[\mathbb{1}_x^\mu] f^2(x,\mu).$$

Proof By Poisson sampling, the multiplicities are independent. Furthermore the variance of sum of independent random variables is the sum of their variances. Hence,

$$\begin{aligned}
\mathrm{Var}\left(\sum_x \sum_\mu \mathbb{1}_x^\mu f(x,\mu)\right) &= \sum_x \mathrm{Var}\left(\sum_\mu \mathbb{1}_x^\mu f(x,\mu)\right) \\
&\le \sum_x \mathbb{E}\left[\left(\sum_\mu \mathbb{1}_x^\mu f(x,\mu)\right)^2\right] \\
&\overset{(a)}{=} \sum_x \mathbb{E}\left[\sum_\mu (\mathbb{1}_x^\mu f(x,\mu))^2\right] \\
&\overset{(b)}{=} \sum_x \sum_\mu \mathbb{E}[\mathbb{1}_x^\mu] f^2(x,\mu).
\end{aligned}$$

For $\mu \neq \mu'$, $\mathbb{E}[\mathbb{1}_x^\mu \mathbb{1}_x^{\mu'}] = 0$ and hence (a). (b) uses the fact that $\mathbb{1}_x^\mu$ is an indicator random variable. ■

Next we prove a concentration inequality for any linear estimator f.

Claim 22 *Let $r \le \sqrt{\frac{\mu_0}{\log n}}$, $\mu_0 \ge \log n$, and $f = \sum_{\mu \in \mathcal{S}_r^{\mu_0}} c_\mu \varPhi_\mu$. For every distribution p if $\mathbb{E}[\varPhi_{\mu_0}] \ge \log \frac{1}{\delta}$, then*

$$|f - \mathbb{E}[f]| \underset{\delta}{=} \mathcal{O}\left(\max_{\mu \in \mathcal{S}_r^{\mu_0}} |c_\mu| \sqrt{\mathbb{E}[\varPhi_{\mu_0}](2r+1)\log\frac{1}{\delta}}\right).$$

Proof By Claim 21,

$$\begin{aligned}
\mathrm{Var}(f) &\le \sum_{\mu \in \mathcal{S}_r^{\mu_0}} \sum_x c_\mu^2 \mathbb{E}[\mathbb{1}_x^\mu] \\
&\le \left(\max_{\mu \in \mathcal{S}_r^{\mu_0}} c_\mu\right)^2 \sum_{\mu \in \mathcal{S}_r^{\mu_0}} \sum_x \mathbb{E}[\mathbb{1}_x^\mu] \\
&\overset{(a)}{=} \left(\max_{\mu \in \mathcal{S}_r^{\mu_0}} c_\mu\right)^2 \sum_{\mu \in \mathcal{S}_r^{\mu_0}} \mathbb{E}[\varPhi_\mu] \\
&= \mathcal{O}\left(\left(\max_{\mu \in \mathcal{S}_r^{\mu_0}} c_\mu\right)^2 (2r+1)\mathbb{E}[\varPhi_{\mu_0}]\right).
\end{aligned}$$

Substituting $\sum_x \mathbb{E}[\mathbb{1}_x^\mu] = \mathbb{E}[\varPhi_\mu]$ results in (a). The last equality follows by repeatedly applying Claim 20. Changing one of the multiplicities changes f by at-most $\max_{\mu \in \mathcal{S}_r^{\mu_0}} |c_\mu|$. Applying Bernstein's inequality with the above calculated bounds on variance, $M = \max_{\mu \in \mathcal{S}_r^{\mu_0}} |c_\mu|$, and $\sum_i \epsilon_i = 0$ yields the claim. ■

Next we prove a concentration bound for Φ_μ in the next claim.

Claim 23 *For every distribution p and every multiplicity μ, if $\mathbb{E}[\Phi_\mu] \geq \log \frac{1}{\delta}$, then*

$$|\Phi_\mu - \mathbb{E}[\Phi_\mu]| \underset{\delta}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_\mu] \log \frac{1}{\delta}}\right).$$

Proof Since $\Phi_\mu = \sum_x \mathbb{1}_x^\mu$, by Claim 21, $\mathrm{Var}(\Phi_\mu) \leq \mathbb{E}[\Phi_\mu]$. Furthermore $|\mathbb{1}_x^\mu - \mathbb{E}(\mathbb{1}_x^\mu)| \leq 1$. Applying Bernstein's inequality with $M = 1$, $\mathrm{Var}(\Phi_\mu) \leq \mathbb{E}[\Phi_\mu]$, and $\sum_i \epsilon_i = 0$ proves the claim. ■

Appendix B. Proofs of results in Section 5

B.1. Proof of Lemma 8

If a distribution is sampled $n' = \mathrm{Poi}(n)$ times, with probability $e^{-n}\frac{n^n}{n!} \geq \frac{1}{e\sqrt{n}}$, $n' = n$. Conditioned on the fact that $n' = n$, Poisson sampling is same as sampling the distribution exactly n times. Therefore, if P fails with probability $> \delta \cdot e\sqrt{n}$ with exactly n samples, then P fails with probability $> \delta$ when sampled $\mathrm{Poi}(n)$ times. ■

B.2. Proof of Claim 9

The proof follows from the fact that each multiplicity is a Poisson random variable under Poisson sampling.

$$\mathbb{E}[M_\mu] = \mathbb{E}\Big[\sum_x p_x \cdot \mathbb{1}_x^\mu\Big] = \sum_x p_x \cdot e^{-np_x}\frac{(np_x)^\mu}{\mu!} = \frac{\mu+1}{n}\sum_x e^{-np_x}\frac{(np_x)^{\mu+1}}{(\mu+1)!} = \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}].$$

■

B.3. Proof of Lemma 10

Let $\epsilon = \frac{20\sqrt{\mu+1}\log\frac{n}{\delta}}{n}$. Since $\varphi_\mu = \sum_x \mathbb{1}_x^\mu$ and $M_\mu = \sum_x p_x \mathbb{1}_x^\mu$,

$$\Pr\left(\left|M_\mu - \Phi_\mu \frac{\mu}{n}\right| \geq \Phi_\mu \epsilon\right) \leq \Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{\mu}{n}\right| > \epsilon, \mathbb{1}_x^\mu = 1\right).$$

If $p_x \geq \frac{\mu}{n} + \epsilon$, then by the Chernoff bound $\Pr(\mathbb{1}_x^\mu = 1) \leq \delta/2n$. Therefore by the union bound,

$$\Pr\left(\exists\, x \text{ s.t. } p_x - \frac{\mu}{n} > \epsilon, \mathbb{1}_x^\mu = 1\right) \leq n\frac{\delta}{2n} \leq \frac{\delta}{2}.$$

Now consider the set of symbols such that $p_x \leq \frac{\mu}{n} - \epsilon$. Since $p_x \geq 0$, we have $\mu \geq 20\sqrt{\mu+1}\log\frac{n}{\delta}$. Group symbols x with probability $\leq 1/4n$ in to smallest number of groups such that $\Pr(g) \leq 1/n$ for each group g. By Poisson sampling, for each group g, $\mu_g =$

$\sum_{x \in g} \mu_x$ and μ_g is a Poisson random variable with mean $\Pr(g)$. Observe that for any two (or more) symbols x and x', $\Pr(\mu_x \geq \mu \vee \mu_{x'} \geq \mu) \leq \Pr(\mu_x + \mu_{x'} \geq \mu)$. Therefore

$$\begin{aligned}\Pr\left(\exists\, x \text{ s.t. } \frac{\mu}{n} - p_x > \epsilon, \mathbb{1}_x^\mu = 1\right) &\leq \Pr\left(\exists\, x \text{ s.t. } \mu_x \geq \mu, p_x \leq \frac{\mu}{n} - \epsilon\right)\\ &\leq \Pr\left(\exists\, g \text{ s.t. } \mu_g \geq \mu \vee \exists x \text{ s.t. } \quad \mu_x \geq \mu, \frac{1}{4n} \leq p_x \leq \frac{\mu}{n} - \epsilon\right).\end{aligned}$$

It is easy to see that the number of groups and the number of symbols with probabilities $\geq 1/4n$ is at most $n+1+4n \leq 6n$. Therefore by the union bound and the Chernoff bound the above probability is $\leq \delta/2$. Adding the error probabilities for cases $p_x \geq \frac{\mu}{n} + \epsilon$ and $p_x \leq \frac{\mu}{n} - \epsilon$ results in the lemma. ■

B.4. Proof of Lemma 11

By Claim 9, $\mathbb{E}\left[M_\mu - \Phi_{\mu+1}\frac{\mu+1}{n}\right] = 0$. Recall that $M_\mu = \sum_x p_x \mathbb{1}_x^\mu$ and $\Phi_{\mu+1} = \sum_x \mathbb{1}_x^{\mu+1}$. Hence by Claim 21 (stated and proved in Appendix A),

$$\begin{aligned}\text{Var}\left(M_\mu - \Phi_{\mu+1}\frac{\mu+1}{n}\right) &\leq \sum_x \mathbb{E}[\mathbb{1}_x^\mu]p_x^2 + \mathbb{E}[\mathbb{1}_x^{\mu+1}]\frac{(\mu+1)^2}{n^2}\\ &\overset{(a)}{=} \sum_x \mathbb{E}[\mathbb{1}_x^{\mu+2}]\frac{(\mu+1(\mu+2)}{n^2} + \mathbb{E}[\mathbb{1}_x^{\mu+1}]\frac{(\mu+1)^2}{n^2}\\ &\overset{(b)}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_{\mu+1}]+1)(\mu+1)^2\log n}{n^2}\right).\end{aligned}$$

$\mathbb{E}[\mathbb{1}_x^\mu] = e^{-np_x}(np_x)^\mu/\mu!$ and $\mathbb{E}[\mathbb{1}_x^{\mu+2}] = e^{-np_x}(np_x)^{\mu+2}/\mu+2!$, and hence (a). (b) follows from Claim 20 (stated and proved in Appendix A) and the fact that $\sum_x \mathbb{E}[\mathbb{1}_x^{\mu+2}] = \mathbb{E}[\Phi_{\mu+2}]$. By the proof of Lemma 10,

$$\Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{\mu}{n}\right| > \frac{20\sqrt{\mu+1}\log\frac{n}{\delta'}}{n}, \mathbb{1}_x^\mu = 1\right) \leq \delta'.$$

Choosing $\delta' = \delta/2$ we get $\forall x, |\mathbb{1}_x^\mu p_x - \mathbb{1}_x^{\mu+1}\frac{\mu+1}{n}| = \mathcal{O}\left(\frac{\sqrt{\mu+1}\log\frac{n}{\delta}}{n} + \frac{\mu+1}{n}\right)$ with probability $1-\delta/2$. The lemma follows from Bernstein's inequality with $M = \mathcal{O}\left(\frac{\sqrt{\mu+1}\log\frac{n}{\delta}}{n} + \frac{\mu+1}{n}\right)$, $\sum_i \epsilon_i = \delta/2$, and above calculated bound on the variance. ■

B.5. Proof of Lemma 13

By Claim 9,

$$\mathbb{E}[M_\mu] - \mathbb{E}[\Phi_\mu]\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]} = 0.$$

We now bound the variance. By definition, $M_\mu = \sum_x p_x \mathbb{1}_x^\mu$ and $\Phi_{\mu+1} = \sum_x \mathbb{1}_x^{\mu+1}$. Using Claim 21,

$$\begin{aligned}
\operatorname{Var}\left(M_\mu - \frac{(\mu+1)\Phi_\mu}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\right) &\le \sum_x \mathbb{E}[\mathbb{1}_x^\mu]\left(p_x - \frac{(\mu+1)\mathbb{E}[\Phi_{\mu+1}]}{n\mathbb{E}[\Phi_\mu]}\right)^2 \\
&= \sum_x \mathbb{E}[\mathbb{1}_x^\mu]\left(p_x - \frac{\mu+1}{n} + \frac{(\mu+1)(\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu+1}])}{n\mathbb{E}[\Phi_\mu]}\right)^2 \\
&\overset{(a)}{\le} \sum_x 2\mathbb{E}[\mathbb{1}_x^\mu]\left(p_x - \frac{\mu+1}{n}\right)^2 + 2\mathbb{E}[\mathbb{1}_x^\mu]\left(\frac{(\mathbb{E}[(\Phi_{\mu+1}] - \mathbb{E}[\Phi_\mu])(\mu+1)}{n\mathbb{E}[\Phi_\mu]}\right)^2 \\
&\overset{(b)}{=} \mathcal{O}\left(\frac{\mathbb{E}[\Phi_\mu]\mu\log^2 n}{n^2}\right),
\end{aligned}$$

where (a) follows from the fact that $(x+y)^2 \le 2x^2 + 2y^2$. Similar to the proof of Claim 20, one can show that the first term in (a) is $\mathcal{O}\left(\frac{\mathbb{E}[\Phi_\mu]\mu\log^2 n}{n^2}\right)$. The second term can be bounded by $\mathcal{O}\left(\frac{\mathbb{E}[\Phi_\mu]\mu\log^2 n}{n^2}\right)$ using Claim 20, hence (b). We now bound the maximum value of each individual term in the summation. By the proof of Lemma 10,

$$\Pr\left(\exists\, x \text{ s.t. } \left|p_x - \frac{\mu}{n}\right| > \frac{c\sqrt{\mu+1}\log\frac{n}{\delta'}}{n}, \mathbb{1}_x^\mu = 1\right) \le \delta' \tag{5}$$

Choosing $\delta' = \delta/2$ we get that with probability $1 - \delta/2$, $\forall x$

$$\begin{aligned}
\mathbb{1}_x^\mu\left|p_x - \frac{(\mu+1)\mathbb{E}[\Phi_{\mu+1}]}{n\mathbb{E}[\Phi_\mu]}\right| &\le \mathbb{1}_x^\mu\left|p_x - \frac{\mu+1}{n}\right| + \left|\frac{(\mu+1)\mathbb{E}[\Phi_{\mu+1}] - \mathbb{E}[\Phi_\mu]}{n\mathbb{E}[\Phi_\mu]}\right| \\
&\overset{(a)}{=} \mathcal{O}\left(\frac{\sqrt{\mu+1}\log\frac{n}{\delta}}{n} + \frac{(\mu+1)\log n}{n}\right) \\
&= \mathcal{O}\left(\frac{(\mu+1)\log\frac{n}{\delta}}{n}\right).
\end{aligned}$$

where the (a) follows from Lemma 20 and Equation (5). The lemma follows from Bernstein's inequality with the calculated variance, $M = \mathcal{O}\left(\frac{(\mu+1)\log\frac{n}{\delta}}{n}\right)$, and $\sum_i \epsilon_i = \delta/2$. ■

B.6. Proof of Lemma 14

By assumption, $|r(x-1)| \le \min(1,x)$. Hence $|r\ln x| < 2|r(x-1)|$ and $|r\ln x| \le 1$. Therefore

$$\begin{aligned}
|B_r(x)| &= \left|1 - \gamma_r(0) - \sum_{i=1}^r \gamma_r(i) 2\cosh(i\ln x)\right| \\
&= \left|1 - \gamma_r(0) - 2\sum_{i=1}^r \gamma_r(i)\Big(1 + \frac{(i\ln x)^2}{2!} + \frac{(i\ln x)^4}{4!} + \frac{(i\ln x)^6}{6!} + \cdots\Big)\right| \\
&\overset{(a)}{=} \left|2\sum_{i=1}^r \gamma_r(i)\Big(\frac{(i\ln x)^4}{4!} + \frac{(i\ln x)^6}{6!} + \cdots\Big)\right| \\
&\overset{(b)}{\le} 2\sum_{i=1}^r \left|\gamma_r(i)\right| 2\frac{(i\ln x)^4}{4!},
\end{aligned}$$

where in (a) we use that $\gamma_r(0) + 2\sum_{i=1}^{r} \gamma_r(i) = 1$ and $\sum_{i=1}^{r} \gamma_r(i)i^2 = 0$. (b) follows from the fact that $|r \ln x| \leq 1$. Now using $r|\ln(x)| \leq 2r|x-1|$, and $|\gamma_r(i)| = \mathcal{O}\left(\frac{1}{r+1}\right)$ (can be shown), the result follows. ■

B.7. Proof of Lemma 15

The proof is technically involved and we prove it in steps. We first observe the following property of $a_{\mu}^{\mu_0}$. The proof follows from the definition.

Claim 24 *For every distribution p and multiplicities μ, μ_0,*

$$a_{\mu}^{\mu_0} \mathbb{E}[\mathbb{1}_x^{\mu}] = \mathbb{E}[\mathbb{1}_x^{\mu_0}] \left(\frac{np_x}{\mu_0}\right)^{\mu-\mu_0}.$$

Next we bound $\widehat{\mathbb{E}[\Phi_\mu]} - \mathbb{E}[\Phi_\mu]$. The proposed estimators for $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu+1}]$ have positive bias. Hence we analyze $\widehat{\mathbb{E}[\Phi_\mu]} - \widehat{\mathbb{E}[\Phi_{\mu+1}]}$ to prove tighter bounds for the ratio.

Lemma 25 *Let $r \leq \frac{\sqrt{\mu_0}}{\log n}$ and $\mu_0 \geq \log n$. For every distribution p, if $\mathbb{E}[\Phi_{\mu_0}] \geq \log \frac{1}{\delta}$, then*

$$\left|\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \mathbb{E}[\Phi_{\mu_0}]\right| \underset{\delta}{=} \mathcal{O}\left(\frac{r^4 \log^2 n \mathbb{E}[\Phi_{\mu_0}]}{\mu_0^2} + \sqrt{\frac{\mathbb{E}[\Phi_{\mu_0}] \log \frac{1}{\delta}}{r+1}}\right),$$

and

$$\left|\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]} - \mathbb{E}[\Phi_{\mu_0} - \Phi_{\mu_0+1}]\right| \underset{\delta}{=} \mathcal{O}\left(\frac{r^4 \mathbb{E}[\Phi_{\mu_0}] \log^{2.5} n}{\mu_0^{2.5}} + \frac{\sqrt{\mathbb{E}[\Phi_{\mu_0}] \log \frac{1}{\delta}}}{(r+1)^{1.5}}\right).$$

Proof By triangle inequality, $\left|\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \mathbb{E}[\Phi_{\mu_0}]\right| \leq |\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}]| + \left|\mathbb{E}[\Phi_{\mu_0}] - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}]\right|$. We first bound $|\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}]|$.

Since $r \leq \sqrt{\mu_0}$ it can show that $a_{\mu}^{\mu_0} \leq e$ and $|\gamma_r(|\mu - \mu_0|)| = \mathcal{O}((r+1)^{-1})$. Therefore each coefficient in $\widehat{\mathbb{E}[\Phi_{\mu_0}]}$ is $\mathcal{O}((r+1)^{-1})$. Hence by Claim 22 (stated and proved in Appendix A),

$$\left|\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}]\right| \underset{\delta}{=} \mathcal{O}\left(\sqrt{\frac{\mathbb{E}[\Phi_{\mu_0}] \log \frac{1}{\delta}}{r+1}}\right).$$

Next we bound the bias, *i.e.*, $\left|\mathbb{E}[\Phi_{\mu_0}] - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}]\right|$. Recall that $a_{\mu}^{\mu_0} \mathbb{E}[\mathbb{1}_x^{\mu}] = \mathbb{E}[\mathbb{1}_x^{\mu_0}] \left(\frac{np_x}{\mu_0}\right)^{\mu-\mu_0}$. Therefore by the linearity of expectation and the definition of $B_r(x)$,

$$\mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]}] - \mathbb{E}[\Phi_{\mu_0}] = \sum_x \mathbb{E}[\mathbb{1}_x^{\mu_0}] B_r\left(\frac{np_x}{\mu_0}\right).$$

For $r = 0$, the bias is 0. For $r \geq 1$, by the Chernoff bound and the grouping argument similar to that in the proof of empirical estimator 10, it can be shown that there is a constant

c such that if $|np_x - \mu_0| \geq c\sqrt{\mu_0 \log n}$, then $\sum_{x \in \mathcal{X}} \mathbb{E}[\mathbb{1}_x^{\mu_0}] B_r\left(\frac{np_x}{\mu_0}\right) \leq n^{-3}$. If not, then by Lemma 14, $B_r\left(\frac{np_x}{\mu_0}\right) = \mathcal{O}\left(\frac{r^4 \log^2 n}{\mu_0^2}\right)$. Bounding $\mathbb{E}[\mathbb{1}_x^{\mu_0}] B_r\left(\frac{np_x}{\mu_0}\right)$ for each alphabet x and using the fact that $\mathbb{E}[\Phi_{\mu_0}] \geq \log \frac{1}{\delta}$, we get

$$\left| E[\widehat{\mathbb{E}[\Phi_{\mu_0}]}] - \mathbb{E}[\Phi_{\mu_0}] \right| = \sum_x \mathbb{E}[\mathbb{1}_x^{\mu_0}] \mathcal{O}\left(\frac{r^4 \log^2 n}{\mu_0^2}\right) + \frac{1}{n^3} = \mathcal{O}\left(\mathbb{E}[\Phi_{\mu_0}] \frac{r^4 \log^2 n}{\mu_0^2}\right).$$

The first part of the lemma follows by the union bound. The proof of the second part is similar. We will prove the concentration of $\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]}$ and then quantify the bias. We first bound the coefficients in $\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]}$. The coefficient of Φ_μ is bounded by

$$a_\mu^{\mu_0} \frac{|\gamma_r(|\mu_0 + 1 - \mu|)|}{\mu_0 + 1} + a_\mu^{\mu_0} \left|\gamma_r(|\mu_0 + 1 - \mu|) - \gamma_r(|\mu_0 - \mu|)\right| = \mathcal{O}\left(\frac{1}{(r+1)^2}\right).$$

Applying Claim 22, we get

$$\left| \widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]} - \mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]}] \right| \underset{\delta}{=} \mathcal{O}\left(\frac{\sqrt{\mathbb{E}[\Phi_{\mu_0}] \log \frac{1}{\delta}}}{(r+1)^{1.5}}\right).$$

Next we bound the bias.

$$\mathbb{E}[\widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]}] - \mathbb{E}[\Phi_{\mu_0} - \Phi_{\mu_0+1}] = \sum_x \mathbb{E}[\mathbb{1}_x^{\mu_0}] \left(1 - \frac{np_x}{\mu_0 + 1}\right) B_r\left(\frac{np_x}{\mu_0}\right).$$

As before, bounding $\mathbb{E}[\mathbb{1}_x^{\mu_0}] \left(1 - \frac{np_x}{\mu_0+1}\right) B_r\left(\frac{np_x}{\mu_0}\right)$ for each x yields the lemma. ■

Now we have all the tools to prove Lemma 15.
Proof [Lemma 15.] If $|\Delta b| \leq 0.9b$, then

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

Let $b = \mathbb{E}[\Phi_{\mu_0}]$, $a = \mathbb{E}[\Phi_{\mu_0+1} - \Phi_{\mu_0}]$, $\Delta b = \widehat{\mathbb{E}[\Phi_{\mu_0+1}]} - \mathbb{E}[\Phi_{\mu_0}]$ and $\Delta a = \widehat{\mathbb{E}[\Phi_{\mu_0}]} - \widehat{\mathbb{E}[\Phi_{\mu_0+1}]} - \mathbb{E}[\Phi_{\mu_0} - \Phi_{\mu_0+1}]$. By Lemma 25, if $\mathbb{E}[\Phi_{\mu_0}] \geq \log^2 \frac{n}{\delta}$ and $\mu_0 \geq r^2 \log^{1.5} n$, then $|\Delta b| \leq 0.9b$. Therefore by Lemma 25, Claim 20, and the union bound,

$$\left| \frac{\widehat{\mathbb{E}[\Phi_{\mu_0+1}]}}{\widehat{\mathbb{E}[\Phi_{\mu_0}]}} - \frac{\mathbb{E}[\Phi_{\mu_0+1}]}{\mathbb{E}[\Phi_{\mu_0}]} \right| \underset{2\delta'}{=} \mathcal{O}\left(\frac{r^4 \log^{2.5} n}{\mu_0^{2.5}} + \frac{\log^{0.5} \frac{n}{\delta'}}{(r+1)^{1.5}\sqrt{\mathbb{E}[\Phi_{\mu_0}]}}\right). \tag{6}$$

By Claim 23 (stated and proved in Appendix A), if $\mathbb{E}[\Phi_{\mu_0}] \geq \log^2 \frac{n}{\delta}$, then with probability $1 - \delta/2$, $\Phi_{\mu_0} \in [0.5\mathbb{E}[\Phi_{\mu_0}], 2\mathbb{E}[\Phi_{\mu_0}]]$. Hence,

$$r_{\mu_0} \in \mathcal{R} \overset{\text{def}}{=} \left[\left\lfloor \frac{\sqrt{\mu_0}}{(2\mathbb{E}[\Phi_{\mu_0}]\sqrt{\mu_0})^{1/11} \log n} \right\rfloor, \left\lfloor \frac{\sqrt{\mu_0}}{(0.5\mathbb{E}[\Phi_{\mu_0}]\sqrt{\mu_0})^{1/11} \log n} \right\rfloor\right].$$

Therefore if we prove the concentration bounds for all $r \in \mathcal{R}$, the lemma would follow by the union bound. If $\max_r \mathcal{R} < 1$, then substituting $r = 0$ in Equation (6) yields the result for the case $\mathbb{E}[\Phi_{\mu_0}] \geq \frac{2}{\log n}\left(\frac{\mu_0}{\log^2 n}\right)^5$. If $\min_r \mathcal{R} \geq 1$, then substituting $r = \Theta\left(\frac{\sqrt{\mu}}{(\mathbb{E}[\Phi_{\mu_0}]\sqrt{\mu_0})^{1/11}\log n}\right)$ in Equation (6) yields the result for the case $\mathbb{E}[\Phi_{\mu_0}] \leq \frac{0.5}{\log n}\left(\frac{\mu_0}{\log^2 n}\right)^5$. A similar analysis proves the result for the case $1 \in \mathcal{R}$. Choosing $\delta' = \delta/2$ in Equation (6) and using the union bound we get the total error probability $\leq \delta$. ■

B.8. Proof of Lemma 16

The proof uses the bound on the error of F_μ, which is given below.

Lemma 26 *For every distribution p and $\mu \geq \log^2 n$, if $\frac{1}{\log n}\left(\frac{\mu}{\log^2 n}\right)^5 \geq E[\Phi_\mu] \geq \log^2 \frac{n}{\delta}$, then*

$$|M_\mu - F_\mu| \underset{2\delta}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_\mu]\sqrt{\mu})^{7/11}\log^2\frac{n}{\delta}}{n} + \frac{\sqrt{\mathbb{E}[\Phi_\mu]\mu}\log^2\frac{n}{\delta}}{n}\right),$$

and if $E[\Phi_\mu] \geq \frac{1}{\log n}\left(\frac{\mu}{\log^2 n}\right)^5$, then

$$|M_\mu - F_\mu| \underset{2\delta}{=} \mathcal{O}\left(\frac{\mu\sqrt{\mathbb{E}[\Phi_\mu]}\log^2\frac{n}{\delta}}{n} + \frac{\sqrt{\mathbb{E}[\Phi_\mu]\mu}\log^2\frac{n}{\delta}}{n}\right).$$

Proof is a simple application of triangle inequality and the union bound. It follows from Lemmas 13 and 15. ■

Proof [Lemma 16] We first show that $\mathbb{E}[\Phi_\mu]$ and $\mathbb{E}[\Phi_{\mu+1}]$ in the bounds of Lemmas 26 and 11 can be replaced by Φ_μ. By Claim 20, if $\mathbb{E}[\Phi_{\mu+1}] \geq 1$,

$$|\mathbb{E}[\Phi_\mu] - \mathbb{E}[\Phi_{\mu+1}]| = \mathcal{O}\left(\mathbb{E}[\Phi_\mu]\max\left(\frac{\log n}{\mu+1}, \sqrt{\frac{\log n}{\mu+1}}\right)\right) + \frac{1}{n} = \mathcal{O}\left(\mathbb{E}[\Phi_\mu]\log n\right).$$

Hence $\mathbb{E}[\Phi_{\mu+1}] = \mathcal{O}(\mathbb{E}[\Phi_\mu]\log n)$. Hence by Lemma 11, for $\mathbb{E}[\Phi_\mu] \geq 1$,

$$|M_\mu - G_\mu| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_{\mu+1}] + 1}\frac{(\mu+1)\log^2 n}{n}\right) = \mathcal{O}\left(\sqrt{\mathbb{E}[\Phi_\mu]}\frac{(\mu+1)\log^3 n}{n}\right).$$

Furthermore by Claim 23, if $\mathbb{E}[\Phi_\mu] \leq 0.5\log^2 n$, then $\Phi_\mu \leq \log^2 n$ with probability $\geq 1 - 0.5n^{-3}$, and we use the empirical estimator. Therefore with probability $\geq 1 - 0.5n^{-3}$, F_μ and G_μ are used only if $\mathbb{E}[\Phi_\mu] \geq 0.5\log^2 n$. If $\mathbb{E}[\Phi_\mu] \geq 0.5\log^2 n$, then by Claim 23 $\mathbb{E}[\Phi_\mu] \underset{0.5n^{-3}}{=} \mathcal{O}(\Phi_\mu)$. Therefore by the union bound, if $\Phi_\mu \geq \log^2 n$, then

$$|M_\mu - G_\mu| \underset{n^{-3}}{=} \mathcal{O}\left(\sqrt{\Phi_\mu}\frac{(\mu+1)\log^3 n}{n}\right).$$

Similarly by Lemma 26, for $\mu \geq \log^2 n$ and $\mathit{\Phi}_\mu \geq \log^2 n$, if $\frac{1}{\log n}\left(\frac{\mu}{\log^2 n}\right)^5 \geq E[\mathit{\Phi}_\mu] \geq \log^2 n$, then

$$|M_\mu - F_\mu| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\mathit{\Phi}_\mu]\sqrt{\mu})^{7/11}\log^2 n}{n} + \frac{\sqrt{\mathbb{E}[\mathit{\Phi}_\mu]\mu}\log^2 n}{n}\right) \underset{n^{-3}}{=} \mathcal{O}\left(\frac{\mathit{\Phi}_\mu^{7/11}\sqrt{\mu}\log^2 n}{n}\right),$$

and if $E[\mathit{\Phi}_\mu] \geq \frac{1}{\log n}\left(\frac{\mu}{\log^2 n}\right)^5$, then

$$|M_\mu - F_\mu| \underset{0.5n^{-3}}{=} \mathcal{O}\left(\frac{\mu\sqrt{\mathbb{E}[\mathit{\Phi}_\mu]}\log^2 n}{n} + \frac{\sqrt{\mathbb{E}[\mathit{\Phi}_\mu]\mu}\log^2 n}{n}\right) \underset{n^{-3}}{=} \mathcal{O}\left(\frac{\mu\sqrt{\mathit{\Phi}_\mu}\log^2 n}{n}\right).$$

Using the above mentioned modified versions of Lemmas 11, 26 and Lemma 10, it can be easily shown that the lemma is true for $\mu \geq 1$.

By Lemma 11, $|F_0'^{\text{un}} - M_0| \underset{n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{\sqrt{\mathit{\Phi}_1}}{n}\right)$. By the Chernoff bound with probability $\geq 1 - e^{-n/4}$, $\mathit{\Phi}_1 \leq n' \leq 2n$. Hence, $|F_0'^{\text{un}} - M_0| \underset{4n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$. Note that the error probabilities are not optimized. ■

B.9. Proof of Lemma 17

By triangle inequality, $|N-1| = |\sum_\mu F_\mu'^{\text{un}} - M_\mu| \leq \sum_\mu |F_\mu'^{\text{un}} - M_\mu|$. By Lemma 16, for $\mu = 0$, $|M_0 - F_0'^{\text{un}}| \underset{4n^{-3}}{=} \widetilde{\mathcal{O}}\left(n^{-1/2}\right)$. We now use Lemma 16 to bound $|F_\mu'^{\text{un}} - M_\mu|$ for $\mu \geq 1$. Since $\sum_\mu \mu\mathit{\Phi}_\mu = n'$ is a Poisson random variable with mean n, $\Pr(\sum_\mu \mu\mathit{\Phi}_\mu \leq 2n) \geq 1 - e^{-n/4}$. For $\mu \geq 1$, applying Cauchy Schwarz inequality repeatedly with the above constraints we get

$$\begin{aligned}
|N-1| &\underset{10n^{-2}}{=} \sum_{\mu=1}^{2n} \mathcal{O}\left(\min\left(\frac{\mathit{\Phi}_\mu^{7/11}\sqrt{\mu}}{n}, \frac{\sqrt{\mathit{\Phi}_\mu}\mu}{n}\right)\text{polylog}(n)\right) \\
&= \sum_{\mu=1}^{2n} \widetilde{\mathcal{O}}\left(\frac{\sqrt{\mu}}{n}\mathit{\Phi}_\mu^{7/11}\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\sum_{\mu=1}^{2n}\frac{\mu\mathit{\Phi}_\mu}{n}\sum_{\mu=1}^{2n}\frac{\mathit{\Phi}_\mu^{3/11}}{n}}\right) \overset{(a)}{=} \widetilde{\mathcal{O}}\left(\sqrt{\sum_{\mu=1}^{2n}\frac{\mathit{\Phi}_\mu^{1/2}}{n}}\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\sqrt{\sum_{\mu=1}^{2n}\frac{\mathit{\Phi}_\mu\mu}{n}\sum_{\mu=1}^{2n}\frac{1}{n\mu}}}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/4}}\right).
\end{aligned}$$

$\mathit{\Phi}_\mu$ takes only integer values, hence (a). Note that by the union bound, the error probability is bounded by

$$\Pr\left(\sum_\mu \mu\mathit{\Phi}_\mu > 2n\right) + \sum_{\mu=0}^{2n}\Pr\left(|M_\mu - F_\mu'^{\text{un}}| \neq \widetilde{\mathcal{O}}\left(\min\left(\frac{\mathit{\Phi}_\mu^{7/11}\sqrt{\mu}}{n}, \frac{\sqrt{\mathit{\Phi}_\mu}\mu}{n}\right)\right)\right).$$

By the concentration of Poisson random variables (discussed above) the first term is $\le e^{-n/4}$. By Lemma 16, the second term is $2n(4n^{-3})$. Hence the error probability is bounded by $e^{-n/4} + 2n(4n^{-3}) \le 10n^{-2}$. ■

B.10. Proof of Theorem 2

It is easy to show that if $\Phi_\mu > \log^2 n$, with probability $\ge 1 - n^{-3}$, $\max(G_\mu, 1/n) = G_\mu$ and $\min(\max(F_\mu, n^{-3})1) = F_\mu$. For the clarity of proofs we ignore these modifications and add an additional error probability of n^{-3}.

Recall that $F'_\mu = \frac{F'^{\text{un}}_\mu}{N}$. By Jensen's inequality, $\sum_\mu M_\mu \log \frac{M_\mu}{F'_\mu} \le \log \sum_\mu \frac{M_\mu^2}{F'_\mu}$. Furthermore $\sum_\mu \frac{M_\mu^2}{F'_\mu} = 1 + \frac{(M_\mu - F'_\mu)^2}{F'_\mu}$. Substituting $F'_\mu = F'^{\text{un}}_\mu / N$ and rearranging, we get

$$\sum_\mu \frac{(M_\mu - F'_\mu)^2}{F'_\mu} \le 2(N-1)^2 + \sum_\mu 2N \frac{(M_\mu - F'^{\text{un}}_\mu)^2}{F'^{\text{un}}_\mu}.$$

By Lemma 17, $N = 1 + \widetilde{\mathcal{O}}(n^{-1/4})$. Therefore,

$$\sum_\mu \frac{(M_\mu - F'_\mu)^2}{F'_\mu} = \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/2}}\right) + \sum_\mu \mathcal{O}\left(\frac{(M_\mu - F'^{\text{un}}_\mu)^2}{F'^{\text{un}}_\mu}\right).$$

To bound the second term in the above equation, we bound $|F'^{\text{un}}_\mu - M_\mu|$ and F'^{un}_μ separately. We first show that $F'^{\text{un}}_\mu \underset{n^{-3}}{=} \widetilde{\Omega}\left(\frac{\mu\Phi_\mu}{n}\right)$.

If empirical estimator is used for estimation, then $F'^{\text{un}}_\mu = \Phi_\mu \frac{\mu}{n}$. If Good-Turing or F_μ is used, then $\Phi_\mu \ge \log^2 n$. If $\mathbb{E}[\Phi_\mu] \le 0.5 \log^2 n$, then $\Pr(\Phi_\mu \ge \log^2 n) \le 0.5n^{-3}$. If $\mathbb{E}[\Phi_\mu] \ge 0.5 \log^2 n$, then using Claim 20 and Lemma 15 it can be shown that $F'^{\text{un}}_\mu \underset{0.5n^{-3}}{=} \widetilde{\Omega}\left(\frac{\mu\Phi_\mu}{n}\right)$. By the union bound, $F'^{\text{un}}_\mu \underset{n^{-3}}{=} \widetilde{\Omega}\left(\frac{\mu\Phi_\mu}{n}\right)$.

Now using bounds on $|F'^{\text{un}}_\mu - M_\mu|$ from Lemma 16 and the fact that $F'^{\text{un}}_\mu = \widetilde{\Omega}(\Phi_\mu \mu / n)$, we bound the KL divergence. Observe that $\sum_\mu \mu\Phi_\mu = n'$ is a Poisson random variable with mean n, therefore $\Pr(\sum_\mu \mu\Phi_\mu \le 2n) \ge 1 - e^{-n/4}$. Applying Cauchy Schwarz inequality repeatedly with the above constraint and using bounds on $|F'^{\text{un}}_\mu - M_\mu|$ (Lemma 16) and F'_μ we get

$$\begin{aligned}
\sum_{\mu=1}^{2n} \frac{(M_\mu - F'^{\text{un}}_\mu)^2}{F'^{\text{un}}_\mu} &\underset{2n(4n^{-3}+n^{-3})}{=} \sum_{\mu=1}^{2n} \mathcal{O}\left(\min\left(\frac{\mu}{n}, \frac{\Phi_\mu^{3/11}}{n}\right) \text{polylog}(n)\right) \\
&= \sum_{\mu=1}^{2n} \widetilde{\mathcal{O}}\left(\frac{\Phi_\mu^{1/2}}{n}\right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\sum_{\mu=1}^{2n} \frac{\Phi_\mu \mu}{n} \sum_{\mu=1}^{2n} \frac{1}{n\mu}}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{n^{1/2}}\right).
\end{aligned}$$

For $\mu = 0$, by Lemma 11, $(M_0 - F_0'^{\text{un}})^2 \underset{n^{-3}}{=} \mathcal{O}\left(\frac{\Phi_1 \text{polylog}(n)}{n^2} + \frac{\text{polylog}(n)}{n^2}\right)$ and hence,

$$\frac{(M_0 - F_0'^{\text{un}})^2}{F_0'^{\text{un}}} \underset{n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

Similar to the proof of Lemma 17, by the union bound the error probability is $\le e^{-n/4} + 10n^{-2} + 2n(4n^{-3} + n^{-3}) + n^{-3} + n^{-3} \le 22n^{-2} \le e^{-1}n^{-1.5}$ for $n \ge 4000$. Hence with Poi(n) samples, error probability is $\le e^{-1}n^{-1.5}$. Therefore by Lemma 8, with exactly n samples, error probability is $\le n^{-1}$. ■

B.11. Lower bounds on Good-Turing and empirical estimates

We prove that the following distribution achieves the lower bound in Lemma 1.

Let p be a distribution with $\frac{\sqrt{n}}{\log^3 n}$ symbols with probability $p_i \overset{\text{def}}{=} \frac{n^{1/3}\log^3 n}{cn} + i\frac{n^{1/6}\log^3 n}{cn}$ for $1 \le i \le n^{1/6}$. c is chosen such that the sum of probabilities adds up to 1. We provide a proof sketch and the detailed proof is deferred to the full version of the paper. It can be shown that p has the following properties.

- Let $\mathcal{R} \overset{\text{def}}{=} \cup_{i=1}^{n^{1/6}}[np_i + n^{1/6}, np_i + 2n^{1/6}]$. For every $\mu \in \mathcal{R}$, $\mathbb{E}[\Phi_\mu] = \widetilde{\Theta}(n^{1/3})$.
- Since the probabilities are $\widetilde{\Theta}\left(\frac{n^{1/3}}{n}\right)$, symbols occur with multiplicity $\widetilde{\Theta}(n^{1/3})$ with high probability.
- The distribution is chosen such that both empirical and Good-Turing bounds in Lemmas 10 and 11 are tight.

Hence for each $\mu \in \mathcal{R}$, both the Good-Turing and empirical estimators makes an error of

$$\widetilde{\Omega}\left(\frac{\mu\sqrt{\mathbb{E}[\Phi_\mu]}}{n}\right) = \widetilde{\Omega}\left(\frac{\sqrt{\mu}\mathbb{E}[\Phi_\mu]}{n}\right) = \widetilde{\Omega}\left(\frac{\sqrt{n^{1/3}}n^{1/3}}{n}\right) = \widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right).$$

Number of multiplicities in the range $\mathcal{R}$ is $n^{1/6} \cdot n^{1/6} = n^{1/3}$. Adding the error over all the multiplicities yields an total error of $\widetilde{\Omega}\left(\frac{1}{n^{1/2}}\right) \cdot n^{1/3} = \widetilde{\Omega}\left(\frac{1}{n^{1/6}}\right)$. ■.

Appendix C. Prediction

In this section, we prove Corollary 4. By definition $\Pr(\Psi^n) \overset{\text{def}}{=} \sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)$. Let ψ appear μ times in Ψ^n. Using the fact that sampling is *i.i.d.*, and the definition of pattern, each of the Φ_μ integers (in the pattern) are equally likely to appear as Ψ_{n+1}. This leads to,

$$P(\Psi^{n+1}, \Psi_{n+1} = \psi) = \sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)\frac{M_\mu(x^n)}{\Phi_\mu},$$

and hence

$$\Pr(\Psi_{n+1}|\Psi^n) = \frac{\sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)\frac{M_\mu(x^n)}{\Phi_\mu}}{\sum_{x^n|\Psi(x^n)=\Psi^n} \Pr(x^n)}.$$

Proof [Corollary 4] Any label-invariant estimator including the proposed estimator assigns identical values for F'_μ to all sequences with the same pattern. Hence

$$
\begin{aligned}
\mathbb{E}\left[\sum_\mu M_\mu \log \frac{M_\mu}{F'_\mu}\right] &= \sum_{x^n} p(x^n) \sum_\mu M_\mu(x^n) \log \frac{M_\mu(x^n)}{F'_\mu(x^n)} \\
&= \sum_{\Psi^n} \sum_\mu \sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) M_\mu(x^n) \log \frac{p(x^n)M_\mu(x^n)}{p(x^n)F'_\mu(x^n)} \\
&\overset{(a)}{\ge} \sum_{\Psi^n} \sum_\mu \Big(\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) M_\mu(x^n) \Big) \log \frac{\Big(\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n) M_\mu(x^n)\Big)}{\Big((\sum_{x^n|\Psi(x^n)=\Psi^n} p(x^n))F'_\mu(x^n)\Big)} \\
&= \sum_{\Psi^n} \sum_\mu \Big(P(\Psi^n) P(\Psi_{n+1}|\Psi^n) \Big) \log \frac{P(\Psi^{n+1})}{P(\Psi^n)F'_\mu} \\
&= \mathbb{E}_{\Psi^n \sim P}\left[\sum_{\Psi_{n+1}=1}^{m+1} P(\Psi_{n+1}|\Psi^n) \log \left(\frac{P(\Psi_{n+1}|\Psi^n)}{q(\Psi_{n+1}|\Psi^n)} \right) \right],
\end{aligned}
$$

where in (a) we used the log-sum inequality and the fact that our estimator F'_μ is identical for all sequences with the same pattern. ∎

Appendix D. Label invariant classification

In this section, we extend the combined-probability estimator to joint-sequences and propose a competitive classifier. First introduce profiles, a sufficient statistic for label-invariant classifiers. Then we relate the problem of classification to that of estimation in joint sequences. Motivated by the techniques in probability estimation, we then develop a joint-sequence probability estimator and prove its convergence rate, thus proving an upper bound on the error of the proposed classifier. Finally we prove a non-tight lower bound of $\widetilde{\Omega}(n^{-1/3})$.

D.1. Joint-profiles

Let the training sequences be X^n and Y^n and the test sequence be Z^1. It is easy to see that a sufficient statistic for label invariant classifiers is the *joint profile* φ of X^n, Y^n, Z^1, that counts how many elements appeared any given number of times in the three sequences (Acharya et al., 2011). For example, for $\overline{X} = aabcd$, $\overline{Y} = bacde$ and $Z = a$, the profiles are $\varphi(\overline{X}, \overline{Y}) = \{(2,1),(1,1),(1,1),(0,1)\}$ and $\varphi(\overline{X}, \overline{Y}, Z) = \{(2,1,1),(1,1,0),(1,1,0),(1,1,0),(0,1,0)\}$. $\varphi(\overline{X}, \overline{Y})$ indicates that there is one symbol appearing twice in first sequence and once in second, two symbols appearing once in both and so on. The profiles for three sequences can be understood similarly. Any label invariant test is only a function of the joint profile.

By definition, the probability of a profile is the sum of the probabilities of all sequences with that profile *i.e.,* for profiles of $(\overline{x}, \overline{y}, z)$, $\Pr(\varphi) = \sum_{\overline{x},\overline{y},z|\varphi(\overline{x},\overline{y},z)} \Pr(\overline{x}, \overline{y}, z)$. $\Pr(\varphi)$ is difficult to compute due to the permutations involved. Various techniques to compute profile probabilities are studied in Acharya et al. (2010). Still the proposed classifier we derive runs in linear time.

D.2. Classification via estimation

Let $\mu_x(\overline{x},\overline{y})$ denote the number of multiplicities symbol x in $(\overline{x},\overline{y})$. Let

$$M^p_{\mu,\mu'}(\overline{x},\overline{y}) \overset{\text{def}}{=} \sum_{x:\mu_x(\overline{x},\overline{y})=(\mu,\mu')} p_x$$

be the sum of the probabilities of all elements in p such that $\mu_x(\overline{x},\overline{y}) = (\mu,\mu')$. $M^q_{\mu,\mu'}(\overline{x},\overline{y})$ is defined similarly.

Let $\varphi = \varphi(\overline{x},\overline{y})$ be the joint profile of $(\overline{x},\overline{y})$. If z is generated according to p, then the probability of observing the joint profile $\varphi(\overline{x},\overline{y},z)$, where z is an element appearing μ and μ' times respectively in $\overline{x}$ and $\overline{y}$ is

$$\begin{aligned}\Pr{}^p(\varphi(\overline{x},\overline{y},z)) &= \sum_{\overline{x},\overline{y}|\varphi(\overline{x},\overline{y})=\varphi} P(\overline{x})Q(\overline{y})M^p_{\mu,\mu'}(\overline{x},\overline{y}),\\ &= \Pr(\varphi(\overline{x},\overline{y}))\mathbb{E}_\varphi[M^p_{\mu,\mu'}],\end{aligned}$$

where $\mathbb{E}_\varphi[M^p_{\mu,\mu'}] \overset{\text{def}}{=} \mathbb{E}[M^p_{\mu,\mu'}|\varPhi = \varphi]$ is the expected value of $M^p_{\mu,\mu'}$ given that φ is the profile.

When the two distributions are known and the observed joint profile is $\varphi(\overline{x},\overline{y},z)$, then the classification problem becomes a hypothesis testing problem. The optimal solution to the hypothesis testing when both hypotheses are equally likely is the one that assigns higher probability to the observation (joint profile in our case). So the optimal classifier is

$$\begin{aligned}&\Pr{}^p(\varphi(\overline{x},\overline{y},z)) \underset{q}{\overset{p}{\gtrless}} \Pr{}^q(\varphi(\overline{x},\overline{y},z))\\ \Rightarrow\qquad &\mathbb{E}_\varphi[M^p_{\mu,\mu'}] \underset{q}{\overset{p}{\gtrless}} \mathbb{E}_\varphi[M^q_{\mu,\mu'}].\end{aligned}$$

We will develop variants of F'_μ for joint profiles, denoted by $F'^p_{\mu,\mu'}$, and $F'^q_{\mu,\mu'}$. We use these estimators in place of the expected values. Our classifier S assigns z to $\overline{x}$ if $F'^p_{\mu,\mu'} > F'^q_{\mu,\mu'}$ and to $\overline{y}$ if $F'^p_{\mu,\mu'} < F'^q_{\mu,\mu'}$. Ties are broken at random. There is an additional error in classification with respect to the optimal label-invariant classifier when $\mathbb{E}_\varphi[M^p_{\mu,\mu'}] < \mathbb{E}_\varphi[M^q_{\mu,\mu'}]$ but $F'^p_{\mu,\mu'} \geq F'^q_{\mu,\mu'}$ or vice versa.

Let $\mathbb{1}^\epsilon_{\mu,\mu'}$ be an indicator random variable that is 1 if

$$|\mathbb{E}_\varphi[M^p_{\mu,\mu'}] - \mathbb{E}_\varphi[M^q_{\mu,\mu'}]| \leq \sum_{s\in\{p,q\}} |F'^s_{\mu,\mu'} - \mathbb{E}_\varphi[M^s_{\mu,\mu'}]|. \tag{7}$$

It is easy to see that if there is an additional error, then $\mathbb{1}^\epsilon_{\mu,\mu'} = 1$. Using these conditions the following lemma provides a bound on the additional error with respect to the optimal.

Lemma 27 (Classification via estimation) *For every (p,q) and every classifier S,*

$$\mathcal{E}^S_{p,q}(n) \leq \mathcal{E}^{S_{p,q}}_{p,q}(n) + \sum_{\mu,\mu'}\sum_{t\in\{p,q\}} \mathbb{E}[\mathbb{1}^\epsilon_{\mu,\mu'}|F'^t_{\mu,\mu'} - M^t_{\mu,\mu'}|].$$

Proof For a joint profile φ, S assigns z to the wrong hypothesis, if $F'^{p}_{\mu,\mu'} > F'^{q}_{\mu,\mu'}$ and $\mathbb{E}_{\varphi}[M^{p}_{\mu,\mu'}] < \mathbb{E}_{\varphi}[M^{q}_{\mu,\mu'}]$ or vice versa. Hence $\mathbb{1}^{\epsilon}_{\mu,\mu'} = 1$. If $\mathbb{1}^{\epsilon}_{\mu,\mu'} = 1$, then the increase in error is $\Pr(\varphi)\mathbb{1}^{\epsilon}_{\mu,\mu'}|\mathbb{E}_{\varphi}[M^{p}_{\mu,\mu'}]-\mathbb{E}_{\varphi}[M^{q}_{\mu,\mu'}]|$. Using Equation (7) and summing over all profiles results in the lemma. ■

In the next section we develop estimators for $M^{p}_{\mu,\mu'}$ and $M^{q}_{\mu,\mu'}$.

D.3. Conventional estimation and the proposed approach

Empirical and Good-Turing estimators can be naturally extended to joint sequences as $E^{p}_{\mu,\mu'} \overset{\text{def}}{=} \varPhi_{\mu,\mu'}\frac{\mu}{n}$ and $G^{p}_{\mu,\mu'} \overset{\text{def}}{=} \varPhi_{\mu+1,\mu'}\frac{\mu+1}{n}$. As with probability estimation, it is easy to come up with examples where the rate of convergence of these estimates is not optimal. The rate of convergence of Good-Turing and empirical estimators are quantified in the next lemma.

Lemma 28 (Empirical and Good-Turing for joint sequences) *For every* (p, q) *and* μ *and* μ',

$$\left|M^{p}_{\mu,\mu'} - G^{p}_{\mu,\mu'}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\sqrt{\mathbb{E}[\varPhi_{\mu+1,\mu'}]+1}\,\frac{(\mu+1)\log^{2} n}{n}\right),$$

and if $\max(\mu,\mu') > 0$, *then*

$$\left|M^{p}_{\mu,\mu'} - E^{p}_{\mu,\mu'}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\varPhi_{\mu,\mu'}\frac{\sqrt{\mu+1}\log n}{n}\right).$$

Similar results hold for $M^{q}_{\mu,\mu'}$.

The proof of the above lemma is similar to those of Lemmas 10 and 11 and hence omitted. Note that the error probability in the above lemma can be any polynomial in $1/n$. n^{-4} has been chosen to simplify the analysis. Motivated by combined probability estimation, we propose $F^{p}_{\mu_0,\mu'_0}$ for joint sequences as

$$F^{p}_{\mu_0,\mu'_0} = \varPhi_{\mu_0,\mu'_0}\frac{\mu_0+1}{n}\frac{\widehat{\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]}}{\widehat{\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]}},$$

where $\widehat{\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]}$ and $\widehat{\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]}$ are estimators for $\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]$ and $\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]$ respectively. Let $\mathcal{S}^{\mu_0,\mu'_0}_{r} = \{(\mu,\mu') \mid |\mu-\mu_0| \le r, |\mu'-\mu'_0| \le r\}$ and $r_{\mu_0} = \left\lfloor \frac{\sqrt{\mu_0}}{(\mu_0\varPhi_{\mu_0,\mu'_0})^{1/12}\log n} \right\rfloor$. The estimators $\widehat{\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]}$ and $\widehat{\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]}$ are given by

$$\widehat{\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]} = \sum_{\mu,\mu' \in \mathcal{S}^{\mu_0,\mu'_0}_{r_{\mu_0}}} c_{\mu,\mu'}\varPhi_{\mu,\mu'}, \text{ and } \widehat{\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]} = \sum_{\mu,\mu' \in \mathcal{S}^{\mu_0+1,\mu'_0}_{r_{\mu_0}}} d_{\mu,\mu'}\varPhi_{\mu,\mu'},$$

where $c_{\mu,\mu'} = \gamma_{r_{\mu_0}}(|\mu-\mu_0|)\gamma_{r_{\mu_0}}(|\mu'-\mu'_0|)a^{\mu_0}_{\mu}a^{\mu'_0}_{\mu'}$ and $d_{\mu,\mu'} = \gamma_{r_{\mu_0}}(|\mu-\mu_0-1|)\gamma_{r_{\mu_0}}(|\mu'-\mu'_0|)\,\frac{\mu_0}{\mu_0+1}a^{\mu_0}_{\mu}a^{\mu'_0}_{\mu'}$. γ_r and $a^{\mu_0}_{\mu}$ are defined in Section 5. The estimator$F^{q}_{\mu_0,\mu'_0}$ can be obtained similarly.

The next lemma shows that the estimate for the ratio of $\mathbb{E}[\varPhi_{\mu_0+1,\mu'_0}]$ and $\mathbb{E}[\varPhi_{\mu_0,\mu'_0}]$ is close to the actual ratio. The proof is similar to that of Lemma 15 and hence omitted.

Lemma 29 *For every (p,q) and every $\mu_0 \geq \log^2 n$, if $\frac{1}{\mu_0}\left(\frac{\mu_0}{\log^2 n}\right)^6 \geq \mathbb{E}[\Phi_{\mu_0,\mu'_0}] \geq \log^2 n$, then*

$$\left|\frac{\mathbb{E}[\widehat{\Phi_{\mu_0+1,\mu'_0}}]}{\mathbb{E}[\widehat{\Phi_{\mu_0,\mu'_0}}]} - \frac{\mathbb{E}[\Phi_{\mu_0+1,\mu'_0}]}{\mathbb{E}[\Phi_{\mu_0,\mu'_0}]}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\log^3 n}{\sqrt{\mu_0}(\mathbb{E}[\Phi_{\mu_0,\mu'_0}]\mu_0)^{1/3}}\right),$$

and if $\mathbb{E}[\Phi_{\mu_0,\mu'_0}] \geq \frac{1}{\mu_0}\left(\frac{\mu_0}{\log^2 n}\right)^6$, then

$$\left|\frac{\mathbb{E}[\widehat{\Phi_{\mu_0+1,\mu'_0}}]}{\mathbb{E}[\widehat{\Phi_{\mu_0,\mu'_0}}]} - \frac{\mathbb{E}[\Phi_{\mu_0+1,\mu'_0}]}{\mathbb{E}[\Phi_{\mu_0,\mu'_0}]}\right| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\log^3 n}{\sqrt{\mathbb{E}[\Phi_{\mu_0,\mu'_0}]}}\right).$$

Using the previous lemma, we bound the error of $F^p_{\mu,\mu'}$ in the next lemma. The proof is similar to that of Lemma 26 and hence omitted.

Lemma 30 *For every (p,q) and $\mu \geq \log^2 n$, if $\frac{1}{\mu}\left(\frac{\mu}{\log^2 n}\right)^6 \geq \mathbb{E}[\Phi_{\mu,\mu'}] \geq \log^2 n$, then*

$$\left|M^p_{\mu,\mu'} - F^p_{\mu,\mu'}\right| \underset{2n^{-4}}{=} \mathcal{O}\left(\frac{(\mathbb{E}[\Phi_{\mu,\mu'}]^{2/3}\mu^{1/6}\log^3 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_{\mu,\mu'}]\mu}\log^2 n}{n}\right),$$

and if $\mathbb{E}[\Phi_{\mu,\mu'}] > \frac{1}{\mu}\left(\frac{\mu}{\log^3 n}\right)^6$, then

$$\left|M^p_{\mu,\mu'} - F^p_{\mu,\mu'}\right| \underset{2n^{-4}}{=} \mathcal{O}\left(\frac{\mu\sqrt{\mathbb{E}[\Phi_{\mu,\mu'}]}\log^3 n}{n} + \frac{\sqrt{\mathbb{E}[\Phi_{\mu,\mu'}]\mu}\log^2 n}{n}\right).$$

Similar results hold for $M^q_{\mu,\mu'}$.

D.4. Competitive classifier

The proposed classifier is given below. It estimates $M^p_{\mu,\mu'}$ (call it $F'^p_{\mu,\mu'}$) and $M^q_{\mu,\mu'}$ (call it $F'^q_{\mu,\mu'}$) and assigns z to the hypothesis that has the higher estimate. Let μ and μ' be the multiplicities of the z in $\overline{x}$ and $\overline{y}$ respectively. If $|\mu - \mu'| \geq \sqrt{\mu+\mu'}\log^2 n$, then the classifier uses empirical estimates. Since μ and μ' are far apart, by the Chernoff bound such an estimate provides us good bounds for the purposes of classification. In other cases, it uses the estimate with the lowest error bounds, given by Lemma 28 for $E^p_{\mu,\mu'}, G^p_{\mu,\mu'}$, and Lemma 30 for $F^p_{\mu,\mu'}$. We also set $F'^p_{\mu,\mu'} = \min(F'^p_{\mu,\mu'}, 1)$ and $F'^q_{\mu,\mu'} = \min(F'^q_{\mu,\mu'}, 1)$, to help in the analysis and ensure that the estimates are always ≤ 1.

Classifier $S(\overline{x}, \overline{y}, z)$
Input: Two sequences $\overline{x}$ and $\overline{y}$ and a symbol z.
Output: `x` or `y`.

1. Let $\mu = \mu_z(\overline{x})$ and $\mu' = \mu_z(\overline{y})$.
2. If $\max(\mu,\mu') = 0$, then $F'^p_{\mu,\mu'} = G^p_{\mu,\mu'}$ and $F'^q_{\mu,\mu'} = G^q_{\mu,\mu'}$.

3. If $\max(\mu,\mu') > 0$ and $|\mu-\mu'| \ge \sqrt{\mu+\mu'}\log^2 n$ or $\mathit{\Phi}_{\mu,\mu'} \le \log^2 n$, then $F'^p_{\mu,\mu'} = E^p_{\mu,\mu'}$ and $F'^q_{\mu,\mu'} = E^q_{\mu,\mu'}$.
4. If $\max(\mu,\mu') > 0$, $|\mu-\mu'| < \sqrt{\mu+\mu'}\log^2 n$, and $\mathit{\Phi}_{\mu,\mu'} > \log^2 n$, then
   (a) If $\mu \ge 4\log^4 n$, then $F'^p_{\mu,\mu'} = F^p_{\mu,\mu'}$ and $F'^q_{\mu,\mu'} = F^q_{\mu,\mu'}$.
   (b) If $\mu < 4\log^4 n$, then $F'^p_{\mu,\mu'} = G^p_{\mu,\mu'}$ and $F'^q_{\mu,\mu'} = G^q_{\mu,\mu'}$.
5. Set $F'^p_{\mu,\mu'} = \min(F'^p_{\mu,\mu'},1)$ and $F'^q_{\mu,\mu'} = \min(F'^q_{\mu,\mu'},1)$.
6. If $F'^p_{\mu,\mu'} > F'^q_{\mu,\mu'}$, then return x. If $F'^p_{\mu,\mu'} < F'^q_{\mu,\mu'}$, then return y. If $F'^p_{\mu,\mu'} = F'^q_{\mu,\mu'}$ return x or y with equal probability.

D.5. Proof of Theorem 6

The analysis of the classifier is similar to that of the combined probability estimation, and we outline few key steps. The error in estimating $M^p_{\mu,\mu'}$ (and $M^q_{\mu,\mu'}$) is quantified in the following lemma.

Lemma 31 *For every (p,q), $|M^p_{0,0} - F'^p_{0,0}| \underset{10n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$ and for $(\mu,\mu') \ne (0,0)$ and $|\mu-\mu'| \le \sqrt{\mu+\mu'}\log^2 n$,*

$$|M^p_{\mu,\mu'} - F'^p_{\mu,\mu'}| \underset{10n^{-3}}{=} \widetilde{\mathcal{O}}\left(\frac{\min\left(\mathit{\Phi}^{2/3}_{\mu,\mu'}\sqrt{\mu+1}, \mathit{\Phi}^{1/2}_{\mu,\mu'}(\mu+1)\right)}{n}\right).$$

Similar results hold for $M^q_{\mu,\mu'}$.

The analysis of the lemma is similar to that of Lemma 16 and hence omitted. We now prove Theorem 6 using the above set of results.

Proof [Theorem 6] Let $\mathcal{R} = \{(\mu,\mu') \mid |\mu-\mu'| \le \sqrt{\mu+\mu'}\log^2 n\}$. By Lemma 27,

$$\mathcal{E}^S_{p,q}(n) \le \mathcal{E}^{S_{p,q}}_{p,q}(n) + 2\max_p\left(\sum_{(\mu,\mu')\in\mathcal{R}} \mathbb{E}[\mathbb{1}^\epsilon_{\mu,\mu'}|F'^p_{\mu,\mu'} - M^p_{\mu,\mu'}|] + \sum_{(\mu,\mu')\in\mathcal{R}^c} \mathbb{E}[\mathbb{1}^\epsilon_{\mu,\mu'}|F'^p_{\mu,\mu'} - M^p_{\mu,\mu'}|]\right).$$

We first show that the second term is $\mathcal{O}(n^{-1.5})$. By Lemma 28,

$$|M^p_{\mu,\mu'} - E^p_{\mu,\mu'}| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\mathit{\Phi}_{\mu,\mu'}\sqrt{\mu}\log n}{n}\right) \quad \text{and} \quad |M^q_{\mu,\mu'} - E^q_{\mu,\mu'}| \underset{n^{-4}}{=} \mathcal{O}\left(\frac{\mathit{\Phi}_{\mu,\mu'}\sqrt{\mu'}\log n}{n}\right).$$

If $|\mu-\mu'| \ge \sqrt{\mu+\mu'}\log^2 n$, then

$$|M^p_{\mu,\mu'} - M^q_{\mu,\mu'}| \ge \frac{\mathit{\Phi}_{\mu,\mu'}\sqrt{\mu+\mu'}\log^2 n}{n}.$$

Hence $\mathbb{1}^\epsilon_{\mu,\mu'} \underset{2n^{-4}}{=} 0$. Since with Poi$(n)$ samples, the bounds hold with probability $1-\mathcal{O}(n^{-4})$, by Lemma 8, with exactly n samples, they hold with probability $1-\mathcal{O}(n^{-3.5})$. Observe that (μ, μ') takes at most $n \cdot n = n^2$ values. Therefore, by the union bound $\Pr(\mathbb{1}^{\epsilon}_{\mu,\mu'} = 1) \leq \mathcal{O}(n^{-1.5})$. Hence $\max_p \sum_{(\mu,\mu') \in \mathcal{R}^c} \mathbb{E}[|F'^{p}_{\mu,\mu'} - M^p_{\mu,\mu'}|] = \mathcal{O}(n^{-1.5})$.

We now consider the case $(\mu, \mu') \in \mathcal{R}$. In Lemma 31, the bounds on $|F'^{p}_{\mu,\mu'} - M^p_{\mu,\mu'}|$ hold with probability $\geq 1 - \mathcal{O}(n^{-3})$, with Poi$(n)$ samples. Therefore by Lemma 8, with exactly n samples, they hold with probability $\geq 1 - \mathcal{O}(n^{-2.5})$, *i.e.*, $|F'^{p}_{\mu,\mu'} - M^p_{\mu,\mu'}| \underset{\mathcal{O}(n^{-2.5})}{=} \widetilde{\mathcal{O}}\left(\frac{\Phi^{2/3}_{\mu,\mu'}(\mu+\mu')^{1/2}}{n}\right)$. Observe that (μ, μ') takes at most $n \cdot n = n^2$ values, hence by the union bound, the probability that the above bound holds for all $(\mu, \mu') \in \mathcal{R}$ is at least $1 - \mathcal{O}(n^{-0.5})$. Since $|F'^{p}_{\mu,\mu'} - M^p_{\mu,\mu'}| \leq 1$, we get

$$\max_p \sum_{(\mu,\mu') \in \mathcal{R}} \mathbb{E}[|F'^{p}_{\mu,\mu'} - M^p_{\mu,\mu'}|] \leq \sum_{(\mu,\mu') \in \mathcal{R}} \widetilde{\mathcal{O}}\left(\frac{\Phi^{2/3}_{\mu,\mu'}(\mu+\mu')^{1/2}}{n}\right) + \mathcal{O}\left(\frac{1}{n^{1/2}}\right).$$

Using techniques similar to those in the proofs Lemma 17 and Theorem 2, it can be shown that the above quantity is $\leq \widetilde{\mathcal{O}}(n^{-1/5})$, thus proving the theorem. ■

D.6. Lower bound for classification

We prove a non-tight converse for the additional error in this section.

Theorem 32 *For any classifier S there exists (p, q) such that*

$$\mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}\left(\frac{1}{n^{1/3}}\right).$$

We construct a distribution q and a collection of distributions $\mathcal{P}$ such that for any distribution $p \in \mathcal{P}$, the optimal label-invariant classification error for (p, q) is $\frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right)$. We then show that any label-invariant classifier incurs an additional error of $\widetilde{\Omega}(n^{-1/3})$ for at least one pair (p', q), where $p' \in \mathcal{P}$. Similar arguments have been used in LeCam (1986); Paninski (2008).

Let q be a distribution over $i = 1, 2, \ldots, \frac{n^{1/3}}{\log n}$ such that $q_i = \frac{3i^2 \log^3 n}{cn}$, and $c \leq 2$ is the normalization factor.

Let $\mathcal{P}$ to be a collection of $2^{\frac{n^{1/3}}{2\log n}}$ distributions. For every $p \in \mathcal{P}$, for all odd i, $p_i = q_i \pm \frac{i \log n}{n}$ and $p_{i+1} = q_{i+1} \mp \frac{i \log n}{n}$, such that, $p_i + p_{i+1} = q_i + q_{i+1}$. For every $p \in \mathcal{P}$. $||p - q||_1 = \Theta\left(\frac{1}{n^{1/3}\log n}\right)$. The next lemma, proved in the full version of the paper states that every distribution $p \in \mathcal{P}$ and q can be classified by a label-invariant classifier with error $\frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right)$.

Lemma 33 *For every $p \in \mathcal{P}$ and q,*

$$\mathcal{E}^{S_{p,q}}_{p,q}(n) = \frac{1}{2} - \Theta\left(\frac{1}{n^{1/3}\log n}\right).$$

Proof [sketch of Theorem 32] We show that for any classifier S, $\max_{p \in \mathcal{P}} \mathcal{E}^{S}_{p,q}(n) = \mathcal{E}^{S_{p,q}}_{p,q}(n) + \widetilde{\Omega}(n^{-1/3})$ for some $p \in \mathcal{P}$, thus proving the theorem. Since extra information reduces the error probability, we aid the classifier with a genie that associates the multiplicity with the probability of the symbol. Using ideas similar to LeCam (1986); Acharya et al. (2012a), one can show that the worst error probability of any classifier between q and the set of distribution $\mathcal{P}$ is lower bounded by error probability between q and any mixture on $\mathcal{P}$. We choose the mixture p_0 such that each $p \in \mathcal{P}$ is chosen uniformly at random. Therefore for any classifier S,

$$\max_{p} \mathcal{E}^{S}_{p,q}(n) \geq \sum_{\overline{x},\overline{y},z} \frac{\min\left(q(\overline{x})p_0(\overline{y},z), p_0(\overline{y})q(\overline{x},z)\right)}{2}.$$

Using techniques similar to Acharya et al. (2012a), it can be shown that difference between above error and $\mathcal{E}^{S_{p,q}}_{p,q}(n)$ is $\widetilde{\Omega}(n^{-1/3})$. The complete analysis is deferred to the full version of the paper. ■