Author: Ido Guy
Website: https://tinyurl.com/3cxthmys

Detailed Description of the Research Program: Learning Over Large Alphabets

1 Scientific Background

Let p be a probability distribution over a countable alphabet $\mathcal{X}$. Let X^n be a sample of n independent observations from p. In this work we study the basic problem of estimating p from X^n. That is, given a sample X^n, the goal is to find an estimator $\hat{p}(X^n)$ that is 'as close as possible' to p. We focus on the large alphabet regime, where the alphabet size m is much greater than (or at least comparable with) the number of samples n. Large alphabet modelling is of special interest in many domains. Its applications span many disciplines such as information retrieval [1], spelling correction [2], word-sense disambiguation [3], language modeling for speech recognition, [4], learning theory [5], and many others.

The missing mass is an essential building block for estimating the complete probability distribution. The missing mass is defined as the probability of all the symbols that do not appear in the sample. Perhaps the first major contribution to the missing mass problem dates back to Laplace in the 18^{th} century [6]. In his work, Laplace studied the *sunrise problem*; given that the sun raised every morning until today, what is the probability that it will rise tomorrow? Laplace addressed the problem of unseen symbols by adding a single count to all the symbols in the alphabet (including unobserved symbols). Then, the missing mass estimator is simply the empirical distribution of all the symbols with a single count. This method is known as the *rule of succession*.

Let $N_x(X^n)$ be the number of appearances of the symbol $x \in \mathcal{X}$ in X^n. Let

$$\Phi_j(X^n) = \sum_{x \in \mathcal{X}} \mathbb{1}(N_x(X^n) = j) \tag{1}$$

be the number of symbols that appear j times in X^n, for $0 \leq j \leq n$, where $\mathbb{1}(\cdot)$ is the indicator function. We denote the collection $\{\Phi_j(X^n)\}_{j=0}^n$ as the *frequency of frequencies* (FoF's).

The Laplace estimator was later generalized to a family of *add-constant* estimators. An add-c estimator assigns to a symbol that appeared j times a probability proportional to $j + c$, where c is a pre-defined constant. Then, the add-c missing mass estimator is

$$\hat{M}^{AC}(X^n) = \frac{c\Phi_0(X^n)}{n + ck}, \tag{2}$$

Add-constant estimators hold many desirable properties, mostly in terms of their simplicity and interpretability. From a Bayesian point of view, the add-c estimator corresponds to the expected value of the posterior distribution, using a symmetric Dirichlet distribution with parameter c as a prior distribution. They also hold several

asymptotic minimax properties. However, when the alphabet size m is large compared to the sample size n, add-constant estimators perform quite poorly [7].

In the 1940's, I.J. Good and A.M. Turing achieved a significant milestone in the study of the missing mass problem while trying to break the Enigma Cipher [8] during World War II. The celebrated Good-Turing (GT) estimator introduced a new approach to the missing mass problem. The idea behind their work was surprisingly simple. Instead of using $\Phi_0(X^n)$ as a statistic for the missing mass, they suggest using $\Phi_1(X^n)$, the number of symbols with a single appearance in the sample. Specifically, the GT missing mass estimator is defined as $\hat{M}^{GT}(X^n) = \Phi_1(X^n)/n$. It is important to emphasize that while add-constant estimators depend on the alphabet size m, the GT estimator does not assume any knowledge of m, which makes it more robust[1].

The GT and add-c estimators can be generalized to estimate the j-th mass for any j appearances (later defined in (3)). This can be used to estimate the entire distribution. However, they are known to be sub-optimal for greater j's. Consequently, several modifications have been proposed, including the Jelinek-Mercer, Katz, Witten-Bell and Kneser-Ney estimators [9, 4].

2 Research Objectives and Expected Significance

In this research we first study a generalized framework for missing mass estimation that depends on $\Phi_0(X^n)$ and $\Phi_1(X^n)$. In other words, while Laplace utilizes $\Phi_0(X^n)$ as a statistic for the missing mass, and GT focuses on $\Phi_1(X^n)$, we suggest a hybrid estimator which applies them both. We focus on minimax mean square error (MSE), where we minimize the maximal risk (in terms of MSE) over all distributions of a countable alphabet $\mathcal{X}$.

Going forward, we suggest a family of hybrid estimators that depend on the number of events with j and $j+1$ appearances, $\Phi_j(X^n)$, $\Phi_{j+1}(X^n)$. Given X^n, our goal is to estimate the j-th mass,

$$M_j(X^n) \triangleq \sum_{x\in\mathcal{X}} p(x)\mathbb{1}(N_x(X^n) = j). \tag{3}$$

The j-th mass is a random variable which refers to the sum of probabilities of symbols that were seen j times in a given set of samples.

As discussed above, the missing mass holds a key role in estimating the complete probability distribution. Therefore, the expected significance of this work stems from the importance of the probability estimation and the missing mass problems. By improving currently known estimators we expect a significant impact on a variety of scientific disciplines, as discussed in Section 1. Ultimately, our goals are:

- Derive efficient and robust missing mass estimators that improve upon currently known methods.
- Develop probability estimation schemes that utilize the enhanced performance of our proposed estimators.
- Introduce new risk bounds for the j-th mass estimation.
- Apply our proposed scheme to existing applications.

[1]There exists a variant of add-c estimators that is independent of m, which adds a single count to the entire collection of unseen symbols (as opposed to every unseen symbol). However, it is known to perform quite poorly for large alphabets [7].

3 Preliminary Results

As a first step, we study the minimax risk for an estimator which uses only $\Phi_0(X^n)$. Let

$$\hat{M}_\beta(X^n) = \beta\Phi_0(X^n) \tag{4}$$

be an estimator of the missing mass that is linear in $\Phi_0(X^n)$. We denote this as GL (Generalized Laplace) estimator. Notice that the Laplace estimator is a special case of (4), where $\beta = \frac{1}{n+m}$. The mean square error of this estimator satisfies

$$\mathbb{E}_{X^n\sim p}(\hat{M}_\beta(X^n) - M_0(X^n))^2 = \mathbb{E}_{X^n\sim p}\left(\sum_u \mathbb{1}\{N_u = 0\}p(u) - \sum_u \mathbb{1}\{N_u = 0\}\beta\right) \tag{5}$$

Theorem 1 below introduces an upper bound for (5), which is required for our following steps.

Theorem 1 *Let p be a probability distribution over an alphabet $\mathcal{X}$ of a known size. Let $0 \leq \beta \leq \frac{1}{m}$. The following holds for every p:*

$$\mathbb{E}_{X^n\sim p}(\hat{M}_\beta(X^n) - M(X^n))^2 \leq 2\beta(\frac{2m}{n} + 4) + \left(\sum_{u\in\mathcal{X}} e^{-np(u)}(p(u) - \beta)\right)^2 + \sum_{u\in\mathcal{X}} e^{-np(u)}(p(u) - \beta)^2 \tag{6}$$

The complete proof for this theorem is provided in the Supplemental Material, located on the author's web-page.

Theorem 1 introduces an upper bound that holds for a range of β values. This means we may seek a value for β that minimizes it, for any p over an alphabet $\mathcal{X}$ of a known size. Each term of (6) can be numerically bounded from above quite efficiently. Specifically, Theorem 2 in the Supplemental Material shows that the distribution which maximizes each term of (6) has no more than four degrees of freedom, regardless of n, m or β.

As a final step, for every β that satisfies the conditions of Theorem 1 we attain a missing mass estimation bound, that holds for every p. Thus, we examine different values, and seek the lowest risk bound. Algorithm 1 in the Supplemental Material provides a pseudo-code for our suggested scheme. Notice that eventually, we obtain a value β, and a corresponding bound for the missing mass estimation risk, for every n and m, as desired.

To find a lower bound, we consider two distinct probability distributions - the uniform distribution and the degenerate distribution. For both distributions it is simple to explicitly evaluate the missing mass for a known m, so explicitly calculating the risk (5) is also straightforward.

Let us now demonstrate our suggested bounds. Let us fix the number of samples $n = 100$, and evaluate the risk for a growing alphabet size m. The left panel of Figure 1 illustrates two different upper bounds (blue and yellow curves) and a lower bound (red curve). The blue curve corresponds to our suggested scheme. The yellow curve is the Laplace estimator upper bound. The red curve is an approximate lower bound for estimators that are linear in Φ_0. This tight bound gives confidence that further attempts to devise such an estimator would yield very little improvements. We are currently working on an analytical expression for the β value as function of n and m.

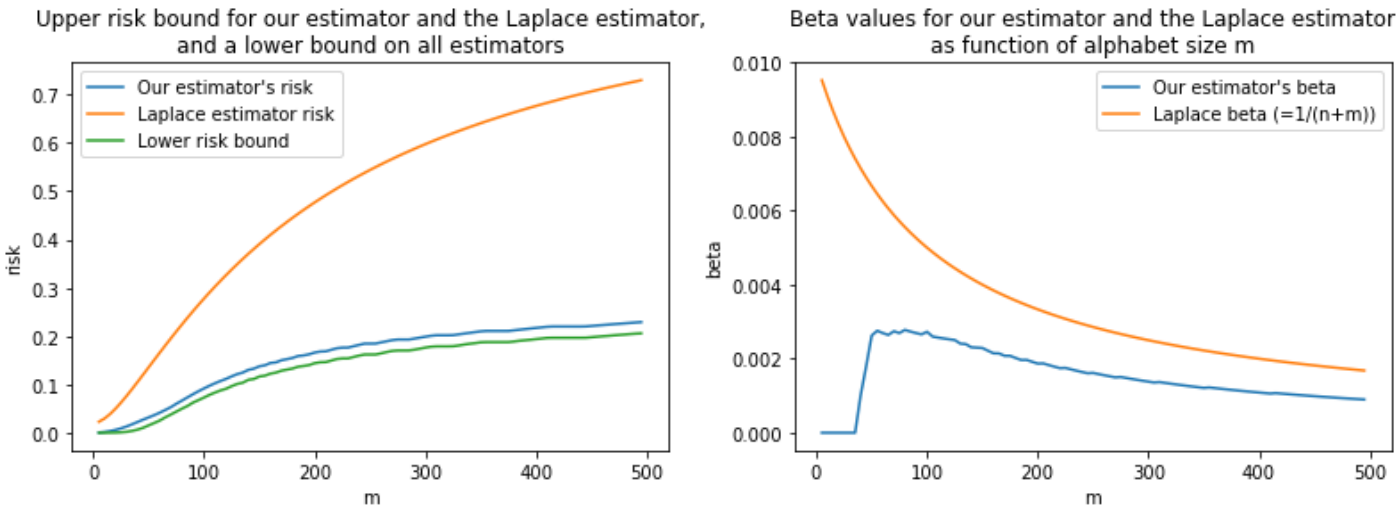

Figure 1: Left: Missing mass estimation risk for $n = 100$ and an increasing m, including a lower bound. Right: β values in our experiment compared to Laplace estimator β values.

We observe that even for a relatively small m, The GL risk introduces a significant improvement over the Laplace risk. As m grows, we notice that the difference between the GL bound and the Laplace bound increases.

The right panel of Figure 1 illustrates β for every m, as described in Algorithm 1. The chart corresponds to $n = 100$. For comparison, we add the Laplace estimator's β value, $\frac{1}{n+m}$. We observe that our β is zero for smaller m values, and then introduces a polynomial decay. This behavior is fundamentally different than add-c estimators.

4 Expected Results

In Section 3 we introduce the GL estimator for missing mass estimation. Our preliminary results introduce new upper and lower risk-bounds that demonstrate improved performance guarantees, compared to add-c estimators.

Our current results are obtained by an estimator that only considers one FoF (specifically, $\Phi_0(X^n)$). We will extend our analysis and consider hybrid estimators using two FoF's. First, we will research a hybrid estimator using $\Phi_0(X^n)$ and $\Phi_1(X^n)$. The use of two FoF's should allow us to further improve the missing mass minimax bounds and the corresponding estimators.

Further, our preliminary results serve as a guideline for additional estimation tasks. This includes the j-th mass (3), and more importantly, the estimation of the complete probability distribution. It is important to emphasize that the probability estimation risk is typically measured in Kullback Leibler (KL) divergence or total variation (TV) [10], as opposed to MSE in the missing mass case. Therefore, our current results may not be immediately applied to this setup, and require additional analysis.

As mentioned above, probability estimation is a widely applied task in many scientific fields, including natural sciences, exact sciences and engineering. Therefore, introducing a new scheme that is both easy to use, and provably improves upon currently known methods, has quite a significant expected impact. This is one of the main motivations of the proposed research.

References

[1] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management*. ACM, 1999, pp. 316–321.

[2] K. W. Church and W. A. Gale, "Probability scoring for spelling correction," *Statistics and Computing*, vol. 1, no. 2, pp. 93–103, 1991.

[3] W. A. Gale, K. W. Church, and D. Yarowsky, "A method for disambiguating word senses in a large corpus," *Computers and the Humanities*, vol. 26, no. 5-6, pp. 415–439, 1992.

[4] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[5] A. Makur, G. W. Wornell, and L. Zheng, "On estimation of modal decompositions," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2717–2722.

[6] P.-S. Laplace, *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*. Springer Science & Business Media, 1825, vol. 13.

[7] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[8] A. Hodges, *Alan Turing: The Enigma*. Random House, 2012.

[9] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[10] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.