Estimating the Entropy of Linguistic Distributions

Aryaman Arora **Clara Meister** **Ryan Cotterell**
Georgetown University ETH Zürich
aa2190@georgetown.edu {clara.meister,ryan.cotterell}@inf.ethz.ch

Abstract

Shannon entropy is often a quantity of interest to linguists studying the communicative capacity of human language. However, entropy must typically be estimated from observed data because researchers do not have access to the underlying probability distribution that gives rise to these data. While entropy estimation is a well-studied problem in other fields, there is not yet a comprehensive exploration of the efficacy of entropy estimators for use with *linguistic* data. In this work, we fill this void, studying the empirical effectiveness of different entropy estimators for linguistic distributions. In a replication of two recent information-theoretic linguistic studies, we find evidence that the reported effect size is over-estimated due to over-reliance on poor entropy estimators. Finally, we end our paper with concrete recommendations for entropy estimation depending on distribution type and data availability.

1 Introduction

There is a natural connection between information theory, the mathematical study of communication systems, and linguistics, the study of human language—the primary vehicle that humans employ to communicate. Researchers have exploited this connection since information theory's inception (Shannon, 1951; Cherry et al., 1953; Harris, 1991). With the advent of modern computing, the number of information-theoretic linguistic studies has risen, exploring claims about language such as the optimality of the lexicon (Piantadosi et al., 2011; Pimentel et al., 2021), the complexity of morphological systems (Cotterell et al., 2019; Wu et al., 2019; Rathi et al., 2021), and the correlation between surprisal and language processing time (Smith and Levy, 2013; Bentz et al., 2017; Goodkind and Bicknell, 2018; Cotterell et al., 2018; Meister et al., 2021, *inter alia*).

In information-theoretic linguistics, a fundamental quantity of research interest is entropy. Entropy

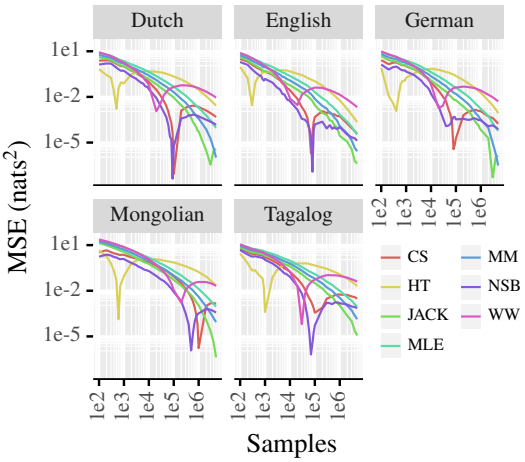

Figure 1: A comparison of several estimators of the entropy of the unigram distribution across 5 languages. Minima in all the graphs indicate sign changes in the error of the estimate, from an under- to an over-estimate.

is both useful to linguists in its own right, and is necessary for estimating other useful quantities, e.g., mutual information. However, the estimation of entropy from raw data can be quite challenging (Paninski, 2003; Nowozin, 2015), e.g., in expectation, the plug-in estimator *underestimates* entropy (Miller, 1955). Linguistic distributions often present additional challenges. For instance, many linguistic distributions, such as the unigram distribution, follow a power law (Zipf, 1935; Mitzenmacher, 2004).[1] Linguistics is not the only field with such nuances, and so a large number of entropy estimators have been proposed in other fields (Chao and Shen, 2003; Archer et al., 2014, *inter alia*). However, no work to date has attempted a practical comparison of these estimators on *natural language* data. This work fills this empirical void.

Our paper offers a large empirical comparison of the performance of 6 different entropy estimators

[1] As Nemenman et al. (2002) highlight, when estimating the entropy of a distribution that follows a power law, it is often possible to get an effectively meaningless estimate that is completely determined by the estimator's hyperparameters.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 175 - 195

on both synthetic and natural language data, an example of which is shown in Figure 1. We find that Chao and Shen's (2003) is the best estimator when very few data are available, but Nemenman et al.'s (2002) is superior as more data become available. Both are significantly better (in terms of mean-squared error) than the naïve plug-in estimator. Importantly, we also show that two recent studies (Williams et al., 2021; McCarthy et al., 2020) show smaller effect sizes when a better estimator is employed; however, we are able to reproduce a significant effect in both replications. We recommend that future studies carefully consider their choice of entropy estimators, taking into account data availability and the nature of the underlying distribution.[2]

2 Entropy and Language

Shannon entropy is a quantification of the uncertainty in a random variable. Given a (discrete) random variable X with probability distribution p over K possible outcomes $\mathcal{X} = \{x_k\}_{k=1}^{K}$, the Shannon entropy of X is defined as

$$\mathrm{H}(X) = \mathrm{H}(p) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} p(x_k) \log p(x_k) \quad (1)$$

Entropy has many uses throughout science and engineering; for instance, Shannon (1948) originally proposed entropy as a lower bound on the compressibility of a stochastic source.

Yet the application of information-theoretic techniques to linguistics is not so straightforward: Information-theoretic measures are defined over probability distributions and, in the study of natural language, we typically only have access to *samples* from the distribution of interest, e.g., the phonotactic distribution in English, which permits word we cannot find in a corpus, like *blick*, rather than the true probabilities required in the computation of Eq. (1). Indeed, it is often the case that not all elements of $\mathcal{X}$ are even observed in available data—such as words that were coined after the a corpus was collected.

Rather, p must be approximated in order to estimate $\mathrm{H}(p)$. One solution is **plug-in estimation**: Given samples from p, the maximum-likelihood estimate for p is "plugged" into Eq. (1). However, as originally noted by Miller (1955), this strategy generally yields poor estimates.[3] It is thus necessary to derive more nuanced estimators.

3 Statistical Estimation Theory

Statistical estimation theory provides us with the tools for estimating various quantities of interest based on samples from a distribution.

Central to this theory is the **estimator**: A statistic that approximates a property of the distribution our data is drawn from. More formally, let $\mathcal{D} = \{\widetilde{x}^{(n)}\}_{n=1}^{N}$ be samples from an unknown distribution p. Suppose we are interested in a quantity θ that can be computed as a function of the distribution p. An estimator $\widehat{\theta}(\mathcal{D})$ for θ is then a function of the data $\mathcal{D}$ that provides an approximation of θ.

Two properties of an estimator are often of interest: **bias**—the difference between the true value of θ and the expected value of our estimator $\widehat{\theta}(\mathcal{D})$ under p—and **variance**—how much $\widehat{\theta}(\mathcal{D})$ fluctuates from sample set to sample set:

$$\mathrm{bias}(\widehat{\theta}(\mathcal{D})) \stackrel{\text{def}}{=} \mathbb{E}_p[\widehat{\theta}(\mathcal{D})] - \theta \quad (2)$$

$$\mathrm{var}(\widehat{\theta}(\mathcal{D})) \stackrel{\text{def}}{=} \mathbb{E}_p[(\widehat{\theta}(\mathcal{D}) - \mathbb{E}_p[\widehat{\theta}(\mathcal{D})])^2] \quad (3)$$

It is desirable to construct an estimator that has both low bias and low variance. However, the **bias–variance** trade-off tells us that we often have to pick one, and we should focus on a balance between the two. This trade-off is evinced through mean-squared error (MSE), a metric oft-employed for assessing estimator quality:

$$\mathrm{MSE}(\widehat{\theta}(\mathcal{D})) = \mathrm{bias}(\widehat{\theta}(\mathcal{D}))^2 + \mathrm{var}(\widehat{\theta}(\mathcal{D})) \quad (4)$$

To recognize the trade-oft note that, for any fixed MSE, a decrease in bias must be compensated with an increase in variance and vice versa. Indeed, it is important to recognize that there is typically no single estimator that is seen as "best." Different estimators balance the bias–variance trade-off differently, making their perceived quality specific to one's use-case. Importantly, the effectiveness of an estimator also depends on the domain of interest. Consequently, an empirical study of various entropy estimators, which this paper provides, is necessary in order to determine which entropy estimators are best suited for linguistic distributions.

3.1 Plug-in Estimation of Entropy

A simple, two-step approach for estimating entropy is **plug-in** estimation. In the first step, we compute the maximum-likelihood estimate for p from our

[2] Our code is available at https://github.com/aryamanarora/entropy-estimation.

[3] A proof of this result in given in full in Proposition 1.

dataset $\mathcal{D}$ as follows

$$\widehat{p}_{\text{MLE}}(x_k) \stackrel{\text{def}}{=} \frac{\sum_{n=1}^{N} \mathbb{1}\{\widetilde{x}^{(n)} = x_k\}}{N} \tag{5}$$

In the second step, we plug Eq. (5) into Eq. (1) directly, which results in the estimator $\widehat{\text{H}}_{\text{MLE}}(\mathcal{D})$. So why is this a bad idea? While our probability estimates themselves are unbiased, entropy is a concave function. Consequently, by Jensen's inequality, this estimator is, in expectation, a *lower bound* on the true entropy (see App. E.1 for proof). Moreover, when $N \ll K$, which is often the case in power-law distributed data, the estimate becomes quite unreliable (Nemenman et al., 2002).

3.2 An Ensemble of Entropy Estimators

MM—Miller (1955) and Madow (1948). The first innovation in entropy estimation known to the authors is a simple fix derived from a first-order Taylor expansion of MLE (described above). The Miller–Madow estimator only involves a simple additive correction, which is shown below:

$$\widehat{\text{H}}_{\text{MM}}(\mathcal{D}) \stackrel{\text{def}}{=} \widehat{\text{H}}_{\text{MLE}}(\mathcal{D}) + \frac{K-1}{2N} \tag{6}$$

where K is size of the support of $\mathcal{X}$. The Miller–Madow correction should seem intuitive in that we add $\frac{K-1}{2N} \geq 0$ to compensate for the negative bias of the estimator. A full derivation of the Miller–Madow estimator is given in Proposition 2.

JACK—Zahl (1977). Next we consider the jackknife, which is a common strategy used to correct for the bias of statistical estimators. In the case of entropy estimation, we can apply the jackknife out of the box to correct the bias inherent in the MLE estimator. Explicitly, this is done by averaging plug-in entropy estimates $\widehat{\text{H}}_{\text{MLE}}(\mathcal{D})$ albeit with the n^{th} sample from the data removed; we denote this held-out plug-in estimator as $\widehat{\text{H}}^{\setminus n}_{\text{MLE}}(\mathcal{D})$. Averaging these "held-out" plug-in estimators results in the following simple entropy estimator

$$\widehat{\text{H}}_{\text{JACK}}(\mathcal{D}) \stackrel{\text{def}}{=} N\,\widehat{\text{H}}_{\text{MLE}}(\mathcal{D}) - \frac{N-1}{N}\sum_{n=1}^{N} \widehat{\text{H}}^{\setminus n}_{\text{MLE}}(\mathcal{D}) \tag{7}$$

Note that the jackknife is applicable to any estimator, not just $\widehat{\text{H}}_{\text{MLE}}(\mathcal{D})$, and, thus, can be combined with any of the other approaches mentioned.

HT—Horvitz and Thompson (1952). Horvitz–Thompson is a general scheme for building estimators that employs importance weighting in order to more efficiently estimate a function of a random variable. Importantly, this estimator gives us the ability to compensate for situations where the probability of an outcome is so low that it is often not observed in a sample, which is often the case for e.g., power-law distributions.

While a full exposition of HT estimators is outside of the scope of this work, in essence, we can divide the expected probability of a class by each class's estimated inclusion probability to compensate for such situations. Given the true probability of an outcome $p(x_k)$, the probability that it occurs at least once in a sample of size N is $1-(1-p(x_k))^N$. The HT estimator for entropy is then defined as

$$\widehat{\text{H}}_{\text{HT}}(\mathcal{D}) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} \frac{\widehat{p}_{\text{MLE}}(x_k)\log \widehat{p}_{\text{MLE}}(x_k)}{1-(1-\widehat{p}_{\text{MLE}}(x_k))^N} \tag{8}$$

using our MLE probability estimates $\widehat{p}_{\text{MLE}}(x_k)$.

CS—Chao and Shen (2003). Chao–Shen modifies HT by multiplying the MLE probability estimates by an estimate of sample coverage. Formally, let f_1 be the number of observed singletons[4] in sample; our sample coverage can be estimated as $\widehat{C} = 1 - \frac{f_1}{N}$. The CS estimator is then computed as:

$$\widehat{\text{H}}_{\text{CS}}(\mathcal{D}) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} \frac{\widehat{C}\cdot\widehat{p}_{\text{MLE}}(x_k)\log \widehat{C}\cdot\widehat{p}_{\text{MLE}}(x_k)}{1-(1-\widehat{C}\cdot\widehat{p}_{\text{MLE}}(x_k))^N} \tag{9}$$

In the case that $f_1 = N$, we set $f_1 = N-1$ to ensure the estimated entropy is not 0.

WW—Wolpert and Wolf (1995). One family of entropy estimators in information theory is based on Bayesian principles. The first of these was the Wolpert–Wolf estimator, which uses a Dirichlet prior (with concentration parameter α and a uniform base distribution). This Bayesian estimator has a clean, closed form:

$$\widehat{\text{H}}_{\text{WW}}(\mathcal{D} \mid \boldsymbol{\alpha}) \stackrel{\text{def}}{=} \psi\left(\widetilde{A}+1\right) - \sum_{k=1}^{K} \frac{\widetilde{\alpha}_k}{\widetilde{A}}\psi(\widetilde{\alpha}_k+1) \tag{10}$$

where $\widetilde{\alpha}_k = c(x_k) + \alpha_k$ (for the histogram count $c(x_k)$ of class k in the sample; this is analogous to Laplace smoothing), $\widetilde{A} = \sum_{k=1}^{K}\widetilde{\alpha}_k$, and ψ is the digamma function. A full derivation of Eq. (10) is given in Proposition 3. Unfortunately, Eq. (10) is

[4] A singleton (*hapax legomenon*) is an outcome which is observed only once in the sample.

	MAB				MSE			
	10^2	10^3	10^4	10^5	10^2	10^3	10^4	10^5
English	HT	HT	NSB	NSB	HT	HT	NSB	NSB
German	HT	HT	NSB	CS	HT	HT	NSB	CS
Dutch	HT	HT	NSB	CS	HT	HT	NSB	CS
Mongolian	NSB	HT	NSB	NSB	NSB	HT	NSB	NSB
Tagalog	HT	HT	NSB	NSB	HT	HT	NSB	NSB

Table 1: The best unigram entropy estimators on the corpora studied, tested on various N averaged over 100 samples. All differences are statistically significant on the permutation test; lighter color indicates fewer statistically significant comparisons on the Tukey test. *Scale*: significantly better than 6 5 4 3 2 1 0 other estimators.

very dependent on the choice of $\boldsymbol{\alpha}$: For large K, $\boldsymbol{\alpha}$ almost completely determines the final entropy estimate, an observation first made by Nemenman et al. (2002) which motivated their improved estimator described below.

NSB—Nemenman et al. (2002). Nemenman et al. (NSB) attempt to alleviate the Wolpert–Wolf estimator's dependence on $\boldsymbol{\alpha}$. They take $\boldsymbol{\alpha} = \alpha \cdot \mathbf{1}$, enforcing that the Dirichlet prior is symmetric, and develop a hyperprior over α that results in a near-uniform distribution over entropy. The hyperprior is given by

$$p_{\text{NSB}}(\alpha) \stackrel{\text{def}}{=} \frac{K\psi_1(K\alpha+1) - \psi_1(\alpha+1)}{\log K} \quad (11)$$

where ψ_1 is the trigamma function. A full derivation of Eq. (11) is given in Proposition 4. This choice of hyperprior mitigates the effect that the chosen α has on the entropy estimate. Nemenman et al.'s (2002) entropy estimator is then the posterior mean of the Wolpert–Wolft estimator taken under p_{NSB}:

$$\widehat{\mathrm{H}}_{\text{NSB}}(\mathcal{D}) = \int_0^\infty \widehat{\mathrm{H}}_{\text{WW}}(\mathcal{D} \mid \alpha \cdot \mathbf{1})\, p_{\text{NSB}}(\alpha)\, \mathrm{d}\alpha \quad (12)$$

Typically, numerical integration is used to quickly compute the unidimensional integral.

4 Experiments

Here we provide an evaluation of the entropy estimators presented in §3.2 on linguistic data.

4.1 Entropy of the Unigram Distribution

We start our study with a controlled experiment where we estimate the entropy of the truncated unigram distribution, the (finite) distribution over the frequent word tokens in a language without regard to context (Baayen et al., 2016; Diessel, 2017; Divjak, 2019; Nikkarinen et al., 2021). We renormalize the frequency counts of corpora in English, German, and Dutch (taken from CELEX; Baayen et al., 1995), as well as Mongolian and Tagalog (from Wikipedia[5]). We take this renormalization as a gold standard distribution, since we cannot access the underlying unigram distribution. We then draw samples of varying sizes ($N \in \{10^2, 10^3, 10^4, 10^5\}$) from the distribution of renormalized frequency counts to test the estimators' ability to recover the underlying distributions' entropy. While the renormalized frequency counts are not necessarily representative of the *true* unigram distribution, they nevertheless provide us with a controlled setting to benchmark various entropy estimators.

We evaluate the estimators on both bias and MSE, as defined in (2) and (4), as well as mean absolute bias (MAB). To test the statistical significance of differences in metrics between entropy estimators, we use paired permutation tests (Good, 2000) (sampling $1,000$ permutations) between pairs of estimators, checking MAB and MSE. We run Tukey's test (1949) to judge the statistical significance of differences in MAB and MSE between all pairs of estimators, which found only a few insignificant comparisons when N was large.

Results are shown in Table 1 and Figure 1. We find that NSB (followed closely by CS) converges almost to the true entropy from below using with only a few samples. HT is the best estimator for $N < 2,000$, but as N increases it tends to overestimate entropy to the point where its bias is greater than that of MLE. Besides HT, all estimators at all tested sample sizes N have lower MAB and MSE than MLE.

4.2 Replication of Williams et al. (2021)

Next, we turn to a replication of Williams et al.'s (2021) information-theoretic study on the associa-

[5]We used dumps from November 1, 2021: Mongolian and Tagalog; the extracted counts are available in our repository.

Language	n	MLE	CS	MM	JACK	WW	NSB
Italian	16,856	20.00%	15.56%	16.43%	14.09%	19.67%	11.41%
Polish	15,525	30.52%	23.48%	25.49%	21.75%	34.68%	17.07%
Portuguese	7,409	27.60%	20.76%	22.51%	18.81%	33.32%	14.18%
Spanish	21,408	20.50%	15.17%	16.44%	13.80%	21.04%	10.50%
Arabic	2,483	45.31%	38.49%	40.99%	37.93%	49.09%	34.82%
Croatian	13,856	31.35%	26.04%	26.62%	23.08%	35.66%	19.06%
Greek	3,305	41.58%	33.17%	36.39%	32.32%	48.80%	27.00%

Table 2: Normalized mutual information, calculated with several estimators, between adjectives and the inanimate nouns they modify based on UD corpora. Colored-in cell means statistically significant NMI value.

tion between gendered inanimate nouns and their modifying adjectives. They estimate mutual information by using its familiar decomposition as the difference of two entropies: $\mathrm{MI}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y)$. The entropies $\mathrm{H}(X)$ and $\mathrm{H}(X \mid Y)$ are estimated independently and then their difference is computed. We replicate Williams et al.'s (2021) experiments using gold-parsed Universal Dependencies corpora, filtering out animate nouns with Multilingual WordNet (Bond and Foster, 2013). We rerun their experimental set-up using our full suite of entropy estimators to determine whether the relationship they posit remains significant, checking 3 more languages not in the original study.

We report results for normalized mutual information (dividing MI by maximum possible MI) in Table 2. We find that using NSB (the estimator we found most effective in §4.1) instead of MLE, nearly halves the measured effect in all languages. However, the effect remains statistically significant in 5 of 7 languages tested, including the 4 that were also in the original study.

4.3 Replication of McCarthy et al. (2020)

Finally, we turn our attention to McCarthy et al.'s (2020) study on the similarity between grammatical gender partitions between languages. Using information-theoretic measures, they found that closely related languages have more similar gender groupings of core lexical items. We replicate their experiment on Swadesh lists (Swadesh, 1955) for 10 European languages with different estimators, and find that hierarchical clustering over both mutual (MI) and variational information (VI) produces the same trees as the original study. In this case, using NSB, our recommended estimator, results in a reduced estimate of MI (e.g. Croatian–Slovak: 0.54 with MLE $\rightarrow 0.46$ with NSB), but significance testing with 1,000 permutations finds the same pairs were statistically significant for both MI and VI regardless of estimator: all pairs of Slavic languages and Romance languages, and Bulgarian–Spanish (see Figure 2). Thus, we see a similar result here as in the previous replication.

5 Conclusion

This work presents the first empirical study comparing the performance of various entropy estimators for use with natural language distributions. From experiments on synthetic data (appendix) and natural data (CELEX), and two replication studies of recent papers in information-theoretic linguistics, we find that the oft-employed plug-in estimator of entropy can cause misleading results, e.g., the overestimates of effect sizes seen in both replication studies. The recommendation of our paper is that researchers should carefully consider their choice of entropy estimator based on data availability and the nature of the underlying distribution.

Ethics Statement

The authors foresee no ethical concerns with the research presented in this paper.

Acknowledgments

We thank Adina Williams, Lucas Torroba Hennigen, Tiago Pimentel, and the anonymous reviewers for feedback on the manuscript.

References

Evan Archer, Il Memming Park, and Jonathan W. Pillow. 2014. Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, 15(81):2833–2868.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. *CELEX2*. Linguistic Data Consortium, Philadelphia.

R. Harald Baayen, Petar Milin, and Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Anne Chao and Tsung-Jen Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443.

E. Colin Cherry, Morris Halle, and Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language*, pages 34–46.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Holger Diessel. 2017. Usage-based linguistics. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Dagmar Divjak. 2019. *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press.

Herwig Friedl and Erwin Stampfer. 2002. Jackknife resampling. In *Encyclopedia of Environmetrics*, volume 2, pages 1089–1098. Wiley Chichester.

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

Phillip I. Good. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd edition. Springer.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Zellig Harris. 1991. *A Theory of Language and Information: A Mathematical Approach*, 1 edition. Clarendon Press.

D. G. Horvitz and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

William G. Madow. 1948. On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, pages 535–545.

Simone Marsili. 2016. simomarsili/ndd: Bayesian entropy estimation in Python - via the Nemenman-Schafee-Bialek algorithm.

Arya D. McCarthy, Adina Williams, Shijia Liu, David Yarowsky, and Ryan Cotterell. 2020. Measuring the similarity of grammatical gender systems by comparing partitions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5664–5675, Online. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Miller. 1955. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, IL.

Michael Mitzenmacher. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251.

Ilya Nemenman, F. Shafee, and William Bialek. 2002. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Irene Nikkarinen, Tiago Pimentel, Damián Blasi, and Ryan Cotterell. 2021. Modeling the unigram distribution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3721–3729, Online. Association for Computational Linguistics.

Sebastian Nowozin. 2015. Estimating discrete entropy, part 1.

Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1254.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021. How (non-)optimal is the lexicon? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4426–4438, Online. Association for Computational Linguistics.

Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, and Stelios Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 149–160.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa,Italy. Linköping University Electronic Press.

Neil Rathi, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10115–10120, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Claude E. Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague Arabic Dependency Treebank: A word on the million words. In *Proceedings of the Workshop on Arabic and Local Languages*, pages 16–23, Marrakech, Morocco.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sara Tonelli, Rodolfo Delmonte, and Antonella Bristot. 2008. Enriching the venice Italian treebank with dependency and grammatical relations. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

John Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Tim Vieira. 2017. Estimating means in a finite universe.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

David H. Wolpert and David R. Wolf. 1995. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841.

Alina Wróblewska. 2018. Extended and enhanced Polish dependency bank in Universal Dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

Samuel Zahl. 1977. Jackknifing an index of diversity. *Ecology*, 58(4):907–913.

Zhiyi Zhang. 2012. Entropy estimation in Turing's perspective. *Neural Computation*, 24(5):1368–1389.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language*. Houghton-Mifflin, New York, NY, USA.

	MAB				MSE			
	10^1	10^2	10^3	10^4	10^1	10^2	10^3	10^4
2	HT	WW	WW	WW	WW	WW	WW	JACK
5	MM	WW	WW	JACK	MM	WW	WW	MM
10	JACK	CS	WW	MM	JACK	WW	WW	MLE
100	CS	CS	JACK	WW	CS	JACK	JACK	WW
1000	CS	HT	CS	JACK	CS	HT	CS	JACK

Table 3: Estimators with least MAB (mean absolute bias) and MSE (mean squared error) for various combinations of N and K sampling from **symmetric Dirichlet**. The lighter the color the fewer estimators the best estimator was found to be statistically significantly better than.

	MAB				MSE			
	10^1	10^2	10^3	10^4	10^1	10^2	10^3	10^4
100	CS	CS	CS	J	CS	CS	CS	J
1000	NSB	HT	NSB	J	CS	HT	NSB	J

Table 4: Estimators with least MAB (mean absolute bias) and MSE (mean squared error) for various combinations of N and K sampling from **Zipfian distributions**.

A Implementation

The code for each of the entropy estimators is implemented in Python using `numpy` (Harris et al., 2020), except for NSB which was taken from an existing efficient implementation in the `ndd` module (Marsili, 2016). We calculated entropies with base e (in nats).

B Experiments with simulated data

In our experiments with simulated data, we explore distributions sampled from a symmetric Dirichlet prior with varying number of classes K and known distributions of Zipfian form with various parameters. Words in natural languages have a roughly Zipfian distribution, with probability inversely proportional to rank (Zipf, 1935), and a symmetric Dirichlet distribution is analogous to e.g. POS tag label distributions in natural language. Thus, studying synthetic data from such distributions as a start is useful.

B.1 Experiment 1: Symmetric Dirichlet distributions

We sample $1,000$ distributions from a symmetric Dirichlet distribution with variable number of classes K, i.e. with paramater $\alpha = [\alpha_1, \ldots, \alpha_K] = [1, \ldots, 1]$. We calculate entropy estimates on different sample sizes N. Since we know the parameters of the true distribution, we can compare estimates with the true entropy. We do pairwise comparisons of the MAB and MSE of estimators, using paired permutation tests to establish significance. Table 3 shows our results, including significance tests. It is clear that when $N \gg K$, all of the estimators have nearly converged to the true value and estimator choice does not matter. However, in the low-sample regime some estimators are indeed significantly better at approximating the true entropy. Our results are mixed as to which estimator is best in what context; the one found to be most frequently significantly better than other estimators was Chao–Shen. What is clear is that MLE is never the best choice.

B.2 Experiment 2: Zipfian distributions

We sample $1,000$ finite Zipfian distributions with K classes which obey Zipf's law, that the probability of an outcome is inverse proportional to its rank. The experimental setup is the same as in Experiment 1. A Zipfian distribution approximates (but is not a perfect model of) the distribution of tokens in natural language text in some languages, including English, which was the basis for the law being proposed. Compare similar experiments on infinite Zipf distributions by Zhang (2012). Results are in Table 4.

C Replication of Williams et al. (2021)

We used the following UD treebanks:

- **Arabic**: PADT (Smrž et al., 2008; Taji et al., 2017);
- **Greek**: GDT (Prokopidis et al., 2005; Prokopidis and Papageorgiou, 2017);
- **Italian**: ISDT (Bosco et al., 2013), VIT (Tonelli et al., 2008);
- **Polish**: PDB (Wróblewska, 2018);
- **Portuguese**: GSD (McDonald et al., 2013), Bosque (Rademaker et al., 2017);
- **Spanish**: AnCora (Taulé et al., 2008), GSD (McDonald et al., 2013).

D Additional Figures

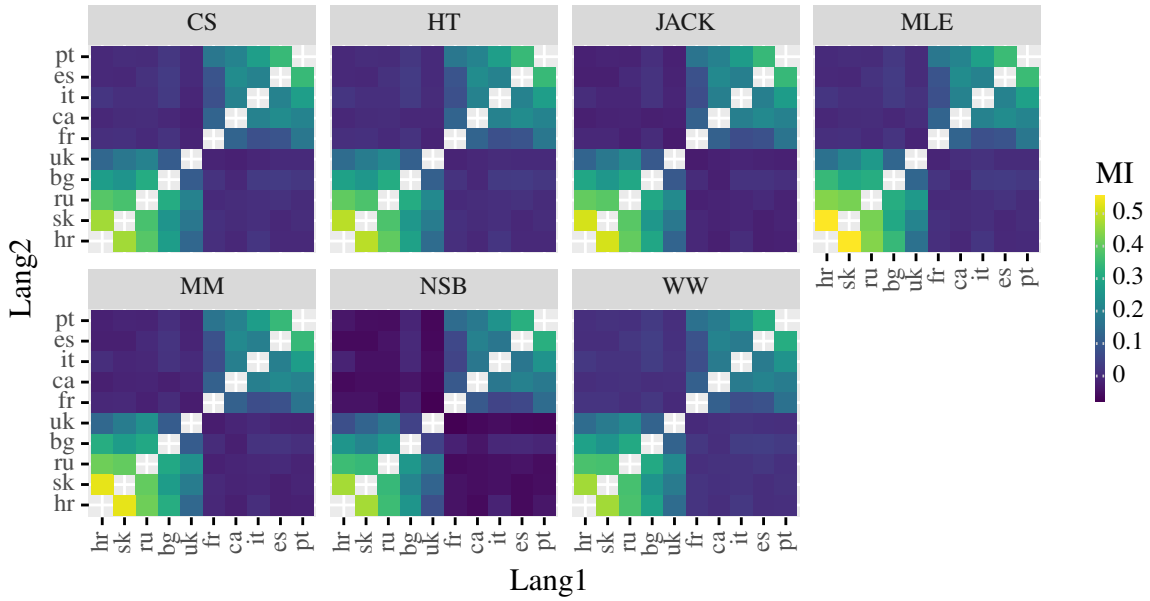

Figure 2: Mutual information between the gender partitions of language pairs with various estimators, replicating McCarthy et al. (2020).

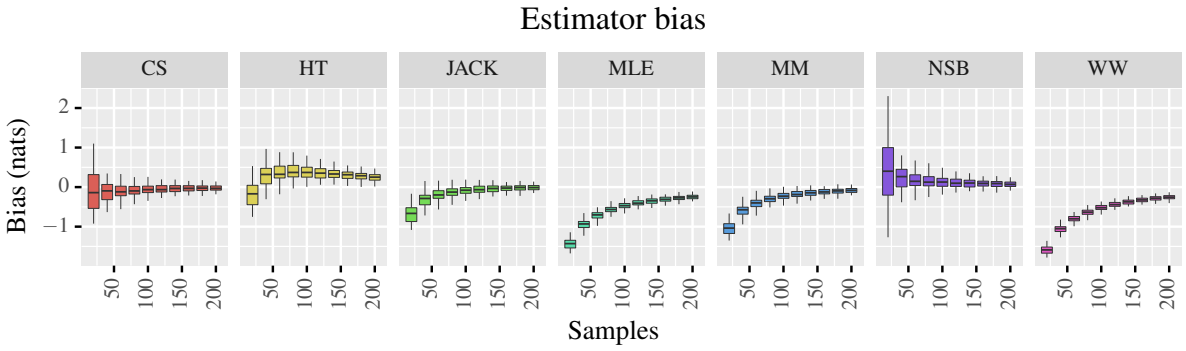

Figure 3: The distribution of bias for entropy over several estimators given variable sample size N, sampling from 100 distributions taken from a symmetric Dirichlet prior with $K = 100$.

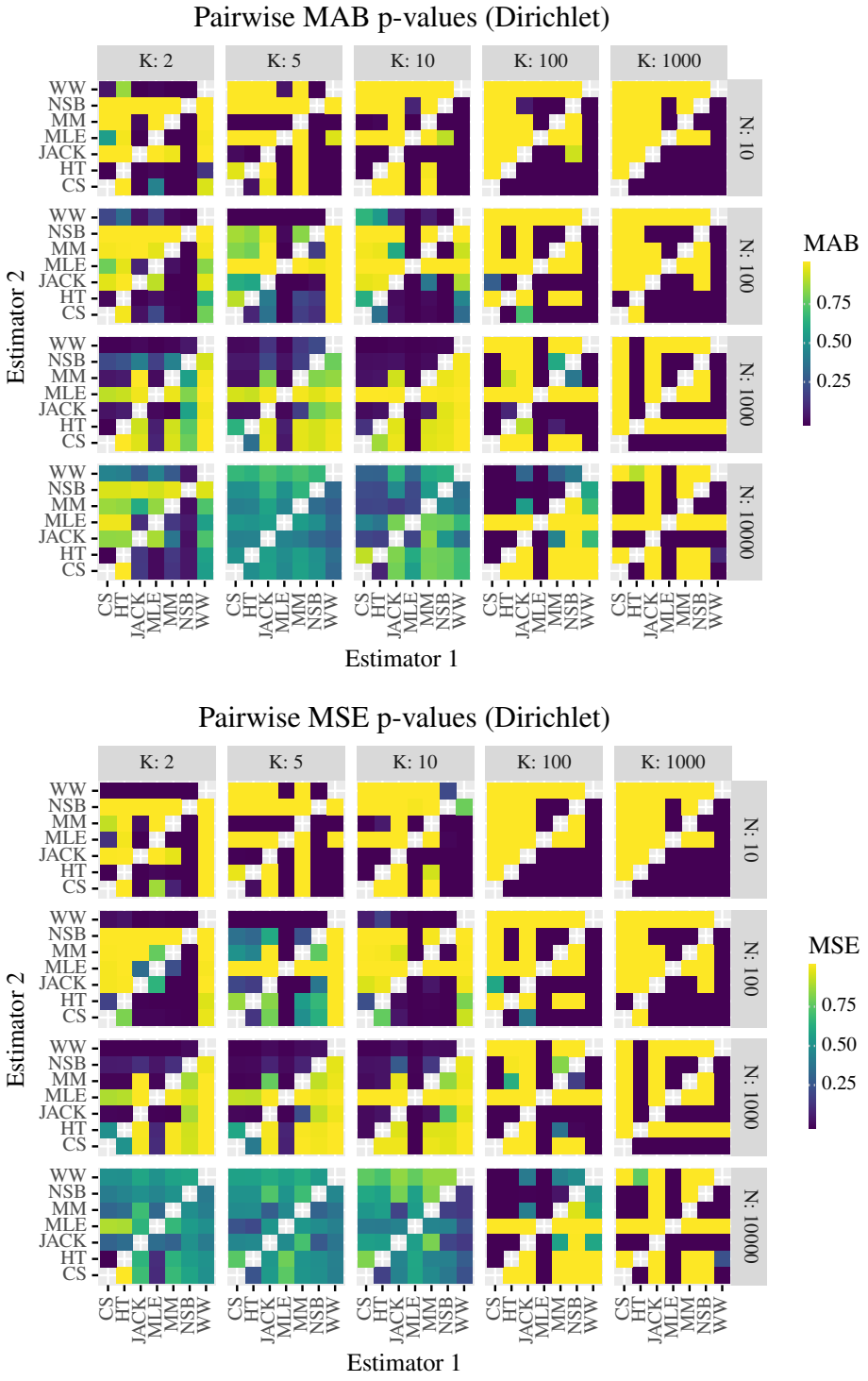

Figure 4: The heatmaps display the p-values calculated between pairs of estimators for mean absolute bias (MAB) and mean squared error (MSE) for Experiment 1. More purple values mean the estimator on the y-axis (Estimator 2) is better than the estimator on the x-axis (Estimator 1). Comparisons tend to become non-significant as N increases, since all the estimators gradually converge to the true entropy.

E Derivation of the Entropy Estimators

Let $\mathcal{X} = \{x_k\}_{k=1}^K$ be a finite set. Let p be a distribution over $\mathcal{X}$. The **entropy** of p is defined as

$$\mathrm{H}(p) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} p_k \log p_k \tag{13}$$

Given a dataset of N samples $\mathcal{D}$ sampled i.i.d. from p, our goal is to estimate the entropy $\mathrm{H}(p)$ from samples $\mathcal{D}$ from the true distribution p. We will denote the count of an item x_k as $c(x_k) = \sum_{n=1}^{N} \mathbb{1}\left\{x_k = \widetilde{x}^{(n)}\right\}$. The **maximum-likelihood estimate** (MLE) of p given $\mathcal{D}$ is denoted $\frac{\sum_{n=1}^{N} \mathbb{1}\{\widetilde{x}^{(n)}=x_k\}}{N}$. The **plug-in estimate** of $\mathrm{H}(p)$ is defined to be the estimate of $\mathrm{H}(p)$ obtained by plugging the MLE estimate $\widehat{p}_{\text{MLE}}$ directly into the definition of entropy, i.e.,

$$\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) = \mathrm{H}(\widehat{p}_{\text{MLE}}) = -\sum_{k=1}^{K} \widehat{p}_{\text{MLE}}(x_k) \log \widehat{p}_{\text{MLE}}(x_k) = -\sum_{k=1}^{K} \frac{c(x_k)}{N} \log \frac{c(x_k)}{N} \tag{14}$$

This section discusses the problems with Eq. (14) as an estimator and provides detailed derivations of improved estimators found in the literature.

E.1 The Plug-in Estimator is Negatively Biased

Proposition 1. *The MLE entropy estimator in expectation underestimates true entropy, i.e.,*

$$\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) = \mathbb{E}\left[\sum_{k=1}^{K} -\widehat{p}_{\text{MLE}}(x_k) \log \widehat{p}_{\text{MLE}}(x_k)\right] \leq \mathrm{H}(p) \tag{15}$$

Proof. The result is a simple consequence of Jensen's inequality and some basic manipulations:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K} -\widehat{p}_{\text{MLE}}(x_k) \log \widehat{p}_{\text{MLE}}(x_k)\right] &= \sum_{k=1}^{K} \mathbb{E}[-\widehat{p}_{\text{MLE}}(x_k) \log \widehat{p}_{\text{MLE}}(x_k)] && \text{(linearity of expectation)} \\
&\leq -\sum_{k=1}^{K} \mathbb{E}[\widehat{p}_{\text{MLE}}(x_k)] \log \mathbb{E}[\widehat{p}_{\text{MLE}}(x_k)] && \text{(Jensen's inequality)} \\
&= -\sum_{k=1}^{K} p(x_k) \log p(x_k) && (\mathbb{E}[\widehat{p}_{\text{MLE}}(x_k)] = p(x_k)) \\
&= \mathrm{H}(p) && \text{(definition of entropy)}
\end{aligned}$$

This completes the result. □

E.2 Miller–Madow

Proposition 2. *Let p be a categorical distribution over $\mathcal{X} = \{x_1, \ldots, x_K\}$, i.e., a categorical distribution with support K. Let $\mathcal{D}$ be our dataset of size N sampled from p. Finally, let $\widehat{p}_{\text{MLE}}$ be the maximum-likelihood estimate computed on $\mathcal{D}$. Then, we have*

$$\begin{aligned}
\operatorname{bias}\left(\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D})\right) &\stackrel{\text{def}}{=} \mathbb{E}_p\left[\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D})\right] - \mathrm{H}(p) && (16) \\
&= -\frac{K-1}{2N} + o\left(N^{-1}\right) && (17)
\end{aligned}$$

Proof. We start by taking a first-order Taylor expansion and take an expectation of both sides.

$$\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) = \underbrace{\mathrm{H}(\widehat{p}_{\text{MLE}}, p)}_{\text{cross-entropy}} - \mathrm{KL}(\widehat{p}_{\text{MLE}} \,||\, p) \qquad \text{(Lemma 1)} \tag{18}$$

$$\mathbb{E}_p\left[\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D})\right] = \mathbb{E}_p\left[\mathrm{H}(\widehat{p}_{\mathrm{MLE}}, p)\right] - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(expectation)} \quad (19)$$

$$= \mathbb{E}_p\left[-\sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k) \log p(x_k)\right] - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(defn. H}(p,q)\text{)} \quad (20)$$

$$= -\sum_{k=1}^{K} \mathbb{E}_p\left[\widehat{p}_{\mathrm{MLE}}(x_k) \log p(x_k)\right] - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(linearity)} \quad (21)$$

$$= -\sum_{k=1}^{K} \mathbb{E}_p\left[\widehat{p}_{\mathrm{MLE}}(x_k)\right] \log p(x_k) - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(algebra)} \quad (22)$$

$$= -\sum_{k=1}^{K} p(x_k) \log p(x_k) - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(unbiased)} \quad (23)$$

$$= \mathrm{H}(p) - \mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(defn. of H}(p)\text{)} \quad (24)$$

$$(25)$$

This gives us:

$$\mathbb{E}_p\left[\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D})\right] - \mathrm{H}(p) = -\mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(subtract H}(p)\text{)} \quad (26)$$

Thus, we may compactly write the bias as:

$$\mathrm{bias}\left(\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D})\right) = \mathbb{E}_p\left[\mathrm{H}(\widehat{p}_{\mathrm{MLE}})\right] - \mathrm{H}(p) \quad \text{(definition of bias)} \quad (27)$$

$$= -\mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right] \quad \text{(above computation)} \quad (28)$$

$$\leq 0 \quad \text{(non-negativity of KL)} \quad (29)$$

Now, we find a simpler expression for the remainder $\mathbb{E}_p\left[\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p)\right]$. Again, we start with a second-order Taylor expansion

$$\mathrm{KL}(p \mid\mid q) = \sum_{x \in \mathcal{X}} \frac{\Delta(x)^2}{2q(x)} + o\left(\Delta(x)^2\right) \quad \text{(Lemma 2)} \quad (30)$$

around the point $\Delta(x) = p(x) - q(x)$. Define $\widehat{p}_{\mathrm{MLE}}(x_k) = \frac{c(x_k)}{N}$ where $c(x_k)$ is the count of x_k in the training set. We now simplify the first term:

$$\mathbb{E}_p\left[\sum_{k=1}^{K} \frac{\Delta(x_k)^2}{2q(x_k)}\right] = \mathbb{E}_p\left[\sum_{k=1}^{K} \frac{(\widehat{p}_{\mathrm{MLE}}(x_k) - p(x_k))^2}{2p(x_k)}\right] \quad \text{(definition of } \Delta(x_k)\text{)} \quad (31)$$

$$= \mathbb{E}_p\left[\sum_{k=1}^{K} \frac{(\frac{c(x_k)}{N} - p(x_k))^2}{2p(x_k)}\right] \quad \text{(definition of MLE)} \quad (32)$$

$$= \mathbb{E}_p\left[\sum_{k=1}^{K} \frac{(c(x_k) - Np(x_k))^2}{2N^2 p(x_k)}\right] \quad (\times N/N) \quad (33)$$

$$= \frac{1}{2N}\mathbb{E}_p\left[\sum_{k=1}^{K} \frac{(c(x_k) - Np(x_k))^2}{Np(x_k)}\right] \quad \text{(pulling out } 1/2N\text{)} \quad (34)$$

$$= \frac{1}{2N}\mathbb{E}_p\left[\sum_{k=1}^{K} \frac{c(x_k)^2 - 2c(x_k)Np(x_k) + N^2 p(x_k)^2}{Np(x_k)}\right] \quad \text{(exp. the binomial)} \quad (35)$$

$$= \frac{1}{2N}\sum_{k=1}^{K} \frac{\mathbb{E}_p\left[c(x_k)^2\right] - 2Np(x_k)\mathbb{E}_p\left[c(x_k)\right] + N^2 p(x_k)^2}{Np(x_k)} \quad \text{(lin. of expect.)} \quad (36)$$

$$= \frac{1}{2N}\sum_{k=1}^{K}\frac{\begin{array}{l}Np_k(1-p(x_k)) + N^2p(x_k)^2 \\ \quad - 2N^2p(x_k)^2 + N^2p(x_k)^2\end{array}}{Np(x_k)} \quad \text{(moments of MLE)} \quad (37)$$

$$= \frac{1}{2N}\sum_{k=1}^{K}\frac{Np_k(1-p(x_k))}{Np(x_k)} + \underbrace{\frac{1}{2N}\sum_{k=1}^{K}\frac{N^2p(x_k)^2 - 2N^2p(x_k)^2 + N^2p(x_k)^2}{Np(x_k)}}_{=0} \quad (38)$$

$$= \frac{1}{2N}\sum_{k=1}^{K}\frac{\cancel{Np(x_k)}(1-p(x_k))}{\cancel{Np(x_k)}} \quad (39)$$

$$= \frac{1}{2N}\sum_{k=1}^{K}(1-p(x_k)) \quad \text{(algebra)} \quad (40)$$

$$= \frac{1}{2N}\underbrace{\sum_{k=1}^{K}1}_{=K} - \frac{1}{2N}\underbrace{\sum_{k=1}^{K}p(x_k)}_{=1} \quad \text{(algebra)} \quad (41)$$

$$= \frac{K-1}{2N} \quad (42)$$

Next, we simplify the second term, $o\left(\Delta(x)^2\right)$, in the MLE case:

$$\mathbb{E}_p\left[o\left(\Delta(x)^2\right)\right] = \mathbb{E}_p\left[o\left((\widehat{p}_{\text{MLE}}(x_k) - p(x_k))^2\right)\right] \quad \text{(definition of } \Delta) \quad (43)$$

$$= \mathbb{E}_p\left[o\left(\left(\frac{c(x_k)}{N} - p(x_k)\right)^2\right)\right] \quad \text{(definition of MLE)} \quad (44)$$

$$= \mathbb{E}_p\left[o\left(\frac{(c(x_k) - Np(x_k))^2}{N^2}\right)\right] \quad (\times^N/_N) \quad (45)$$

$$= \mathbb{E}_p\left[o\left(\frac{c(x_k)^2 - 2c(x_k)Np(x_k) + N^2p(x_k)^2}{N^2}\right)\right] \quad (46)$$

$$= o\left(\frac{\mathbb{E}_p\left[c(x_k)^2 - 2c(x_k)Np(x_k) + N^2p(x_k)^2\right]}{N^2}\right) \quad \text{(push exp. through)} \quad (47)$$

$$= o\left(\frac{\begin{array}{l}Np_k(1-p(x_k)) + N^2p(x_k)^2 \\ \quad - 2N^2p(x_k)^2 + N^2p(x_k)^2\end{array}}{N^2}\right) \quad (48)$$

$$= o\left(\frac{Np(x_k)(1-p(x_k))}{N^2}\right) \quad \text{(cancel terms)} \quad (49)$$

$$= o\left(\frac{p(x_k)(1-p(x_k))}{N}\right) \quad \text{(cancel } N \text{ in fraction)} \quad (50)$$

$$= o\left(N^{-1}\right) \quad \text{(ignore constants)} \quad (51)$$

Putting it all together, we get that $\text{bias}\left(\text{H}(\widehat{p}_{\text{MLE}})\right) = -\frac{K-1}{2N} + o\left(N^{-1}\right)$ which is the desired result. □

Interestingly, it can be seen that the negative bias of the MLE gets worse as the number of classes K grows. Distributions with large K pop up frequently when dealing with natural language.

Corollary 1. *The plug-in estimator of entropy is consistent.*

Proof. From Proposition 2, we have $\operatorname{bias}\left(\mathrm{H}(\widehat{p}_{\mathrm{MLE}})\right) = -\frac{K-1}{2N} + o\left(N^{-1}\right)$. Clearly, as $N \to 0$, we have $\operatorname{bias}\left(\mathrm{H}(\widehat{p}_{\mathrm{MLE}})\right) \to 0$, so the estimator is consistent. One could also prove consistency through a simple application of the continuous mapping theorem. □

Estimator 1 (Miller–Madow)**.** *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Then, the Miller–Madow estimator of $\mathrm{H}(p)$ is given by*

$$\widehat{\mathrm{H}}_{\mathrm{MM}}(\mathcal{D}) \stackrel{\text{def}}{=} \widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D}) + \frac{K-1}{2N} \tag{52}$$

The Miller–Madow estimator is biased, however it is consistent.

Lemma 1. *The the first-order Taylor approximation of $\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D})$ around the distribution p is given by*

$$\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D}) = \mathrm{H}(\widehat{p}_{\mathrm{MLE}}, p) + R(p, \widehat{p}_{\mathrm{MLE}}) \tag{53}$$

where the remainder R is given by

$$R(p, \widehat{p}_{\mathrm{MLE}}) = -\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \mid\mid p) \tag{54}$$

Proof. The result follows from direct computation. We start by taking the Taylor expansion of $\mathrm{H}(\widehat{p}_{\mathrm{MLE}})$ around $\mathrm{H}(p)$:

$$\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D}) = \mathrm{H}(p) + \sum_{k=1}^{K} \frac{\partial}{\partial p(x_k)} \Big[\mathrm{H}(p)\Big] \Big(\widehat{p}_{\mathrm{MLE}}(x_k) - p(x_k)\Big) + \underbrace{R(p, \widehat{p}_{\mathrm{MLE}})}_{\text{remainder}} \tag{55}$$

Our first order term can then be rewritten as follows:

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\partial}{\partial p(x_k)} \Big[\mathrm{H}(p)\Big] \Big(\widehat{p}_{\mathrm{MLE}}(x_k) - p(x_k)\Big) && (56)\\
&\quad= \sum_{k=1}^{K} \frac{\partial}{\partial p(x_k)} \left[\sum_{k'=1}^{K} -p(x_{k'}) \log p(x_{k'})\right] \Big(\widehat{p}_{\mathrm{MLE}}(x_k) - p(x_k)\Big) && (57)\\
&\quad= \sum_{k=1}^{K} \left[\sum_{k'=1}^{K} -\frac{\partial}{\partial p(x_k)} p(x_{k'}) \log p(x_{k'})\right] \Big(\widehat{p}_{\mathrm{MLE}}(x_k) - p(x_k)\Big) && \text{(linearity)} \quad (58)\\
&\quad= \sum_{k=1}^{K} \left[\sum_{k'=1}^{K} \frac{\partial}{\partial p(x_k)} p(x_{k'}) \log p(x_{k'})\right] \Big(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\Big) && \text{(sign)} \quad (59)\\
&\quad= \sum_{k=1}^{K} \Big(1 + \log p(x_k)\Big) \Big(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\Big) && (60)\\
&\quad= \sum_{k=1}^{K} \Big(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\Big) + \log p(x_k) \left(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\right) && (61)\\
&\quad= \sum_{k=1}^{K} \Big(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\Big) + \sum_{k=1}^{K} \log p(x_k) \left(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\right) && (62)\\
&\quad= \underbrace{\sum_{k=1}^{K} p(x_k)}_{=1} - \underbrace{\sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k)}_{=1} + \sum_{k=1}^{K} \log p(x_k) \left(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\right) && \text{(distrib. sum)} \quad (63)\\
&\quad= \sum_{k=1}^{K} \log p(x_k) \left(p(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k)\right) && \text{(simplify)} \quad (64)
\end{aligned}$$

$$= \underbrace{\sum_{k=1}^{K} \log p(x_k) p(x_k)}_{-\mathrm{H}(p)} - \underbrace{\sum_{k=1}^{K} \log p(x_k) \widehat{p}_{\mathrm{MLE}}(x_k)}_{\mathrm{H}(p, \widehat{p}_{\mathrm{MLE}})} \quad \text{(distrib. sum)} \quad (65)$$

$$= \mathrm{H}(p, \widehat{p}_{\mathrm{MLE}}) - \mathrm{H}(p) \quad (66)$$

Plugging this back into our Taylor expansion, we get the following:

$$\widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D}) = \cancel{\mathrm{H}(p)} - \cancel{\mathrm{H}(p)} + \mathrm{H}(p, \widehat{p}_{\mathrm{MLE}}) + R(p, \widehat{p}_{\mathrm{MLE}}) \quad (67)$$

Now, we see that this implies

$$R(p, \widehat{p}_{\mathrm{MLE}}) = \widehat{\mathrm{H}}_{\mathrm{MLE}}(\mathcal{D}) - \mathrm{H}(\widehat{p}_{\mathrm{MLE}}, p) \quad \text{(algebra)} \quad (68)$$

$$= -\sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k) \log \widehat{p}_{\mathrm{MLE}}(x_k) + \sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k) \log p(x_k) \quad \text{(defn.)} \quad (69)$$

$$= -\sum_{k=1}^{K} (\widehat{p}_{\mathrm{MLE}}(x_k) \log \widehat{p}_{\mathrm{MLE}}(x_k) - \widehat{p}_{\mathrm{MLE}}(x_k) \log p(x_k)) \quad \text{(merge sums)} \quad (70)$$

$$= -\sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k)(\log \widehat{p}_{\mathrm{MLE}}(x_k) - \log p(x_k)) \quad \text{(factor out } \widehat{p}_{\mathrm{MLE}}(x_k)) \quad (71)$$

$$= -\sum_{k=1}^{K} \widehat{p}_{\mathrm{MLE}}(x_k) \log \frac{\widehat{p}_{\mathrm{MLE}}(x_k)}{p(x_k)} \quad (\log \text{ algebra}) \quad (72)$$

$$= -\mathrm{KL}(\widehat{p}_{\mathrm{MLE}} \,||\, p) \quad \text{(defn.)} \quad (73)$$

which is the desired result. □

Lemma 2. *Define* $\Delta(x) = p(x) - q(x)$. *The second-order Taylor expansion of* $\mathrm{KL}(p \,||\, q)$ *around* $\Delta(x)$ *is given by*

$$\mathrm{KL}(p \,||\, q) = \sum_{x \in \mathcal{X}} \frac{\Delta(x)^2}{2q(x)} + o\left(\Delta(x)^2\right) \quad (74)$$

Proof. Now we compute the series expansion of the KL-divergence. We first make a tricky substitution:

$$\frac{p(x)}{q(x)} = \frac{q(x) + p(x) - q(x)}{q(x)} = 1 + \frac{p(x) - q(x)}{q(x)} = 1 + \frac{\Delta(x)}{q(x)} \quad (75)$$

Now, we proceed with the derivation:

$$\mathrm{KL}(p \,||\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad \text{(defn. of KL divergence)} \quad (76)$$

$$= \sum_{x \in \mathcal{X}} (q(x) + \Delta(x)) \log \left(1 + \frac{\Delta(x)}{q(x)}\right) \quad \text{(Eq. (75))} \quad (77)$$

$$= \sum_{x \in \mathcal{X}} (q(x) + \Delta(x)) \left(\frac{\Delta(x)}{q(x)} - \frac{\Delta(x)^2}{2q(x)^2} + o\left(\Delta(x)^2\right)\right) \quad \text{(Taylor expansion)} \quad (78)$$

$$= \sum_{x \in \mathcal{X}} \Delta(x) - \frac{\Delta(x)^2}{2q(x)} + \frac{\Delta(x)^2}{q(x)} - \frac{\Delta(x)^3}{2q(x)^2} + o\left(\Delta(x)^2\right) \quad \text{(distribute)} \quad (79)$$

$$= \sum_{x \in \mathcal{X}} \Delta(x) - \frac{\Delta(x)^2}{2q(x)} + \frac{\Delta(x)^2}{q(x)} + o\left(\Delta(x)^2\right) \quad \text{(defn. of } o) \quad (80)$$

$$= \sum_{x \in \mathcal{X}} \Delta(x) + \frac{\Delta(x)^2}{2q(x)} + o\left(\Delta(x)^2\right) \quad \text{(algebra)} \quad (81)$$

$$= \underbrace{\sum_{x \in \mathcal{X}} \Delta(x)}_{=0} + \sum_{x \in \mathcal{X}} \frac{\Delta(x)^2}{2q(x)} + o\left(\Delta(x)^2\right) \qquad \text{(split sums)} \tag{82}$$

$$= \sum_{x \in \mathcal{X}} \frac{\Delta(x)^2}{2q(x)} + o\left(\Delta(x)^2\right) \tag{83}$$

which is the desired result. □

E.3 Jackknife

The jackknife resampling method is used to estimate the bias of an estimator and correct for it, by sampling all subsamples of size $N-1$ from the available sample of size N, computing their average for the statistic being estimated.

Generally, this reduces the order of the bias of an estimator from $O(N^{-1})$ to at most $O(N^{-2})$ (Friedl and Stampfer, 2002).

Estimator 2 (Jackknife)**.** *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Let $\widehat{\mathrm{H}}^{\backslash n}(\mathcal{D})$ be an estimate of the entropy from a sample with the n^{th} observation held out. Then, the **Jackknife estimator** is given by*

$$\widehat{\mathrm{H}}_{\text{JACK}}(\mathcal{D}) \stackrel{\text{def}}{=} N\,\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) - \frac{N-1}{N} \sum_{n=1}^{N} \widehat{\mathrm{H}}^{\backslash n}_{\text{MLE}}(\mathcal{D}) \tag{84}$$

This estimator is derived from the jackknife-resampled estimate of the bias of the MLE estimator, multiplied by $N-1$.

$$\widehat{\mathrm{H}}_{\text{JACK}}(\mathcal{D}) - \widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) = (N-1)\left(\widehat{\mathrm{H}}_{\text{MLE}}(\mathcal{D}) - \frac{1}{N}\sum_{n=1}^{N} \widehat{\mathrm{H}}^{\backslash n}_{\text{MLE}}(\mathcal{D})\right) \tag{85}$$

E.4 Horvitz–Thompson

Horvitz and Thompson (HT; 1952) is a common estimator given a finite universe, which is our case as K is finite. We omit a derivation a full here as it is well documented in other places (Vieira, 2017). However, we note that, in contrast to many applications of HT, the application of HT to entropy estimation results in a biased estimator as the function whose mean we seek to estimate is $\log p(x_k)$, which is dependent on the unknown distribution p.

Estimator 3 (Horvitz–Thompson)**.** *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Then the **Horvitz–Thompson estimator** is defined as*

$$\widehat{\mathrm{H}}_{\text{HT}}(\mathcal{D}) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} \frac{\widehat{p}_{\text{MLE}}(x_k) \log \widehat{p}_{\text{MLE}}(x_k)}{1 - (1 - \widehat{p}_{\text{MLE}}(x_k))^N} \tag{86}$$

*where $1 - (1 - \widehat{p}_{\text{MLE}}(x_k))^N$ is an estimate of the **inclusion probability**, i.e., the probability that x_k appears in a random sample $\mathcal{D}$ of size N.*

We do not know of a simple expression for the bias of the Horvitz–Thompson entropy estimator, but one observation is that $\mathbb{E}_p\left[(1 - \widehat{p}_{\text{MLE}}(x_k))^N\right] > \mathbb{E}_p\left[(1 - p(x_k))^N\right]$ when $N > 1$ (justified by Jensen's inequality, since x^N, $N > 1$ is convex over $[0, 1]$); this is an overestimate of the true inclusion probability.

E.5 Chao–Shen

The Chao–Shen estimator builds upon Horvitz–Thompson by noting that that estimator does not correct for underestimation of number of classes K and resulting effect on estimates of $p(x_k)$; i.e. $1-(1-\widehat{p}_{\text{MLE}}(x_k))^N$ is always 0 for a class not included in the sample even if the class is present in the true distribution. We can reweight the sample probabilities to compensate for missing classes using the notion of sample coverage.

Definition 1 (Sample coverage). *We define the **sample coverage** as*

$$C = \sum_{k=1}^{K} p(x_k) \mathbb{1}\Big\{x_k \in \mathcal{D}\Big\} \tag{87}$$

Definitionally, $(1 - C)$ is then the probability of sampling an x_k not *observed in the sample $\widetilde{\mathcal{X}}$.*

However, exact computation of Eq. (88) is impossible as we do not know the true distribution p. Thus, Chao and Shen (2003) fall back on a well-known estimator of C that uses a technique from Good–Turing (1953) smoothing. Let f_1 be the number of classes with only one observation in the current sample, i.e, the number of singletons, then we can estimate the sample coverage as

$$\widehat{C} \stackrel{\text{def}}{=} 1 - \frac{f_1}{N} \tag{88}$$

The Chao–Shen estimator, described below, simply re-scales the MLE estimate of probability $\widehat{p}_{\text{MLE}}(x_k)$ in the HT estimator by $\widehat{C}$. This corrects for the observed *under*estimation of p's entropy by HT.

Estimator 4 (Chao–Shen). *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Let $\widehat{C}$, an estimate of sample coverage, be defined as in Eq. (88). The **Chao–Shen estimator** is then defined as*

$$\widehat{\mathrm{H}}_{\text{CS}}(\mathcal{D}) \stackrel{\text{def}}{=} -\sum_{k=1}^{K} \frac{\widehat{C} \cdot \widehat{p}_{\text{MLE}}(x_k) \log\left(\widehat{C} \cdot \widehat{p}_{\text{MLE}}(x_k)\right)}{1 - (1 - \widehat{C} \cdot \widehat{p}_{\text{MLE}}(x_k))^N} \tag{89}$$

E.6 Wolpert–Wolf

Fact 1 (Derivative of an exponent).

$$\frac{\mathrm{d}}{\mathrm{d}a} x^a = x^a \log x \tag{90}$$

Fact 2 (Normalizer of a Dirichlet). *The normalizer of a Dirichlet distribution is*

$$\int \delta\left(\sum_{k=1}^{K} x_k - 1\right) \prod_{k=1}^{K} x^{\alpha_k} \,\mathrm{d}\boldsymbol{x} = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \tag{91}$$

A relatively easy proof of this fact makes use of a Laplace transform.

Estimator 5 (Wolpert–Wolf). *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Then, the **Wolpert–Wolf estimator** is given by*

$$\widehat{\mathrm{H}}_{\text{ww}}(\mathcal{D} \mid \boldsymbol{\alpha}) \stackrel{\text{def}}{=} \psi\left(\widetilde{A} + 1\right) - \sum_{k=1}^{K} \frac{\widetilde{\alpha}_k}{\widetilde{A}} \psi(\widetilde{\alpha}_k + 1) \tag{92}$$

where $c(x_k) \stackrel{\text{def}}{=} \sum_{n=1}^{N} \mathbb{1}\{\widetilde{x}_n = x_k\}$, and we additionally define $\widetilde{\alpha}_k \stackrel{\text{def}}{=} c(x_k) + \alpha_k$ and $\widetilde{A} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \widetilde{\alpha}_k$.

Proposition 3 (Wolpert–Wolf). *The expectation of entropy under a Dirichlet posterior $\mathrm{Dirichlet}(\boldsymbol{\alpha})$ where parameter $\boldsymbol{\alpha}$ is given by*

$$\mathbb{E}\left[\mathrm{H}(p) \mid \boldsymbol{\alpha}\right] \stackrel{\text{def}}{=} \int \mathrm{H}(p)\, \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \frac{\Gamma(A)}{\prod_{k=1} \Gamma(\alpha_k)} \prod_{k=1}^{K} p(x_k)^{\alpha_k - 1} \mathrm{d}p \tag{93}$$

$$= \psi(A + 1) - \sum_{k=1}^{K} \frac{\alpha_k}{A} \psi(\alpha_k + 1) \tag{94}$$

where $A \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k$.

Proof. Let $\mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ be a Dirichlet posterior. The result follows by a series of manipulations:

$$\mathbb{E}\left[\mathrm{H}(p) \mid \boldsymbol{\alpha}\right] = \int \mathrm{H}(p)\, \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \prod_{k=1}^{K} p(x_k)^{\alpha_k - 1} \mathrm{d}p \qquad \text{(defn.)} \quad (95)$$

$$= \frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \int \mathrm{H}(p)\, \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \prod_{k=1}^{K} p(x_k)^{\alpha_k - 1} \mathrm{d}p \qquad (96)$$

$$= \frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \int \left(-\sum_{k=1}^{K} p(x_k) \log p(x_k)\right) \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \prod_{k=1}^{K} p_k^{\alpha_k - 1} \mathrm{d}p \qquad \text{(defn. H)} \quad (97)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \int p(x_k) \log p(x_k) \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \prod_{k=1}^{K} p(x_k)^{\alpha_k - 1} \mathrm{d}p \qquad \text{(linear.)} \quad (98)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \int p(x_k)^{\alpha_k} \log p(x_k) \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \prod_{\substack{j=1,\\ j\neq k}}^{K} p(x_j)^{\alpha_j - 1} \mathrm{d}p \qquad \text{(algebra)} \quad (99)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \int \frac{\mathrm{d}}{\mathrm{d}\alpha_k} p(x_k)^{\alpha_k} \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) \prod_{\substack{j=1,\\ j\neq k}}^{K} p(x_j)^{\alpha_j - 1} \mathrm{d}p \qquad \text{(fact \#1)} \quad (100)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \int \frac{\mathrm{d}}{\mathrm{d}\alpha_k} \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) p(x_k)^{\alpha_k} \prod_{\substack{j=1,\\ j\neq k}}^{K} p(x_j)^{\alpha_j - 1} \mathrm{d}p \qquad \text{(algebra)} \quad (101)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \frac{\mathrm{d}}{\mathrm{d}\alpha_k} \int \delta\left(\sum_{k=1}^{K} p(x_k) - 1\right) p(x_k)^{\alpha_k} \prod_{\substack{j=1,\\ j\neq k}}^{K} p(x_j)^{\alpha_j - 1} \mathrm{d}p \qquad (102)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \frac{\mathrm{d}}{\mathrm{d}\alpha_k} \frac{\Gamma(\alpha_k + 1) \prod_{\substack{j=1,\\ j\neq k}}^{K} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)} \qquad \text{(fact \#2)} \quad (103)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \prod_{\substack{j=1,\\ j\neq k}}^{K} \Gamma(\alpha_j) \frac{\mathrm{d}}{\mathrm{d}\alpha_k} \frac{\Gamma(\alpha_k + 1)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)} \qquad (104)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \prod_{\substack{j=1,\\ j\neq k}}^{K} \Gamma(\alpha_j) \frac{\psi(\alpha_k + 1)\Gamma(\alpha_k + 1)\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)^2} \qquad \text{(derivative)} \quad (105)$$

$$- \frac{\psi(\sum_{j=1}^{K} \alpha_j + 1)\Gamma(\alpha_k + 1)\Gamma(\sum_{j=1}^{K} \alpha_k + 1)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)^2}$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \prod_{\substack{j=1,\\ j\neq k}}^{K} \Gamma(\alpha_j) \frac{\begin{array}{c}\psi(\alpha_k + 1)\Gamma(\alpha_k + 1) \\ - \psi(\sum_{j=1}^{K} \alpha_j + 1)\Gamma(\alpha_k + 1)\end{array}}{\Gamma\left(\sum_{j=1}^{K} \alpha_j + 1\right)} \qquad \text{(simplify)} \quad (106)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)} \sum_{k=1}^{K} \prod_{\substack{j=1,\\ j\neq k}}^{K} \Gamma(\alpha_j) \frac{\begin{array}{c}\psi(\alpha_k + 1)\Gamma(\alpha_k)\alpha_k \\ - \psi(\sum_{j=1}^{K} \alpha_j + 1)\Gamma(\alpha_k)\alpha_k\end{array}}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right) A} \qquad \text{(defn. } \Gamma) \quad (107)$$

$$= -\frac{\Gamma(A)}{\prod_{k=1}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(A)}\sum_{k=1}^{K}\left(\frac{\alpha_k}{A}\psi(\alpha_k+1)-\frac{\alpha_k}{A}\psi\left(\sum_{k=1}^{K}\alpha_k+1\right)\right) \quad \text{(distrib.)} \quad (108)$$

$$= -\sum_{k=1}^{K}\left(\frac{\alpha_k}{A}\psi(\alpha_k+1)-\frac{\alpha_k}{A}\psi\left(\sum_{k=1}^{K}\alpha_k+1\right)\right) \quad \text{(cancel)} \quad (109)$$

$$= -\sum_{k=1}^{K}\left(\frac{\alpha_k}{A}\psi(\alpha_k+1)-\frac{\alpha_k}{A}\psi\left(A+1\right)\right) \quad \text{(defn. } A\text{)} \quad (110)$$

$$= -\sum_{k=1}^{K}\frac{\alpha_k}{A}\psi(\alpha_k+1)+\sum_{k=1}^{K}\frac{\alpha_k}{A}\psi\left(A+1\right) \quad \text{(distrib.)} \quad (111)$$

$$= -\sum_{k=1}^{K}\frac{\alpha_k}{A}\psi(\alpha_k+1)+\psi\left(A+1\right) \quad (\textstyle\sum a_k = A) \quad (112)$$

$$= \psi\left(A+1\right)-\sum_{k=1}^{K}\frac{\alpha_k}{A}\psi(\alpha_k+1) \quad \text{(rearr.)} \quad (113)$$

which proves the result. □

E.7 Nemenman–Shafee–Bialek

Estimator 6 (Nemenman–Shafee–Bialek)**.** *Let p be a categorical over K categories. We seek to estimate the entropy $\mathrm{H}(p)$. Let $\mathcal{D}$ be our dataset of size N sampled from p. Define the NSB density as*

$$p_{\text{NSB}}(\alpha) \stackrel{\text{def}}{=} \frac{K\psi_1\left(K\alpha+1\right)-\psi_1(\alpha+1)}{\log K} \tag{114}$$

*where ψ_1 is the trigramma function. Then, the **NSB estimator** is given by*

$$\widehat{\mathrm{H}}_{\text{NSB}}(\mathcal{D}) \stackrel{\text{def}}{=} \int_0^\infty \widehat{\mathrm{H}}_{\text{WW}}(\mathcal{D} \mid \alpha\cdot\mathbf{1})\, p_{\text{NSB}}(\alpha)\,\mathrm{d}\alpha \tag{115}$$

The integral in Eq. (115) is typically computed by numerical integration.

To derive the Nemenman–Shafee–Bialek (NSB) estimator, we start with the idea that we would like a prior over distributions such that the distribution over expected entropy is uniform. In other words, we are looking for a p_{NSB} such that for $\alpha \sim p_{\text{NSB}}$, the values of $\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right]$ are uniformly distributed over $[0, \log K]$. This is a good idea since, a-priori, we do not know entropy of p and, in the absence of any insight, we should assume the entropy could be anywhere in the range $[0, \log K]$. We make the above intuition formal with the following proposition.

Proposition 4. *Let p_{NSB} be the NSB density given in Eq. (114). Then the following conditional expectation*

$$\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right] \stackrel{\text{def}}{=} \int \mathrm{H}(p)\,\delta\left(\sum_{k=1}^{K}p(x_k)-1\right)\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}p(x_k)^{\alpha-1}\,\mathrm{d}p \tag{116}$$

$$= \psi\left(K\alpha+1\right)-\psi(\alpha+1) \qquad \textit{(Proposition 3)} \tag{117}$$

is uniformly distributed over $[0, \log K]$ when $\alpha \sim p_{\text{NSB}}(\cdot)$, defined in Eq. (114).

Proof. First, we note that $\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right]$ is a continuous, increasing function in α. We will not prove this formally, but it should make intuitive sense: α is a smoothing parameter and the more the distribution is smoothed, the more entropic it should be. From basic analysis, we know that a strictly continuous, increasing function has an inverse. The above means that we can view $\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right]$ as a bijection from $\mathbb{R}_{\geq 0}$ to the interval $[0, \log K]$. Our goal is to reparameterize the Uniform distribution in terms of α. To that end, we

define the function $g^{-1}(\alpha) \stackrel{\text{def}}{=} \mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right] : \mathbb{R}_{\geq 0} \to [0, \log K]$ and perform a change-of-variables transform on Eq. (118) using g^{-1}. We start with the continuous uniform over $[0, \log K]$, which is show below

$$p(H) \stackrel{\text{def}}{=} \underbrace{\frac{1}{\log K} \mathbb{1}\left\{H \in [0, \log K]\right\}}_{\text{uniform over } [0, \log K]} \qquad \text{(defn. of uniform dist)} \tag{118}$$

Note H is a random variable and unrelated to the functional $\mathrm{H}(\cdot)$; the choice of letter intentionally reminds one that the variable represents the expected entropy of under a random distribution. Now we apply the change-of-variables formula at $H = g^{-1}(\alpha)$ and manipulate:

$$\begin{aligned}
p(H) &= p(g^{-1}(\alpha)) \left|\frac{\mathrm{d}g^{-1}}{\mathrm{d}\alpha}(\alpha)\right| && \text{(change of variable)} && (119)\\
&= \frac{1}{\log K} \mathbb{1}\left\{g^{-1}(\alpha) \in [0, \log K]\right\} \left|\frac{\mathrm{d}g^{-1}}{\mathrm{d}\alpha}(\alpha)\right| && \text{(definition of } p\text{)} && (120)\\
&= \frac{1}{\log K} \left|\frac{\mathrm{d}g^{-1}}{\mathrm{d}\alpha}(\alpha)\right| && \text{(redundant indicator)} && (121)\\
&= \frac{1}{\log K} \frac{\mathrm{d}g^{-1}}{\mathrm{d}\alpha}(\alpha) && \text{(derivative is positive)} && (122)\\
&= \frac{K\psi_1\left(K\alpha + 1\right) - \psi_1(\alpha + 1)}{\log K} && \text{(Lemma 3)} && (123)\\
&\stackrel{\text{def}}{=} p_{\text{NSB}}(\alpha) && \text{(definition)} && (124)
\end{aligned}$$

By construction, the prior $p_{\text{NSB}}(\alpha)$ has the property that the expected entropy $\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right]$ where $\alpha \sim p_{\text{NSB}}(\cdot)$ is uniformly distributed over $[0, \log K]$, which we can see by reversing the above derivation. This proves the result. □

Nemenman et al. (2002) interpreted Proposition 4 in the following manner: As the variance of $\mathbb{E}_p\left[\mathrm{H}(p) \mid \alpha\right]$, which is treated as a random variable since α is random, approaches 0, then the the NSB estimator implies a uniform prior over the entropy.

Lemma 3 (NSB Derivative)**.**

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\psi(K\alpha + 1) - \psi(\alpha + 1)\right] = K\psi_1(K\alpha + 1) - \psi_1(\alpha + 1) \tag{125}$$

Proof. The proof follows by a straightforward computation:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\psi(K\alpha + 1) - \psi(\alpha + 1)\right] &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\psi(K\alpha + 1)\right] - \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\psi(\alpha + 1)\right] && \text{(linearity)} && (126)\\
&= K\psi_1(K\alpha + 1) - \psi_1(\alpha + 1) && \text{(definition)} && (127)
\end{aligned}$$

where $\psi_1(x) \stackrel{\text{def}}{=} \frac{\mathrm{d}}{\mathrm{d}x}\psi(x)$. □