Title

Learning Structured Distributions: Power-Law and Low-Rank

Author

Falahatgar, Moein

Publication Date

2019

UNIVERSITY OF CALIFORNIA SAN DIEGO

Learning Structured Distributions: Power-Law and Low-Rank

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Machine Learning and Data Science)

by

Moein Falahatgar

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Sanjoy Dasgupta
Professor Tara Javidi
Professor Ndapa Nakashole
Professor Ken Zeger

2019

The dissertation of Moein Falahatgar is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To my Parents: Azam and Baba.

EPIGRAPH

God has shown me that it is a scientific fact that gratitude reciprocates.

—Matthew McConaughey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my parents and my brothers, Mohammad and Mostafa, whose support by no means can be put into words to be described and acknowledged. My family has continuously supported me during my studies. Acknowledging them here is the simplest I can do to thank them for all their cheering of every single accomplishment I had in my life.

I was so fortunate to have numerous caring friends. Friendships initiated in my undergraduate studies as well as friendships made after moving to the US and starting my PhD. There is a long list of friends to whom I am indebted for their genuine friendships and it is not simply possible to list all here. I would like to show my earnest appreciation to Masoud Behzadi, Navid Naderi, and Arian Shahnazari for their invaluable support, advice, and care. I would like to express my sincere thanks to my friends whom I have lived with and shared memorable moments during the past six years: Erfan Sayyari, Shahab Sarmashghi, Hamid Fard, and Armin Ahmadi. I would also like to thank Ali Akbari, Borhan Vasli, Hossein Fard, Aref Shiran, and Roham Razzaghi for all the pleasant moments we had together and specially thank Nasrin Shariati for her considerate support and care.

At UC San Diego, I was fortunate to have great teachers and mentors. I am thankful to my PhD committee members: Prof. Sanjoy Dasgupta, Prof. Tara Javidi, Prof. Young-Han Kim, Prof. Ndapa Nakashole, and Prof. Ken Zeger.

I would like to sincerely thank my colleague and collaborator, Prof. Mesrob Ohannessian for sharing his wisdom and knowledge with me and teaching me the concepts of true research through our joyful collaborations. I will be indebted to him for all his support. I also want to thank my current and previous colleagues: Ashkan Jafarpour, Ananda Theertha Suresh, Jayadev Acharya, Yi Hao, Vaishakh Ravindrakumar, and Ayush Jain. I could not show how thankful I am for all the fruitful discussions I had with all my colleagues. I would like to show my special gratitude to my colleague and friend, Dheeraj Pichapati. For the entire of my PhD years, I was fortunate to have Dheeraj by my side as my colleague, collaborator, and a friend.

My sincerest acknowledgements will go to Alon. I met him when I took the Information Theory course in my first quarter at UC San Diego. I immediately fell in love with his teaching. Joining his research group shortly after, I realized that his research method was as pleasant as his teaching. Being fortunate to be part of his research group, I was provided the opportunity to meet many intellectual researchers and get involved in one of the biggest conferences in Information Theory and Applications, ITA. My appreciation of interacting with Alon goes beyond academic aspects as I constantly learned from him in every aspect of life. However, the space here cannot do the justice to show my gratitude of having him as my advisor.

Chapter 2, in part, is a reprint of the material as it appears in *2015 IEEE International Symposium on Information Theory (ISIT)*. Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh, 2015. The dissertation author was the primary investigator and author of this paper. This chapter has also been submitted with the same authors for publication of the material as it may appear in the journal of IEEE Transactions on Information Theory.

Chapter 3, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati, 2017. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Falahatgar, Moein, Mesrob I. Ohannessian, and Alon Orlitsky, 2016. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 5, in full, has been submitted for the publication of the material as it may appear in *Submitted to Neurips*. Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati, 2019. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2013	B. Sc. in Electrical Engineering, Sharif University of Technology, Iran
2016	M.S. in Electrical Engineering (Communication Theory and Systems), University of California San Diego
2019	Ph. D. in Electrical Engineering (Machine Learning and Data Science), University of California San Diego

PUBLICATIONS

Falahatgar, Moein, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. "Maximum selection and ranking under noisy comparisons." *Proceedings of the 34th International Conference on Machine Learning*-Volume 70, pp. 1088-1096. JMLR. org, 2017.

Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. "Faster algorithms for testing under conditional sampling." *Conference on Learning Theory*, pp. 607-636. 2015.

Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. "Estimating the number of defectives with group testing." *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1376-1380. IEEE, 2016.

Falahatgar, Moein, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. "Maxing and ranking with few assumptions." *Advances in Neural Information Processing Systems*, pp. 7060-7070. 2017.

Falahatgar, Moein, Mesrob I. Ohannessian, and Alon Orlitsky. "Near-optimal smoothing of structured conditional probability matrices." *Advances in Neural Information Processing Systems*, pp. 4860-4868. 2016.

Falahatgar, Moein, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. "Learning markov distributions: Does estimation trump compression?." *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 2689-2693. IEEE, 2016.

Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. "Universal compression of power-law distributions." *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2001-2005. IEEE, 2015.

Falahatgar, Moein, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. "The limits of maxing, ranking, and preference learning." *International Conference on Machine Learning*, pp. 1426-1435. 2018.

Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati. “The power of absolute discounting: all-dimensional distribution estimation.” *Advances in Neural Information Processing Systems*, pp. 6660-6669. 2017.

Acharya, Jayadev, Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. “Maximum selection and sorting with adversarial comparators.” *The Journal of Machine Learning Research* 19, no. 1 (2018): 2427-2457.

Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati. “Towards Competitive N-gram Smoothing.” *Submitted to Neurips*, 2019.

Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. “Universal compression of power-law distributions.” *In preparation*.

ABSTRACT OF THE DISSERTATION

Learning Structured Distributions: Power-Law and Low-Rank

by

Moein Falahatgar

Doctor of Philosophy in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2019

Professor Alon Orlitsky, Chair

Utilizing the structure of a probabilistic model can significantly increase its compression efficiency and learning speed. We consider these potential improvements under two naturally-omnipresent structures.

Power-Law: English words and many other natural phenomena are well-known to follow a power-law distribution. Yet this ubiquitous structure has never been shown to help compress or predict these phenomena. It is known that the class of unrestricted distributions over alphabet of size k and blocks of length n can never be compressed with diminishing per-symbol redundancy, when $k > n$. We show that under power-law structure, in expectation we can compress with diminishing per-symbol redundancy for k growing as large as sub-exponential in n.

For learning a power-law distribution, we rigorously explain the efficacy of the absolute-discount estimator using less pessimistic notions. We show that (1) it is *adaptive* to an effective dimension and (2) it is strongly related to the Good–Turing estimator and inherits its *competitive* properties.

Low-Rank: We study learning low-rank conditional probability matrices under expected KL-risk. This choice accentuates smoothing, the careful handling of low-probability elements. We define a loss function, determine sample-complexity bound for its global minimizer, and show that this bound is optimal up to logarithmic terms. We propose an iterative algorithm that extends classical non-negative matrix factorization to naturally incorporate additive smoothing and prove that it converges to the stationary points of our loss function.

Power-Law and Low-Rank: We consider learning distributions in the presence of both low-rank and power-law structures. We study Kneser-Ney smoothing, a successful estimator for the N-gram language models through the lens of competitive distribution estimation. We first establish some competitive properties for the contextual probability estimation problem. This leads to *Partial Low Rank*, a powerful generalization of Kneser-Ney that we conjecture to have even stronger competitive properties. Empirically, it significantly improves the performance on language modeling, even matching the feed-forward neural models, and gives similar gains on the task of predicting attack types for the Global Terrorism Database.

Chapter 1

Introduction

Data Compression and *Distribution Estimation* are two of the most fundamental and classical problems in Information Theory and Learning Theory. These problems have been widely studied when many observations are available, and impossibility results were derived for the few-samples regime.

In this dissertation we exploit structure in the data to better learn and compress it. We study two important structures: *power-law* and *low-rank*. For both problems we derive surprising results that may help explain why, despite the pessimistic theoretical results to the contrary, humans can grasp distributions even over very large domains.

1.1 Universal compression of power-law distributions

The fundamental data-compression theorem limits the compression of any source to its Shannon Entropy, yet in real applications the source distribution is often not known. This requires *Universal Compression* algorithms that applies to all distributions and certainly leads to an increase over entropy known as *redundancy*. It is well known that when the alphabet size k is larger than the sample size n, source symbols cannot be compressed with diminishing per-symbol redundancy [OSZ04].Our hope therefore, has been to exploit the structure of the data to derive efficient data-compression algorithms with vanishing per-symbol redundancy even for the $k > n$ regime.

One of the most common distribution structures, discovered by linguist George Zipf in 1935, is power-law distributions, also known as the rich-get-richer phenomenon, or $80-20$ rule [Zip13]. It has been observed in the flagship application of Natural Language Processing, as well as in distributions of species, genera, rainfall, terror incidents, and many more [New05]. In power-law distribution with power α the i'th largest probability is proportional to $\frac{1}{i^\alpha}$.

We show that under power-law structure we can compress data sources with diminishing per-symbol redundancy, even when the alphabet size grows nearly-exponentially in the sample size, namely, $k = 2^{n^{1-\frac{1}{\alpha}}}$. This may explain why humans can grasp English distribution even when its alphabet size, nearly a million words, is significantly larger than the number of times a person may have seen the given context.

1.2 Learning power-law distributions

Absolute-discounting has long been used in Natural Language Processing to accurately estimate the probability of a word in a context [CG99]. But why it works so well was never properly determined. Classical minimax redundancy results, bound the KL-risk of the whole k-dimension distribution-simplex in terms of the sample size n [BS04, Pan04]. However, when the distribution class is only a small part of the simplex, these guarantees are pessimistic. In this dissertation we analyze the performance of absolute-discounting for a family of distribution classes defined by the expected number of distinct elements, d, that may appear in n samples. In fact, d acts as an effective dimension of the data. We show several results:

- For power-law distributions, absolute-discounting is strongly related to the well-known Good-Turing estimator, therefore, benefiting from its competitive properties. Put differently, for power-law distributions, absolute-discounting matches the performance of the best estimator up to a vanishing additive term. For a power-law distribution with power α, this additive term is negligible: $O(n^{-\frac{2\alpha-1}{2\alpha+1}})$, hence absolute-discounting is min-max optimal for power-law distributions.

- Absolute-discounting adapts to the effective data dimension d. For example, instead of the minimax bound of $\frac{k-1}{2n}$ in the range $n > k$, we derive the bound $O(\frac{d}{n})$ where d is an upper bound on the expected number of distinct elements for distributions in the class.

- This bound recovers classical minimax KL-risk rates in all ranges of k and n. For example, when $n > k$, d can be at most k, and therefore the minimax rate of $\frac{k}{n}$ for the whole simplex is recovered.

These results may explain the long-standing mystery of absolute-discounting success in NLP.

1.3 Learning low-rank probability matrices

The problem of estimating a one-dimensional probability distribution from observations has been widely studied in Information Theory [KOPS15]. However, many applications involve more complex distributions.

For example, in language modeling it is natural to assume a Markov or N-gram language model where each word depends on the preceding context. This leads to estimating a transition probability matrix, P. There are k^{N-1} rows in P, one for each context, and each row is a probability distribution over k elements. It can further be argued that because words may depend on the context through meaning and part of speech, the transition matrix may have low rank.

We therefore study *low-rank structure* of the data where the transition probability matrix P despite being a $k^{N-1} \times k$ matrix, it is of much lower rank, m. We consider learning low-rank conditional probability matrices under expected KL-risk. This choice makes smoothing–the careful handling of low-probability elements, paramount. We design an iterative algorithm that extends classical non-negative matrix factorization to naturally incorporate additive smoothing and prove that it converges to the stationary points of a penalized empirical risk. For a sample size of n, we bound the expected KL-risk by $\tilde{O}\left(\frac{(k+k^{N-1})m}{n}\right)$ for the global minimizer of the penalized risk. The KL-risk bound captures the right dependence on the number of parameters, $(k+k^{N-1})m$, instead of k^N in the absence of low-rank structure.

This framework generalizes to more sophisticated smoothing techniques, including absolute-discounting. Our experiments on real-data show that the resulting algorithms improve over several benchmarks such as the Kneser-Ney estimator.

1.4 Contextual competitive estimators

Recently, the prosperity of the well-celebrated Good-Turing estimator has been justified by the notion of competitive distribution estimation [OS15]. We extend this notion to the contextual case, namely, when we have multiple, possibly related distributions to estimate. Similar to the non-contextual case, we justify the performance a practical estimator, Kneser-Ney, through competitive notions.

Without assuming any explicit low-rank model assumption, Kneser-Ney estimator performs well even when the data is generated according to a low-rank model. Yet, no clear and complete understanding of the Kneser-Ney estimator has been suggested. Perhaps this is partly due to the surge of neural networks and in particular recurrent neural language models, which led to a significant jump in performance [MKB$^+$10].

Neural language models have since continued to achieve better results [MKS17, YDSC17, GHT$^+$18, TSN18, DYY$^+$19]. Interestingly, N-gram techniques are still relevant as they usually run much faster, and, can be used in conjunction with neural models to further improve performance [XWL$^+$17]. Moreover, for low-resource languages, non-neural methods or a combination of neural and non-neural methods are known to achieve the best performance [GML14].

Motivated by these reasons we investigate the first theoretical handle into N-gram models. As evidence of this exploration, we report on a powerful generalization of Kneser-Ney backoff that applies the rank structure only to the rare part of the data, and hence named *Partial Low-Rank*. Kneser-Ney corresponds to the rank-1 special case. We show that Partial Low-Rank uniformly improves over Kneser-Ney on various benchmarks. Also, a nested trigram-level implementation of this approach meets and slightly exceeds the performance of the best feed-forward neural models on the Penn Tree Bank data set. Furthermore, it can be trained with a fraction of the time and space resources required for the neural model.

Part of the novelty of our perspective is to study back-off through the lens of competitive distribution estimation. This notion was expressed most clearly in the result of [OS15], where it was used to give a clear justification to the Good-Turing estimator [Goo53], that is intimately related to Kneser-Ney.

1.5 Dissertation organization

The rest of this dissertation is organized as follows:

- In Chapter 2 we study the universal compression of power-law distribution classes.
- In Chapter 3 we study learning power-law distributions and show the equivalence of Good-Turing and absolute-discounting estimators for power-law classes that leads to competitive properties.
- In Chapter 4 we study learning low-rank conditional probability matrices and propose integrating smoothing with non-negative matrix factorization.
- In Chapter 5 we study competitive distribution estimators for probability matrices. We propose the Partial Low Rank, a generalization of Kneser-Ney estimator, that performs as well as feed-forward neural language models.

Chapter 2

Universal Compression of Power Law Distributions

2.1 Introduction

2.1.1 Definitions

The fundamental data-compression theorem states that every discrete distribution p can be compressed to its entropy $H(p) \overset{\text{def}}{=} \sum p(x) \log \frac{1}{p(x)}$, a compression rate approachable by assigning each symbol x a codeword of roughly $\log \frac{1}{p(x)}$ bits.

In reality, the underlying distribution is seldom known. For example, in text compression, we observe only the words, no one tells us their probabilities. In all these cases, it is not clear how to compress the distributions to their entropy.

The common approach to these cases is *universal compression*. It assumes that while the underlying distribution is unknown, it belongs to a known class of possible distributions, for example, *i.i.d.* or Markov distributions. Its goal is to derive an encoding that works well for all distributions in the class.

To move towards formalizing this notion, observe that every reasonable compression scheme for a distribution over a discrete set $\mathcal{X}$ corresponds to some distribution q over $\mathcal{X}$ where each symbol $x \in \mathcal{X}$ is assigned a codeword of length roughly $\log \frac{1}{q(x)}$. Hence the expected number of bits used to encode the distribution's output is $\sum p(x) \log \frac{1}{q(x)}$, and the penalty over the entropy minimum is $\sum p(x) \log \frac{p(x)}{q(x)}$ bits.

Let $\mathcal{P}$ be a collection of distributions over $\mathcal{X}$. The collection's *expected redundancy*, is the least worst-case increase in the expected number of bits over the entropy, where the worst case is taken over all distributions in $\mathcal{P}$ and the least is minimized over all possible encoders,

$$\bar{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_q \max_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

An even stricter measure of the increased encoding length due to not knowing the distribution is the collection's *worst-case redundancy* that considers the worst increase not just over all distributions, but also over all possible outcomes x,

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_q \max_{p \in \mathcal{P}} \max_{x \in \mathcal{X}} \log \frac{p(x)}{q(x)}.$$

Clearly,

$$\bar{R}(\mathcal{P}) \leq \hat{R}(\mathcal{P}).$$

Interestingly, until now, except for some non-natural examples, all analyzed collections had extremely close expected and worst-case redundancies. One of our contributions is to demonstrate a practical collection where these redundancies vastly differ, hence achieving different optimization goals may require different encoding schemes.

By far the most widely studied are the collections of *i.i.d.* distributions. For every distribution p, the *i.i.d.* distribution p^n assigns to a length-n string $x^n \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_n)$ probability

$p(x^n) = p(x_1) \cdot \ldots \cdot p(x_n)$. For any collection $\mathcal{P}$ of distributions, the length-n *i.i.d.* collection is

$$\mathcal{P}^n \overset{\text{def}}{=} \{p^n : p \in \mathcal{P}\}.$$

2.1.2 Previous results

Let Δ_k denote the collection of all distribution over $\{1,\ldots,k\}$, where Δ was chosen to represent the simplex. In the first few decades of universal compression, researchers studied the redundancy of Δ_k^n [KT81, Kie78, Dav73, DMPW81, WST95, XB00, SW10, OS04, Ris96, Cov91, Szp98, SW12]. In particular, [KT81] showed that for $k = o(n)$,

$$\hat{R}(\Delta_k^n) = \frac{k-1}{2}\log\frac{n}{k} + \frac{k}{2}\log e + o(k),$$

and for the complementary regime where $n = o(k)$, [OS04] showed that

$$\hat{R}(\Delta_k^n) = n\log\frac{k}{n} - \log e + O\left(\frac{1}{n}\right).$$

These results show that redundancy grows only logarithmically with the sequence length n when $k = o(n)$, therefore for long sequences, the per-symbol redundancy diminishes to zero and the underlying distribution needs not to be known to approach entropy. As is also well known, expected redundancy is exactly the same as the log loss of sequential prediction, hence these results also show that prediction can be performed with very small log loss.

However, as intuition suggests, and these equations confirm, redundancy increases sharply with the alphabet size k. In many, and possibly most, important real-life applications, the alphabet size is larger than the block length. This is the case for example in applications involving natural language processing, population estimation, and genetics [CG99]. The redundancy in these cases is therefore very large, and can be even unbounded for any sequence length n.

Over the last decade, researchers therefore considered methods that could cope with compression and prediction of distributions over large alphabets. Two main approaches were taken.

[OSZ04] separated compression of large-alphabet sequences into compression of their *pattern* that indicates the order at which symbols appeared, and *dictionary* that maps the order to the symbols. A sequence of papers [OSZ04, Sha06, Sha04, Gar09, OS04, ADO12, ADJ$^+$13] showed that patterns can be compressed with redundancy sub-linear in block length and most significantly, is uniformly upper bounded regardless of the alphabet size that can be even infinite. Note also, that for pattern redundancy, worst-case and expected redundancy are quite close.

The other approach is to use the properties of the distributions in the class. In most applications it happens that we know a rough behavior of distributions in the class. For example we know that distributions are monotone or follow an envelope function. A series of works studied class of *monotone* distributions [Sha13, AJOS14a]. More closely related to this work are *envelope classes* [BGG09, BGO14]. An *envelope* is a function $f : \mathbb{N}_+ \to \mathbb{R}_{\geq 0}$. For envelope function f,

$$\mathcal{P}_{\leq f} \overset{\text{def}}{=} \{p : p_i \leq f(i) \text{ for all } i \geq 1\}$$

is the collection of distributions where each p_i is at most the corresponding envelope bound $f(i)$. Some canonical examples are the power-law envelopes $f(i) = c \cdot i^{-\alpha}$, and the exponential envelopes $f(i) = c \cdot e^{-\alpha \cdot i}$. Recently, [AJOS14b] showed that

$$\hat{R}(\mathcal{P}_{\leq f}) = \Theta(n^{1/\alpha}).$$

The restricted-distribution approach has the advantage that it considers the complete sequence redundancy, not just the pattern. Yet it has the shortcoming that it may not capture relevant distribution collections. For example, most real distributions are not monotone, words starting with ‘a’ are not necessarily more likely than those starting with ‘b’. Similarly for say power-law envelopes, why should words in the early alphabet have higher upper bound than subsequent ones? Thus, words do not carry frequency order inherently. To address this issue, we propose a new class of distributions to capture the unknown inherent order of the symbols. And we show for the power-law envelop classes, under the new model, we have per-symbol vanishing expected redundancy for k as large as sub-exponential in the block length n.

2.1.3 Distribution model

We focus on power-law distributions, a ubiquitous class that is also of scientific interest. Power-law distribution with exponent α is denoted by zipf(α) and is defined as $\text{zipf}(\alpha)_i = \frac{1}{C_\alpha i^\alpha}$. C_α is the normalization factor required by the fact that the total probability should sum to 1, namely $C_\alpha = \sum_{i=1}^{\infty} \frac{1}{i^\alpha}$. We immediately can infer that if $\alpha \leq 1$ the sum would diverge and thus power-laws with exponent ≤ 1 cannot be normalized and rarely occur in nature, if ever happen [New05].

Perhaps the most well-known instance of power-laws was shown in [Zip13]. In 1935, linguist George Kingsley Zipf observed that when English words are sorted according to their probabilities, namely so that $p_1 \geq p_2 \geq \ldots$, the resulting distribution follows a power law, $p_i \sim \frac{c}{i^\alpha}$ for some constant c and power α.

Started long before Zipf and continued after him, researchers have found a very large number of distributions such as population ranks of cities [PBS$^+$60], earthquakes' magnitudes [GR44], computer files [CB97], the frequency of occurrence of personal names [ZM01], number of citations papers receive [YVdS65], hits on web pages [AH99], number of species in biological taxa [WY22] and people's annual income [Par64] that when sorted follow this *Zipf*-, or *power*-law. Figure 2.1, adapted from [New05] shows Zipf distribution in several natural classes. The plots in Figure 2.1 are for the cumulative distribution function; however, it is easy to show that if p follows a power-law with exponent α, the cumulative distribution function also follows a power-law but with exponent $\alpha - 1$.

Although the reason for power-law behavior has been under question for more than a

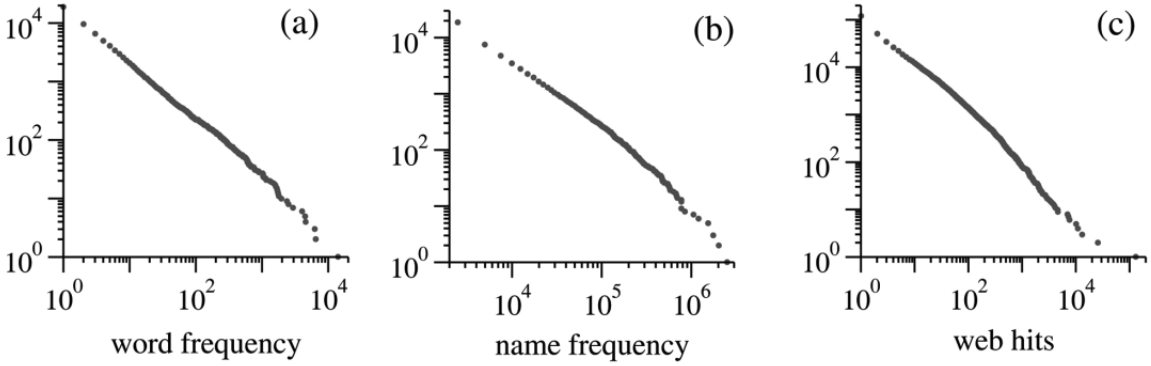


Figure 2.1: (Figure 4 in [New05]) Examples of Zipf law in natural classes. Rank-frequency plots of some quantities well known to follow a power-law. (a) Numbers of occurrences of words in the novel 'Moby Dick' by Hermann Melville. (b) Frequency of occurrence of family names in the US in the year 1990 . (c) Numbers of hits on web sites by 60000 users of the America Online Internet service for the day of 1 December 1997.

century, several explanations for this phenomena have been proposed in [Mit04, New05, Sor06]. In fact, a Google Scholar search for "power-law distribution" returns more than two million results.

2.1.4 Notation

Let $x^n \stackrel{\text{def}}{=} x_1, x_2, .., x_n$ be a sequence of n symbols from alphabet $\mathcal{X}$ of size k. The *multiplicity* μ_x of a symbol $x \in \mathcal{X}$ is the number of times x appears in x^n. Let $[k] = \{1, 2, ..., k\}$ be the indices of elements in $\mathcal{X}$. The type vector of x^n over $[k] = \{1, 2, ..., k\}$, $\tau^k \stackrel{\text{def}}{=} \tau(x^n) = (\mu_1, \mu_2, \ldots, \mu_k)$ is a k-tuple of multiplicities in x^n. The *prevalence* of a multiplicity μ, denoted by φ_μ, is the number of elements appearing μ times in x^n. For example, φ_1 denotes the number of elements which appeared once in x^n. Furthermore, φ_+ denotes the number of distinct elements in x^n.

For a distribution $p = (p_1, p_2, ..., p_k)$ let $p_{(i)}$ be the i^{th} largest probability. Namely, $p_{(1)} \geq p_{(2)} \geq \ldots \geq p_{(k)}$. Zipf distribution with parameter α and support k is denoted by $\text{zipf}(\alpha, k)$,

$$\text{zipf}(\alpha, k)_i = \frac{i^{-\alpha}}{C_{k,\alpha}},$$

where $C_{k,\alpha}$ is the normalization factor. Note that all logarithms in this chapter are in base 2 and we consider only the case $\alpha > 1$.

The rest of the chapter is organized as follows. In Section 2.2 we state the summary of our results. Next, in Section 2.3 we bound the worst-case redundancy for the power-law envelope class. In Section 2.4 we take a novel approach to analyze the expected redundancy. We introduce a new class of distributions with bounded expected number of distinct elements and provide upper and lower bounds on the expected redundancy of this class. In Sections 2.6 and 2.7 we study the expected redundancy of power-law classes.

2.2 Results

Inspired by the widely-appearing Zipf's law in numerous incidents, a natural question therefore is to ask whether this established and commonly trusted empirical observation can be used to better predict or equivalently compress them, and if so, by how much.

Note that if the alphabet size is not bounded, the redundancy will be infinite and therefore we limit ourselves to the alphabet of size k and study how redundancy changes as a function of k. We show that the class of Zipf distributions, namely

$$\mathcal{P}_{(\text{zipf}(\alpha,k))} = \{p : \ p_{(i)} = \text{zipf}(\alpha,k)_i \ \forall 1 \le i \le k\},$$

can be compressed with diminishing per-symbol redundancy. In fact we show that much larger power-law classes can be compressed universally for $k > n$.

In many applications, the exponent α in Zipf distribution is not exactly revealed, if ever known. Therefore we define a larger class $\mathcal{P}_{(\text{zipf}(\ge\alpha,k))}$ which contains all Zipf distributions and their permutations with exponent no less than α.

$$\mathcal{P}_{(\text{zipf}(\geq\alpha,k))} = \bigcup_{a\geq\alpha} \{p: \; p_{(i)} = \text{zipf}(a,k)_i \; \forall 1 \leq i \leq k\}.$$

Enlarging the collection of distributions to $\mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$ solves the problem of unknown exponent; however this modified class is still too restrictive since it is forcing the probabilities to be exactly a power-law function. In other words, the sorted distribution behaves like a power-law but it is a perturbed version of that in almost all cases. [BGG09] used the concept of envelope classes to model this impact. For an envelope f over the alphabet of size k, they define

$$\mathcal{P}_{\leq f} = \{p: \; p_i \leq f(i) \; \forall 1 \leq i \leq k\}.$$

Envelope distributions are very appealing as they represent our belief about the distribution [BGG09]. However the main drawback of the model is that the correspondence between the probabilities and symbols is assumed to be known, namely that $p_i \leq f(i)$ for the same i. We relax this requirement and assume only that an upper envelope on the sorted distribution, not the individual elements, is known. We define

$$\mathcal{P}_{(\leq f)} = \{p: \; p_{(i)} \leq f(i) \; \forall 1 \leq i \leq k\}.$$

Observe that $\mathcal{P}_{(\text{zipf}(\alpha,k))} \subseteq \mathcal{P}_{(\leq c\cdot i^{-\alpha})}$ for some constant c, and $\mathcal{P}_{(\text{zipf}(\alpha,k))} \subseteq \mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$ and therefore any lower bound on the redundancy of $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ is also a lower bound on the redundancy of $\mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$ and $\mathcal{P}_{(\leq c\cdot i^{-\alpha})}$.

To establish an upper bound on the expected redundancy of $\mathcal{P}_{(\leq c\cdot i^{-\alpha})}$, we follow a more general approach and define a larger class containing all *i.i.d.* distributions where the expected number of distinct elements in n samples, φ_+^n is at most d,

$$\mathcal{P}_{\leq d} = \{p : \mathbb{E}_p[\varphi_+^n] \leq d\}.$$

It can be shown that $\forall \mathcal{P} \in \{\mathcal{P}_{(\text{zipf}(\alpha,k))}, \mathcal{P}_{(\text{zipf}(\geq\alpha,k))}, \mathcal{P}_{(\leq c\cdot i^{-\alpha})}\}$ and an appropriate d,

$$\mathcal{P} \subseteq \mathcal{P}_{\leq d},$$

hence any upper bound on the redundancy of $\mathcal{P}_{\leq d}$ is also an upper bound on the redundancy of $\mathcal{P} \in \{\mathcal{P}_{(\text{zipf}(\alpha,k))}, \mathcal{P}_{(\text{zipf}(\geq\alpha,k))}, \mathcal{P}_{(\leq c\cdot i^{-\alpha})}\}$.

We establish an upper and lower bound on $\bar{R}(\mathcal{P}_{\leq d})$ and $\bar{R}(\mathcal{P}_{(\text{zipf}(\alpha,k))})$, respectively. We show that these bounds match up to constants for certain range of k, n, resolving the expected redundancy for all three power-law classes.

2.2.1 Main results

- [Theorem 1] For $k \geq n^{\alpha}$,
$$\hat{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) \geq n \log \frac{k}{n^{\alpha}}.$$
Significance: When $n \ll k$, the worst-case redundancy even for the smallest class, $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ behaves the same as that of general distributions, $n \log \frac{k}{n}$.

- [Theorems 22 and 23] For $n < k^{\frac{\alpha}{1.1}}$ and $\forall \mathcal{P} \in \{\mathcal{P}_{(\text{zipf}(\alpha,k))}, \mathcal{P}_{(\text{zipf}(\geq\alpha,k))}, \mathcal{P}_{(\leq c\cdot i^{-\alpha})}\}$,
$$\bar{R}(\mathcal{P}^n) = \Theta\left(n^{\frac{1}{\alpha}} \log \frac{k}{n^{\frac{1}{\alpha}}}\right).$$
Significance: When k is larger than n, but of the same order, the more practical expected redundancy of Zipf distributions of order $\alpha > 1$ is the $1/\alpha$ power of the expected redundancy of general distributions. For example, for $\alpha = 2$ and $k = 10n$, the redundancy of Zipf

distributions is $\Theta(\sqrt{n}\log n)$ compared to $\Theta(n)$ for Δ_k^n. This sub-linear dependence on n is valid for the entire range of $k \le 2^{n^{1-\frac{1}{\alpha}}}$ and also implies that unordered power-law distributions have vanishing per-symbol expected redundancy for this range. This shows a dichotomy between the worst-case and expected redundancy; making power-law classes the first natural class for which

  - the worst-case and expected redundancy significantly diverge, and
  - the worst-case redundancy is same as that of general distributions, while the expected redundancy differs a lot.

- [Theorem 7] For all k, n, d,

$$\bar{R}(\mathcal{P}^n_{\le d}) = \Theta\left(d \log \frac{\max\{k,n\}}{d}\right).$$

  Interpretation The redundancy of compressing a sequence comes from describing the type vector. First we declare how many distinct elements are in that sequence using $\log n$ bits. In addition, we need $\log\binom{k}{d}$ bits to specify which d distinct elements out of k elements appeared in the sequence. Finally, for the exact number of occurrences of each distinct element we should use $\log\binom{n-1}{d-1}$ bits. This results in the redundancy of at most

$$\log n + \log\binom{k}{d} + \log\binom{n-1}{d-1}.$$

2.2.2 Complementary results

- Redundancy of $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ for $n \ge k^{\frac{\alpha}{1.1}}$

  - [Theorem 20] For $n > k^{\alpha+3}$, and $R \in \{\bar{R}, \hat{R}\}$

$$R(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) = \Theta(k \log k).$$

where the constant for the upper bound is 1 and for the lower bound is 0.5.

- [Theorem 21] For $k^{\frac{\alpha}{1.1}} \le n \le k^{\alpha+3}$

$$\frac{k}{4} \le \bar{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) \le \hat{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) \le k \log k.$$

- [Theorem 8] For $\mathcal{P}_{\le d}$ we present a low-complexity sequential compression algorithm for the range $k > n$ that achieves $\bar{R}(\mathcal{P}_{\le d})$.

2.3 Worst-case redundancy

Let $\hat{p}(x) \overset{\text{def}}{=} \max_{p \in \mathcal{P}} p(x)$ be the maximum probability any distribution in class $\mathcal{P}$ assigns to x. The Shtarkov sum S is

$$S(\mathcal{P}) \overset{\text{def}}{=} \sum_{x \in \mathcal{X}} \hat{p}(x). \tag{2.1}$$

It is well known that Shtarkov sum determines the worst-case redundancy [Sht87]. For any class $\mathcal{P}$

$$\hat{R}(\mathcal{P}) = \log S(\mathcal{P}). \tag{2.2}$$

2.3.1 Small alphabet

Recall that for $k = o(n)$ the leading term in $\hat{R}(\Delta^n_k)$ is $\frac{k-1}{2} \log n$. We now give a simple example showing that un-ordered distribution classes may have much smaller redundancy than Δ^n_k. The Shtarkov sum of a class $\mathcal{P}$ is upper bounded by the number of distributions in the class [OS04]. In particular for a distribution p over k symbols and class $\mathcal{P}_{(p)}$ containing all permutations of p, the Shtarkov sum is upper bounded by $|\mathcal{P}_{(p)}| \le k!$ and therefore $\forall n$,

$$\hat{R}(\mathcal{P}_{(p)}) \le \log k! \le k \log k.$$

Clearly for $n \gg k$, this bound is smaller than $\hat{R}(\Delta_k^n)$.

2.3.2 Large alphabets

As we saw, for any fixed k, knowledge of the underlying-distribution multiset helps in universal compression. It is natural to ask if the same applies for the large-alphabet regime when $n \ll k$. Recall that [AJOS14b, BGG09] showed that for power-law envelopes, $f(i) = c \cdot i^{-\alpha}$, with infinite support size

$$\hat{R}(\mathcal{P}_{\leq f}) = \Theta(n^{\frac{1}{\alpha}}).$$

We show that if the permutation of the distribution is not known then the worst-case redundancy is $\Omega(n) \gg \Theta(n^{\frac{1}{\alpha}})$, and thus the knowledge of the permutation is essential. In particular, we prove that even for the case when the class consists of only one power-law distribution, $\hat{R}$ scales as n.

Theorem 1. *For $k \geq n + C_{k,\alpha} \cdot n^{\alpha}$,*

$$\hat{R}(\mathcal{P}^n_{(\leq c \cdot i^{-\alpha})}) \geq \hat{R}(\mathcal{P}^n_{(zipf(\alpha,k))}) \geq n \log \frac{k-n}{n^{\alpha} C_{k,\alpha}}.$$

Proof. Since $\mathcal{P}_{(\mathrm{zipf}(\alpha,k))} \supseteq \mathcal{P}_{(\leq c \cdot i^{-\alpha})}$, we have

$$\hat{R}(\mathcal{P}^n_{(\leq c \cdot i^{-\alpha})}) \geq \hat{R}(\mathcal{P}^n_{(\mathrm{zipf}(\alpha,k))}).$$

To lower bound $\hat{R}(\mathcal{P}^n_{(\mathrm{zipf}(\alpha,k))})$, Let φ^n_+ be the number of distinct symbols in x^n,

$$\begin{aligned} S(\mathcal{P}^n_{(\mathrm{zipf}(\alpha,k))}) &= \sum_{x^n} \hat{p}(x^n) \\ &\geq \sum_{x^n : \varphi^n_+ = n} \hat{p}(x^n) \\ &\overset{(a)}{\geq} k^{\underline{n}} \prod_{i=1}^{n} \frac{1}{i^{\alpha} C_{k,\alpha}}, \end{aligned}$$

where (a) follows as there are $k^{\underline{n}} = k(k-1)(k-2)\dots(k-n+1)$ sequences with $\varphi_+^n = n$. We lower bound $\hat{p}(x^n)$ for every such sequence. Consider the distribution $q \in \mathcal{P}_{(\text{zipf}(\alpha,k))}$ given by $q(x_i) = \frac{1}{i^\alpha C_{k,\alpha}}$ $\forall 1 \le i \le n$. Clearly $\hat{p}(x^n) \ge q(x^n)$ and thus

$$\ge \left(\frac{k-n}{n^\alpha C_{k,\alpha}}\right)^n.$$

Taking the logarithm yields the result. □

Thus for small values of n, independent of the underlying distribution the per-symbol redundancy is $\log \frac{k}{n^\alpha}$. Since for $n \le k$, $\hat{R}(\Delta_k^n) = n \log \frac{k}{n} - \log e + O(\frac{1}{n})$, we have for $k \ge C_{k,\alpha} \cdot n^\alpha + n$

$$\hat{R}(\mathcal{P}^n_{(ci^{-\alpha},k)}) \le \hat{R}(\Delta_k^n) \le O(n \log \frac{k}{n}).$$

Together with Theorem 1, for $k \ge C_{k,\alpha} \cdot n^\alpha + n$,

$$\Omega(n \log \frac{k}{n^\alpha}) \le \hat{R}(\mathcal{P}^n_{(ci^{-\alpha},k)}) \le O(n \log \frac{k}{n}).$$

2.4 Expected redundancy of $\mathcal{P}_{\le d}$

We start from the most general class defined in Section 2.2 , and upper bound the expected redundancy of $\mathcal{P}_{\le d}$. We also show a matching lower bound up to constants later in this section.

2.4.1 Upper bound

Using an explicit coding scheme, we upper bound the expected redundancy of $\mathcal{P}_{\le d}$ and we use this upper bound for the expected redundancy of the class $\mathcal{P}_{(\text{zipf}(\alpha,k))}$.

Lemma 2. *For all k, n and d,*

$$\bar{R}(\mathcal{P}^n_{\leq d}) \leq d \log \frac{kn}{d^2} + (2\log e + 1)d + \log(n+1).$$

Proof. For a sequence x^n with multiplicity vector of $\mu^k \overset{\text{def}}{=} (\mu_1, \mu_2, \ldots, \mu_k)$, let's assign the probability

$$q(x^n) = \frac{1}{N_{\varphi^n_+}} \cdot \prod_{j=1}^{k} \left(\frac{\mu_j}{n}\right)^{\mu_j},$$

with the normalization factor

$$N_{\varphi^n_+} = n \cdot \binom{k}{\varphi^n_+} \cdot \binom{n-1}{\varphi^n_+ - 1}.$$

Before proceeding, we show that q is a valid coding scheme by showing $\sum_{x^n \in \mathcal{X}^n} q(x^n) \leq 1$,

$$\begin{aligned}
\sum_{x^n \in \mathcal{X}^n} q(x^n) &= \sum_{d'=1}^{n} \sum_{S \subseteq \mathcal{X}:|S|=d'} \sum_{\mu^k:\mu_i=0 \text{ iff } i \notin S} \sum_{x^n:\mu(x^n)=\mu^k} q(x^n) \\
&\overset{(a)}{\leq} \sum_{d'=1}^{n} \sum_{S \subseteq \mathcal{X}:|S|=d'} \sum_{\mu^k:\mu_i=0 \text{ iff } i \notin S} \frac{1}{N_{d'}} \\
&\overset{(b)}{=} \sum_{d'=1}^{n} \frac{\binom{k}{d'} \cdot \binom{n-1}{d'-1}}{N_{d'}} \\
&= \sum_{d'=1}^{n} \frac{1}{n} = 1.
\end{aligned}$$

Where (a) holds since for a given μ^k , the maximum likelihood distribution for all sequences with same values of $\mu_1, \mu_2, \ldots \mu_k$ is same, and (b) follows from the fact that the second summation ranges over $\binom{k}{d'}$ values and the third summation ranges over $\binom{n-1}{d'-1}$ values.

It follows that for any $p \in \mathcal{P}_{\leq d}$,

$$\log \frac{p(x^n)}{q(x^n)} \leq \log N_{\varphi^n_+} + n \cdot \sum_{i=1}^{k} \frac{\mu_i}{n} \log \frac{p_i}{\mu_i/n} \leq \log N_{\varphi^n_+},$$

where the last inequality is by non-negativity of KL divergence. Taking expectations of both sides,

$$
\begin{aligned}
\bar{R}(\mathcal{P}^n_{\le d}) &\le \mathbb{E}[\log N_{\varphi^n_+}] \\
&\le \log n + \mathbb{E}\left[\log \binom{k}{\varphi^n_+} + \log \binom{n-1}{\varphi^n_+ - 1}\right] \\
&\overset{(a)}{\le} \log n + \mathbb{E}\left[\varphi^n_+ \log\left(\frac{k}{\varphi^n_+}\cdot\frac{2n}{\varphi^n_+}\right) + (2\log e)\varphi^n_+\right] \\
&\overset{(b)}{\le} \log n + d\log\frac{kn}{d^2} + (2\log e + 1)d,
\end{aligned}
$$

where (a) follows from $\binom{n}{d} \le \left(\frac{ne}{d}\right)^d$ and (b) follows from Jensen's inequality. □

2.4.2 Lower bound

We present two lower bounds on $\bar{R}(\mathcal{P}_{\le d})$ in Lemmas 3 and 5. The first one is order-wise tight for the range $k \le 10n$ and the second one is order-wise tight for $k > 10n$.

Lemma 3. *For all k, n and d,*

$$\bar{R}(\mathcal{P}^n_{\le d}) \ge \left(\frac{\lfloor d\rfloor - 1}{2}\right)\log\frac{n}{\lfloor d\rfloor}(1+o(1)).$$

Proof. Consider the class $\Delta_{\lfloor d\rfloor}$ of all *i.i.d.* distributions over $\lfloor d\rfloor$ out of k elements. Any distribution in this class will result to expected number of distinct elements $\le d$. Hence, $\Delta^n_{\lfloor d\rfloor} \subseteq \mathcal{P}^n_{\le d}$ and $\bar{R}(\mathcal{P}^n_{\le d}) \ge \bar{R}(\Delta^n_{\lfloor d\rfloor}) = \left(\frac{\lfloor d\rfloor - 1}{2}\right)\log\frac{n}{\lfloor d\rfloor}(1+o(1))$. □

For $k \le 10n$, the leading term in Lemma 2 is of the order of $d\log\frac{n}{d}$ and so is the lower bound in Lemma 3. However, for $k > 10n$, the order-wise leading term in the upper- and lower-bound do not match. In Lemma 5 we show another lower-bound, order-wise tight in the leading term for the range $k > 10n$. Before that, we state a result from the previous works that we later use in the proof of Lemma 5.

Type redundancy: In next lemma we show that the redundancy of the sequence is same as the redundancy of the type vector. Therefore we can focus on compressing the type of the sequence and calculate the expected redundancy of that.

Lemma 4. *(Lemma 9 in [AJOS14b]) Lets define* $\tau(\mathcal{P}^n) = \{\tau(p^n) : p \in \mathcal{P}\}$*, then*

$$\bar{R}(\mathcal{P}^n) = \bar{R}(\tau(\mathcal{P}^n)).$$

Using Lemma 4 we have

$$\begin{aligned}
\bar{R}(\mathcal{P}^n) &= \bar{R}(\tau(\mathcal{P}^n)) \\
&= \min_{q} \max_{p \in \tau(\mathcal{P}^n)} \sum_{\tau^k} p(\tau^k) \log \frac{p(\tau^k)}{q(\tau^k)} \\
&= \min_{q} \max_{p \in \tau(\mathcal{P}^n)} \left[\sum_{\tau^k} p(\tau^k) \log \frac{1}{q(\tau^k)} - H(\tau^k) \right]. \qquad (2.3)
\end{aligned}$$

For any class of distributions with a bound on the expected number of distinct elements, we can lower bound the expected redundancy by Equation (2.3). The first term, $\sum_{\tau^k} p(\tau^k) \log \frac{1}{q(\tau^k)}$ is the number of necessary bits needed to describe the type vector and is related to the expected number of distinct elements in a class. Using the same argument we present the following lower bound on $\mathcal{P}_{\leq d}$.

Lemma 5. *For all* k, n *and any class* $\mathcal{P}_{\leq d}$ *where* $\lfloor d \rfloor > 1$*,*

$$\bar{R}(\mathcal{P}^n_{\leq d}) \geq 0.6 \lfloor d \rfloor \log \frac{k}{\lfloor d \rfloor} - \sqrt{\lfloor d \rfloor} \log \lfloor d \rfloor - \frac{\lfloor d \rfloor}{2} \log \frac{n}{\lfloor d \rfloor} - \frac{\lfloor d \rfloor}{2} \log 2\pi e - O(\frac{\lfloor d \rfloor}{n}).$$

Proof. For the simplicity of notation let $d = \lfloor d \rfloor$ in this proof. Consider the class of all uniform distributions over d elements, denoted by $\mathcal{P}_{unif(d)}$. In this lemma we show a lower bound on $\bar{R}(\mathcal{P}^n_{unif(d)})$ and thus on $\bar{R}(\mathcal{P}^n_{\leq d})$.

First, we show that the expected number of distinct elements for all distributions in $\mathcal{P}_{unif(d)}$

is $d' \overset{\text{def}}{=} d\left(1-(1-\frac{1}{d})^n\right)$,

$$\begin{aligned}\mathbb{E}[\varphi_+^n] &= \sum_{i=1}^{d}\left(1-(1-\frac{1}{d})^n\right)\\ &= d\left(1-(1-\frac{1}{d})^n\right) \overset{\text{def}}{=} d'.\end{aligned}$$

To use (2.3) we upper bound the type entropy for the distributions in $\mathcal{P}_{unif(d)}$:

$$\begin{aligned}H(\tau^k) &\le \sum_{i=1}^{d} H(\tau_i)\\ &\overset{(a)}{=} \frac{d}{2}\log\left(2\pi en\frac{1}{d}(1-\frac{1}{d})\right)+O(\frac{d}{n})\\ &\le \frac{d}{2}\log(2\pi e)+\frac{d}{2}\log(\frac{n}{d})+O(\frac{d}{n}), \qquad (2.4)\end{aligned}$$

where (a) is because each $\tau_i \sim Binomial(n,p_i)$ and thus $H(\tau_i)=\frac{1}{2}\log\left(2\pi enp(1-p)\right)+O(\frac{1}{n})$. Now we have

$$\begin{aligned}\bar{R}(\mathcal{P}^n_{unif(d)}) &\overset{(a)}{=} \bar{R}(\tau(\mathcal{P}^n_{unif(d)}))\\ &\overset{(b)}{\ge} \mathbb{E}\left[\log\binom{k}{\varphi_+^n}\right]-H(\tau^k)\\ &\overset{(c)}{\ge} \mathbb{E}[\varphi_+^n]\cdot\log k-\mathbb{E}[\varphi_+^n\log\varphi_+^n]-H(\tau^k)\\ &\overset{(d)}{=} d'\log k-\mathbb{E}[\left(\varphi_+^n-\mathbb{E}[\varphi_+^n]\right)\log\varphi_+^n]-d'\cdot\mathbb{E}[\log\varphi_+^n]-H(\tau^k)\\ &\overset{(e)}{\ge} d'\log k-\left(\mathbb{E}\left[|\varphi_+^n-\mathbb{E}[\varphi_+^n]|\right]\right)\log d-d'\log\mathbb{E}[\varphi_+^n]-H(\tau^k)\\ &\overset{(f)}{\ge} d'\log\frac{k}{d'}-\sqrt{d}\log d-H(\tau^k)\\ &\overset{(g)}{\ge} 0.6d\log\frac{k}{d}-\sqrt{d}\log d-\frac{d}{2}\log\frac{n}{d}-\frac{d}{2}\log 2\pi e-\Omega(\frac{d}{n}).\end{aligned}$$

Note that (a) is by Lemma 4 and (b) is by (2.3) and the fact that $q(\tau^k)\le\frac{1}{\log\binom{k}{\varphi_+^n}}$ and H is same for all distributions in the class. Also (c) is because $\binom{k}{\varphi_+^n}\ge(\frac{k}{\varphi_+^n})^{\varphi_+^n}$, (d) is by adding and subtracting the term $\mathbb{E}[\varphi_+^n]\mathbb{E}[\log\varphi_+^n]$, (e) is by substituting φ_+^n by a larger value d, using an always-positive term $|\varphi_+^n - \mathbb{E}[\varphi_+^n]|$ instead of $(\varphi_+^n - \mathbb{E}[\varphi_+^n])$, and the concavity of the log function. For (f) we need to first calculate $\mathrm{Var}(\varphi_+^n)$. Let X_j be the event that symbol j appears in x^n,

$$\begin{aligned}
\mathbb{E}[(\varphi_+^n)^2] &= \mathbb{E}\left[\left(\sum_{j=1}^{d} X_j\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{d} X_j^2 + \sum_{j=1,l=1,j\neq l}^{d} X_j X_l\right] \\
&= d\mathbb{E}[X_j^2] + (d^2-d)\mathbb{E}[X_j X_l] \\
&\overset{(1)}{\leq} d\mathbb{E}[X_j] + (d^2-d)\mathbb{E}^2[X_j],
\end{aligned}$$

where (1) is because $\mathbb{E}[X_j^2] = \mathbb{E}[X_j]$ and

$$\begin{aligned}
\mathbb{E}[X_j X_l] &= \Pr[X_j = 1, X_l = 1] \\
&= \Pr[X_j = 1]\Pr[X_l = 1 | X_j = 1] \\
&\leq \Pr[X_j = 1]\Pr[X_l = 1] \\
&= \mathbb{E}[X_j]\mathbb{E}[X_l] = \mathbb{E}^2[X_j].
\end{aligned}$$

Since $\mathbb{E}[\varphi_+^n] = \mathbb{E}\left[\left(\sum_{j=1}^{d} X_j\right)\right] = d\mathbb{E}[X_j]$,

$$\begin{aligned}
\mathrm{Var}(\varphi_+^n) &= \mathbb{E}[(\varphi_+^n)^2] - \mathbb{E}^2[\varphi_+^n] \\
&\leq d\mathbb{E}[X_j] + d^2\mathbb{E}^2[X_j] - d\mathbb{E}^2[X_j] - d^2\mathbb{E}^2[X_j] \\
&= d\left(\mathbb{E}[X_j] - \mathbb{E}^2[X_j]\right) \\
&\leq d,
\end{aligned}$$

where the last line of the equations is by $\mathbb{E}[X_j] \leq 1$. Also by concavity,

$$\mathbb{E}\left[|\varphi_+^n - \mathbb{E}[\varphi_+^n]|\right] \leq \sqrt{\mathbb{E}\left[|\varphi_+^n - \mathbb{E}[\varphi_+^n]|^2\right]} = \sqrt{\mathrm{Var}(\varphi_+^n)} \leq \sqrt{d}.$$

Finally (g) is by (2.4) and $0.6d \leq d' \leq d$. Since $\mathcal{P}_{unif(d)} \subseteq \mathcal{P}_{\leq d}$, we have the lemma. □

Lemma 6. *For all k,n and any class $\mathcal{P}_{\leq d}$,*

$$\bar{R}(\mathcal{P}_{\leq d}^n) \geq \Omega\left(d \log \frac{\max\{k,n\}}{d}\right).$$

Proof. By Lemmas 3 and 5, we have

$$\bar{R}(\mathcal{P}_{\leq d}^n) \geq \max\left\{0.6\lfloor d\rfloor \log \frac{k}{\lfloor d\rfloor} - \sqrt{\lfloor d\rfloor}\log\lfloor d\rfloor - \frac{\lfloor d\rfloor}{2}\log\frac{n}{\lfloor d\rfloor} - \frac{\lfloor d\rfloor}{2}\log 2\pi e - \Omega(\frac{\lfloor d\rfloor}{n}),\right.$$
$$\left.\left(\frac{\lfloor d\rfloor - 1}{2}\right)\log\frac{n}{\lfloor d\rfloor}(1+o(1))\right\}.$$

For $k \geq 10n$, the first term in the maximum is $\geq \Omega(d\log\frac{k}{d})$. For $n \leq k \leq 10n$, the second term is $\geq \Omega(d\log\frac{k}{d})$ and for $k \leq n$, the second term is $\geq \Omega(d\log\frac{n}{d})$, hence the lemma. □

Theorem 7. *For all k,n and any class $\mathcal{P}_{\leq d}$,*

$$\bar{R}(\mathcal{P}_{\leq d}^n) = \Theta\left(d \log \frac{\max\{k,n\}}{d}\right).$$

Proof. Proof follows from Lemmas 2 and 6. □

2.5 Sequential compression

While the compression scheme considered so far was to encode the whole block, in many applications the symbols must be encoded as they arrive, raising the need for a low-complexity sequential encoder. In *sequential* compression we associate with every sequence $x^n \in \mathcal{X}^n$ a probability distribution $q(x|x^n)$ over $[k]$ corresponding to the probability of $x_{n+1} = x$ after observing x^n. Therefore, the probability that a sequential encoder q assigns to x^n can be written as

$$q(x^n) = \prod_{i=1}^{n-1} q(x_i|x^{i-1}).$$

In this section we first introduce *absolute-discount* estimator and then show that it has diminishing per-symbol expected redundancy for $\mathcal{P}_{\le d}$ when $k > n$. In fact we show that we can achieve the order-wise optimal redundancy for the range $k > n$ in a sequential manner.

2.5.1 Absolute-discount estimator

Similar to add-constant estimators, absolute discount modifies the empirical estimator to assign probability to symbols. It does so by removing some probability from the observed symbols and assigning it to the unobserved ones. Let μ_j be the multiplicity of symbol $j \in [k]$ and φ_+^i be the number of distinct elements in x^i. For a fixed discount value δ, the estimator is formally defined as

$$q^{AD}(x_{i+1} = j|x^i) = \begin{cases} \frac{\mu_j - \delta}{i} & \text{if } \mu_j > 0 \\ \frac{\varphi_+^i \cdot \delta}{(k-\varphi_+^i)i} & \text{if } \mu_j = 0 \end{cases}$$

Next we show the performance of the absolute-discount estimator for $\mathcal{P}_{\le d}$. Let

$$\bar{R}(\mathcal{P}, q) \stackrel{\text{def}}{=} \max_{p \in \mathcal{P}} \mathbb{E}_{x^n}[\log \frac{p(x^n)}{q(x^n)}]$$

be the expected redundancy for class $\mathcal{P}$ using the estimator q. Note that for absolute-discount estimator,

$$\begin{aligned} q^{AD}(x^n) = \prod_{i=0}^{n-1} q^{AD}(x_{i+1}|x^i) &= \frac{1}{n!}\frac{1}{k}\frac{\delta}{k-1}\frac{2\delta}{k-2}\cdots\frac{(\varphi_+^i-1)\delta}{k-\varphi_+^i+1}\prod_{j,\mu_j\geq 2}(1-\delta)(2-\delta)\ldots(\mu_j-1-\delta) \\ &= \frac{\delta^{\varphi_+^i-1}(\varphi_+^i-1)!}{k!/(k-\varphi_+^i)!}\frac{1}{n!}\prod_{j,\mu_j\geq 2}(1-\delta)(2-\delta)\ldots(\mu_j-1-\delta). \end{aligned}$$

Theorem 8. *For all $k > n$,*

$$\bar{R}(\mathcal{P}^n_{\leq d}, q^{AD}) \leq d\log\frac{k}{d} + 2d\log n + O(d).$$

Proof.
$$\mathbb{E}\left[\log\frac{p(x^n)}{q^{AD}(x^n)}\right] = \mathbb{E}\left[\log\frac{k!/(k-\varphi_+^n)!}{\delta^{\varphi_+^n-1}(\varphi_+^n-1)!}\right] + \mathbb{E}\left[\log\frac{p(x^n)}{\frac{1}{n!}\prod_{j,\mu_j\geq 2}(1-\delta)(2-\delta)\ldots(\mu_j-1-\delta)}\right]$$

For the first term in the right hand side above,

$$\begin{aligned} \mathbb{E}\left[\log\frac{k!/(k-\varphi_+^n)!}{\delta^{\varphi_+^n-1}(\varphi_+^n-1)!}\right] &= \mathbb{E}\left[\log\binom{k}{\varphi_+^n}\varphi_+^n(\frac{1}{\delta})^{\varphi_+^n-1}\right] \\ &= \mathbb{E}\left[\log\binom{k}{\varphi_+^n}\right] + \mathbb{E}\left[\log\varphi_+^n\right] + \mathbb{E}\left[(\varphi_+^n-1)\log\frac{1}{\delta}\right] \\ &\overset{(a)}{\leq} d\log\frac{k}{d} + \log d + (d-1)\log\frac{1}{\delta}, \end{aligned}$$

where (a) is because of concavity of $x\log x$ and $\log x$. For the second term we have

$$\begin{aligned}
\mathbb{E}\left[\log\frac{p(x^n)}{\frac{1}{n!}\prod_{j,\mu_j\geq 2}(1-\delta)(2-\delta)...(\mu_j-1-\delta)}\right] &= \mathbb{E}\left[\log\frac{\prod_j p_j^{\mu_j}}{\frac{1}{n!}\prod_{j,\mu_j\geq 2}\frac{\Gamma(\mu_j-\delta)}{\Gamma(1-\delta)}}\right]\\
&\overset{(b)}{\leq} \mathbb{E}\left[\log\frac{(\Gamma(1-\delta))^{\varphi_+^n}\prod_j p_j^{\mu_j}}{\frac{1}{n!}\prod_{j,\mu_j\geq 2}(\mu_j-2)!}\right]\\
&\overset{(c)}{\leq} \mathbb{E}\left[\log\frac{(\Gamma(1-\delta))^{\varphi_+^n}\prod_j p_j^{\mu_j}}{\frac{e^{n-1}}{n^{n+\frac{1}{2}}}\prod_{j,\mu_j\geq 2}(\frac{\mu_j-2}{e})^{\mu_j-2}}\right]\\
&\overset{(d)}{\leq} \mathbb{E}\left[\log\frac{(\Gamma(1-\delta))^{\varphi_+^n}\prod_{j,\mu_j\geq 2} p_j^{\mu_j-2}}{\frac{e^{\varphi_+^n-1}}{n^{2\varphi_+^n+\frac{1}{2}}}\prod_{j,\mu_j\geq 2}(\frac{\mu_j-2}{n})^{\mu_j-2}}\right]\\
&\overset{(e)}{=} \mathbb{E}\left[\log\prod_{j,\mu_j\geq 2}\left(\frac{p_j}{\frac{\mu_j-2}{n}}\right)^{\mu_j-2}\right]\\
&+\mathbb{E}\left[\varphi_+^n\log\left(\Gamma(1-\delta)\right)+(2\varphi_+^n+\frac{1}{2})\log n\right]\\
&\overset{(f)}{\leq} 2d+d\log\left(\Gamma(1-\delta)\right)+(2d+\frac{1}{2})\log n.
\end{aligned}$$

where (b) is by $\Gamma(\mu_j-\delta)\geq\Gamma(\mu_j-1)=(\mu_j-2)!$, $|j:\mu_j\geq 2|\leq\varphi_+^n$, and $\Gamma(1-\delta)>1$. Also, (c) is by $\sqrt{2\pi}m^{m+\frac{1}{2}}e^{-m}\leq m!\leq em^{m+\frac{1}{2}}e^{-m}$, and (e) is by Lemma 9 □

Lemma 9. $\log\left(\prod_{j,\mu_j\geq 2}\left(\frac{p_j}{\frac{\mu_j-2}{n}}\right)^{\mu_j-2}\right)\leq 2\varphi_+^n.$

Proof.

$$\log\prod_{j,\mu_j\geq 2}\left(\frac{p_j}{\frac{\mu_j-2}{n}}\right)^{\mu_j-2}=nS\sum_{j,\mu_j\geq 2}\frac{\mu_j-2}{nS}\log\frac{p_j}{\frac{\mu_j-2}{nS}}+nS\log\frac{1}{S},$$

where $S=\sum_{j,\mu_j\geq 2}\frac{\mu_j-2}{n}=1-\frac{2\varphi_+^n-\Phi_1}{n}$. Due to concavity of log function the first term right hand side above is negative and therefore

$$\log\prod_{j,\mu_j\geq 2}\left(\frac{p_j}{\frac{\mu_j-2}{n}}\right)^{\mu_j-2}\leq n(1-\frac{2\varphi_+^n-\Phi_1}{n})\log\frac{1}{1-\frac{2\varphi_+^n-\Phi_1}{n}}\leq n\frac{2\varphi_+^n-\Phi_1}{n}\leq 2\varphi_+^n,$$

and the Lemma follows. □

2.6 Expected redundancy of $\mathcal{P}_{(\mathbf{zipf}(\alpha,k))}$

In this section we derive lower bounds on the expected redundancy of $\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}$ for different ranges of k and n and also use the results in Section 2.4 to derive upper bounds on the expected redundancy.

To show the lower bounds we use a technique introduced in previous works. We reintroduce Poisson sampling and relate the expected redundancy in two cases when we use normal sampling and Poisson sampling.

Poisson sampling: In the standard sampling method, where a distribution is sampled n times, the multiplicities are dependent, for example they add up to n. Hence, calculating redundancy under this sampling often requires various concentration inequalities, complicating the proofs. A useful approach to make them independent and hence simplify the analysis is to sample the distribution n' times, where n' is a Poisson random variable with mean n. Often called as Poisson sampling, this approach has been used in universal compression to simplify the analysis [ADO12, AJOS14b, YB13, ADJ$^+$13].

Under Poisson sampling, if a distribution p is sampled *i.i.d.* $\mathrm{poi}(n)$ times, then the number of times symbol x appears is an independent Poisson random variable with mean np_x, namely, $\Pr(\mu_x = \mu) = \frac{e^{-np_x}(np_x)^\mu}{\mu!}$ [MU05]. Henceforth, to distinguish between two cases of normal sampling and Poisson sampling we specify it with superscripts n and $poi(n)$, respectively.

Lemma 10 lower bounds $\bar{R}(\mathcal{P}^n)$ by the redundancy in the presence of Poisson sampling. The proof is given in Appendix 2.A. We later use this lemma in our lower-bound arguments.

Lemma 10. *For any class $\mathcal{P}$,*

$$\bar{R}(\mathcal{P}^n) \geq \frac{1}{2}\bar{R}(\mathcal{P}^{\mathrm{poi}(n)}).$$

Now we lower bound the expected redundancy of class $\mathcal{P}^n_{(\text{zipf}(\alpha,k))}$. By Lemmas 4 and 10 we have

$$\bar{R}\left(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}\right) \geq \frac{1}{2}\bar{R}\left(\mathcal{P}^{poi(n)}_{(\text{zipf}(\alpha,k))}\right) = \frac{1}{2}\bar{R}\left(\tau(\mathcal{P}^{poi(n)}_{(\text{zipf}(\alpha,k))})\right)$$

For notational simplicity we denote $\tau(\mathcal{P}^{poi(n)}_{(\text{zipf}(\alpha,k))})$ by $\mathcal{P}_\tau$. Therefore it suffices to show a lower bound on $\bar{R}(\mathcal{P}_\tau)$. Similar to the decomposition in (2.3),

$$\begin{aligned}\bar{R}(\mathcal{P}_\tau) &= \min_q \max_{p\in\mathcal{P}_\tau} \sum_{\tau^k} p(\tau^k)\log\frac{p(\tau^k)}{q(\tau^k)} \\ &= \min_q \max_{p\in\mathcal{P}_\tau}\left[\sum_{\tau^k} p(\tau^k)\log\frac{1}{q(\tau^k)} - \sum_{\tau^k} p(\tau^k)\log\frac{1}{p(\tau^k)}\right]. \qquad (2.5)\end{aligned}$$

Therefore, the goal is to lower bound $\sum_{\tau^k} p(\tau^k)\log\frac{1}{q(\tau^k)}$ and upper bound $\sum_{\tau^k} p(\tau^k)\log\frac{1}{p(\tau^k)}$. For the first term, we upper bound $q(\tau^k)$ based on the number of distinct elements in the sequence $x^{poi(n)}$ in Lemmas 11, 12, and 13. Afterwards we consider the second term which is the entropy of the type vectors under Poisson sampling and we upper bound it in Lemma 15.

The following two concentration lemmas from [GHP$^+$07, BHBO14] help us to relate the expected number of distinct elements for class $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ in both normal and Poisson sampling. Lets denote the number of distinct elements in $x^{poi(n)}$ as $\varphi_+^{poi(n)}$, and $d^{poi(n)} = \mathbb{E}[\varphi_+^{poi(n)}]$. Similarly, φ_+^n is the number of distinct elements in x^n and $d = \mathbb{E}[\varphi_+^n]$.

Lemma 11. *([BHBO14]) Let $v = \mathbb{E}[\varphi_1^{poi(n)}]$ be the expected number of elements which appeared once in $x^{poi(n)}$, then*

$$\Pr[\varphi_+^{poi(n)} < d^{poi(n)} - \sqrt{2vs}] \leq e^{-s}.$$

Lemma 12. *(Lemma 1 in [GHP$^+$07]) Let $\mathbb{E}[\varphi_2^{poi(n)}]$ be the expected number of elements which appeared twice in $x^{poi(n)}$, then*

$$|d^{poi(n)} - d| < 2\frac{\mathbb{E}[\varphi_2^{poi(n)}]}{n}.$$

Using Lemmas 11 and 12 we bound the number of non-zero elements in $\tau(x^{poi(n)})$.

Lemma 13. *The number of non-zero elements in* $\tau(x^{poi(n)})$ *is more than* $(1-\varepsilon)d$ *with probability* $> 1-e^{-\frac{d(\varepsilon-2/n)^2}{2}}$. *Also, the number of non-zero elements in* $\tau(x^{poi(n)})$ *is less than* $(1+\varepsilon)d$ *with probability* $> 1-e^{-\frac{d(\varepsilon-2/n)^2}{2}}$.

Proof. The number of non-zero elements in $\tau(x^{poi(n)})$ is equal to the number of distinct elements in $x^{poi(n)}$. Let $v = \mathbb{E}[\varphi_1^{poi(n)}]$, by Lemma 11

$$\begin{aligned}\Pr[\varphi_+^{poi(n)} < d^{poi(n)}(1-\varepsilon)] &\leq e^{-\frac{(d^{poi(n)}\varepsilon)^2}{2v}} \\ &\overset{(a)}{\leq} e^{-\frac{d^{poi(n)}\varepsilon^2}{2}},\end{aligned}$$

where (a) is because $d^{poi(n)} > v$. Lemma 12 implies $d^{poi(n)}(1-\frac{2}{n}) < d < d^{poi(n)}(1+\frac{2}{n})$. Therefore,

$$\begin{aligned}\Pr[\varphi_+^{poi(n)} < d(1-\varepsilon)] &\leq \Pr[\varphi_+^{poi(n)} < d^{poi(n)}\left(1+\frac{2}{n}\right)(1-\varepsilon)] \\ &\leq e^{-\frac{d(\varepsilon-\frac{2}{n})^2}{2}}.\end{aligned}$$

Where the last inequality is by Lemma 13. Proof of the other part is similar and omitted. □

Next we lower bound the number of bits we need to express τ^k based on the number of non-zero elements in it.

Lemma 14. *If the number of non-zero elements in* τ^k *is more than* z, *then*

$$q(\tau^k) \leq \frac{1}{\binom{k}{z}}.$$

Proof. Consider all the type vectors with the same number of non-zero elements as τ^k. It is not hard to see that q should assign same probability to all types with the same profile vector. Number of such type vectors for a given number of non-zero elements z is at least $\binom{k}{z}$. □

Note that the number of non-zero elements in τ^k is same as $\varphi_+^{poi(n)}$. Based on Lemmas 13 and 14 we have

$$\begin{aligned}\sum_{\tau^k} p(\tau^k)\log\frac{1}{q(\tau^k)} &\geq \sum_{\tau^k:\varphi_+^{poi(n)}\geq(1-\varepsilon)d} p(\tau^k)\log\frac{1}{q(\tau^k)}\\ &\geq \sum_{\tau^k:\varphi_+^{poi(n)}\geq(1-\varepsilon)d} p(\tau^k)\log\binom{k}{d(1-\varepsilon)}\\ &\geq \left(1-e^{-\frac{d(\varepsilon-\frac{2}{n})^2}{2}}\right)\log\binom{k}{d(1-\varepsilon)}\\ &= (1+o_d(1))\log\binom{k}{d}. \end{aligned}\tag{2.6}$$

where the last line is by choosing $\varepsilon = d^{-\frac{1}{3}}$. Here we state a lemma to upper bound the entropy of type vectors for the distributions in class $\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}$. The proof is in Appendix 2.B.

Lemma 15. *Let τ^k be the induced type vector by any distribution in $\mathcal{P}_{(zipf(\alpha,k))}$, then for $n < k^{\frac{\alpha}{1.1}}$,*

$$H(\tau^k) \leq c_1 \cdot n^{\frac{1}{\alpha}} + c_2 \cdot \log n,$$

for some constants c_1, c_2 that can be found in Appendix 2.B.

Lemma 16. *For $n < k^{\frac{\alpha}{1.1}}$ and $d = \mathbb{E}[\varphi_+^n]$ there exist constants c_1, and c_2 such that*

$$\bar{R}\left(\mathcal{P}^n_{(zipf(\alpha,k))}\right) \geq \left(\frac{1}{2}+o_d(1)\right)\cdot\log\binom{k}{d} - c_1\cdot n^{\frac{1}{\alpha}} - c_2\cdot\log n.$$

Proof. Proof is by Equations (2.5) and (2.6) and Lemma 15. □

Theorem 17 shows that in fact for $n < k^{\frac{\alpha}{1.1}}$ the upper and lower bounds match up to constant factors.

Theorem 17. *For $n < k^{\frac{\alpha}{1.1}}$*

$$\bar{R}\left(\mathcal{P}^n_{(zipf(\alpha,k))}\right) = \Theta\left(n^{\frac{1}{\alpha}} \log \frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}}\right).$$

Proof. By Lemmas 16 and 28 in Appendix 2.C we have $\bar{R}\left(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}\right) \geq \Omega\left(n^{\frac{1}{\alpha}} \log \frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}}\right)$. Since $\mathcal{P}^n_{(\text{zipf}(\alpha,k))} \subseteq \mathcal{P}^n_{\leq d}$, using Lemma 2 results to $\bar{R}\left(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}\right) \leq O\left(n^{\frac{1}{\alpha}} \log \frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}}\right)$, and hence the Theorem. □

Here we show matching upper and lower bounds on the expected redundancy of $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ for $n > k^{\alpha+2} \log k$. Before stating the theorem, we define the concept of distinguishablity and state a lemma, both used in the theorem's proof.

Lemma 18. *Let $X_1 \sim \text{poi}(\lambda_1)$, $X_2 \sim \text{poi}(\lambda_2)$ with $\lambda_1 < \lambda_2$ and $Y = X_1 - X_2$. Then,*

$$\Pr(Y \geq 0) \leq \exp\left(-(\sqrt{\lambda_1} - \sqrt{\lambda_2})^2\right).$$

Proof. The proof follows from a standard Chernoff bound. □

Definition 19. *Let $\mathcal{P}$ be a class of distributions over the alphabet $\mathcal{X}$. A subclass $\mathcal{S} \subseteq \mathcal{P}$ is $\varepsilon-distinguishable$ if there is a mapping $f : \mathcal{X}^n \to \mathcal{S}$ such that for all sequences $x^n \in \mathcal{X}^n$ generated by $S \in \mathcal{S}$, $\Pr\{f(x^n) \neq S\} \leq \varepsilon$.*

[ADO12] showed that $\bar{R}(\mathcal{P}) \geq \max_{\mathcal{S}}(1-\varepsilon)\log|\mathcal{S}| - h(\varepsilon)$, where $h(\varepsilon)$ is the binary entropy function. Now we formally state the Theorem.

Theorem 20. *For $n \geq \frac{9C^2_{k,\alpha}}{\alpha^2} k^2 (k+1)^\alpha \log \frac{k}{\varepsilon}$,*

$$(1-\varepsilon)\log(k!) \leq \bar{R}\left(\mathcal{P}^n_{(zipf(\alpha,k))}\right) \leq \log(k!).$$

Proof. Using the concept of distinguishablity in Definition 19, we show that if $n \geq \frac{9C^2_{k,\alpha}}{\alpha^2} k^2 (k+1)^\alpha \log \frac{k}{\varepsilon}$ we can pack all distributions of $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ in a set $\mathcal{S}$ such that it is $\varepsilon-$distinguishable.

In words, the probability that a sequence generated by a distribution in the class $\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}$ misidentified to be generated by another distribution in the class is less than ε. The mapping f we use to map the sequences to probability distributions (a permutation of a power-law) is a simple one. We just sort the elements based on their multiplicities and then choose the permutation corresponding to that (ties are broken arbitrarily). For notational simplicity we denote $C_{k,\alpha}^{-1}$ by c in this proof.

$$\begin{aligned}
\Pr\{f(x^n) \neq S\} &\overset{(a)}{\leq} \bigcup_{i=1}^{k} \Pr\{\mu_i \geq \mu_{i+1}, p_i < p_{i+1}\} \\
&\overset{(b)}{\leq} \sum_{i=1}^{k} \exp\left(-(\sqrt{np_i} - \sqrt{np_{i+1}})^2\right) \\
&\leq \sum_{i=1}^{k} \exp\left(-nc\left(\frac{1}{i^{\alpha/2}} - \frac{1}{(i+1)^{\alpha/2}}\right)^2\right) \\
&\overset{(c)}{\leq} \sum_{i=1}^{k} \exp\left(-nc\left(\frac{\alpha/3 \cdot i^{\alpha/2-1}}{i^{\alpha/2}(i+1)^{\alpha/2}}\right)^2\right) \\
&= \sum_{i=1}^{k} \exp\left(-\frac{nc\alpha^2}{9i^2(i+1)^{\alpha}}\right) \\
&\leq k \cdot \exp\left(-\frac{nc\alpha^2}{9k^2(k+1)^{\alpha}}\right)
\end{aligned}$$

where (b) is because of lemma 18, and (c) is because $(1+x)^{\frac{\alpha}{2}} - 1 \geq \frac{\alpha x}{3}$ for $0 < x \leq 1$ and $\alpha \geq 1$. To have $\Pr\{f(x) \neq \mathcal{S}\} \leq \varepsilon$, it is sufficient to have $n \geq \frac{9}{c^2\alpha^2} k^2 (k+1)^{\alpha} \log\frac{k}{\varepsilon}$. Thus, for this range we have $\bar{R}(\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}) \geq (1-\varepsilon)\log|\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}| = (1-\varepsilon)\log(k!)$. Also, $\bar{R}(\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}) \leq \hat{R}(\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}) \leq \log k!$ and hence the theorem. □

The next theorem bounds the expected redundancy for $n \geq C_{k,\alpha}(20k)^{\alpha}$.

Theorem 21. *For $n \geq C_{k,\alpha}(20k)^{\alpha}$,*

$$\left(1 - e^{-nD(\frac{1}{16}||\frac{1}{20})}\right) \cdot D\left(\frac{1}{8}||\frac{1}{2}\right) \frac{\log e}{2} \cdot k \leq \bar{R}(\mathcal{P}_{(zipf(\alpha,k))}) \leq \log(k!)$$

Proof. The upper bound is trivial. To show the lower bounds, similar to Theorem 20, we use distinguishability . Specifically, we construct a subclass of distributions $\mathcal{S}$, such that there exists a mapping $f : \mathcal{X}^n \to \mathcal{S}$ under which $\Pr\{f(x^n) \neq S\} \leq \varepsilon$. Then we lower bound the expected redundancy by $(1-\varepsilon)\log|S|$.

Consider the distribution $P_0 = \left(p_{(1)}, p_{(1+\frac{k}{2})}, p_{(2)}, p_{(2+\frac{k}{2})}, \dots, p_{(\frac{k}{2})}, p_{(k)}\right)$. Recall that $p_{(i)}$ denotes the i'th largest probability. Now we construct class $\mathcal{S}$ with perturbed versions of P_0. The only perturbation allowed is to change the place of two elements in any non-overlapping pairs of consecutive elements. Namely, the first and second elements can be substituted, same for the third and fourth elements and so on. If we allow this change for all $\frac{k}{2}$ non-overlapping consecutive pairs, there will be $2^{\frac{k}{2}}$ different distributions in the class. By allowing only a certain number $h < \frac{k}{2}$ of changes we maintain the property of distinguishability.

Consider any distribution in class $\mathcal{S}$ as a binary codeword of length $\frac{k}{2}$, where 1 at position i means the re-ordering of $p_{(i)}$ and $p_{(i+\frac{k}{2})}$ with respect to P_0 which has the all-zero codeword. By ensuring the minimum Hamming distance of h, we can correct up to $\frac{h-1}{2}$ errors and therefore uniquely determine which probability distribution in $\mathcal{S}$ generates x^n. Using Gilbert-Varshamov bound, we have $|\mathcal{S}| \geq \frac{2^{\frac{k}{2}}}{\sum_{j=0}^{h-1}\binom{\frac{k}{2}}{j}}$. Choosing $h = \frac{k}{16}$, by Chernoff bound we have $|\mathcal{S}| \geq e^{\frac{k}{2}D(\frac{1}{8}||\frac{1}{2})}$, where D is KL-divergence.

Next we show that the set $\mathcal{X}$ is $\varepsilon-$distinguishable. We show the probability of having more than $\frac{h-1}{2}$ changes with respect to P_0 can be made arbitrarily small. Let P_{error} be the probability of re-ordering of $p_{(i)}$ and $p_{(i+\frac{k}{2})}$ with respect to P_0. Let c be $C_{k,\alpha}^{-1}$ for notational simplicity. Based on lemma 18,

$$
\begin{aligned}
P_{error} &\leq \exp\left(-nc\left(\frac{1}{i^{\frac{\alpha}{2}}}-\frac{1}{(i+\frac{k}{2})^{\frac{\alpha}{2}}}\right)^2\right)\\
&\leq \exp\left(-nc\left(\frac{1}{\frac{k}{2}^{\frac{\alpha}{2}}}-\frac{1}{k^{\frac{\alpha}{2}}}\right)^2\right)\\
&\leq \exp\left(-nc\frac{(2^{\frac{\alpha}{2}}-1)^2}{k^{\alpha}}\right)
\end{aligned}
$$

Thus if $k < \frac{1}{20}(cn)^{\frac{1}{\alpha}}$ then $P_{error} < \frac{1}{20}$, and $\Pr\{f(x^n) \neq S\} = \Pr(\text{having more than } \frac{k}{32} \text{ re-orders}) < e^{-nD(\frac{1}{16}||\frac{1}{20})}$ by a simple application of Chernoff bound. Together with the definition of the distinguishability we have $\bar{R}(\mathcal{P}_{(\text{zipf}(\alpha,k))}) \geq \left(1-e^{-nD(\frac{1}{16}||\frac{1}{20})}\right)\cdot D(\frac{1}{8}||\frac{1}{2})\frac{\log e}{2}\cdot k.$ □

2.7 Expected redundancy of power-law envelope classes

In this section we study the expected redundancy of the classes $\mathcal{P}_{(\leq c\cdot i^{-\alpha})}$ and $\mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$.

Theorem 22. *For $n < k^{\frac{\alpha}{1.1}}$,*

$$\bar{R}(\mathcal{P}_{(\leq c\cdot i^{-\alpha})}) = \Theta\left(n^{\frac{1}{\alpha}}\log\frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}}\right).$$

Proof. Since $\mathcal{P}_{(\text{zipf}(\alpha,k))} \subseteq \mathcal{P}_{(\leq c\cdot i^{-\alpha})}$, we have $\bar{R}(\mathcal{P}_{(\leq c\cdot i^{-\alpha})}) \geq \bar{R}(\mathcal{P}_{(\text{zipf}(\alpha,k))})$. Theorem 17 results to $\bar{R}(\mathcal{P}_{(\text{zipf}(\alpha,k))}) > \Omega(n^{\frac{1}{\alpha}}\log\frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}})$. Moreover, $\mathcal{P}_{(\leq c\cdot i^{-\alpha})} \subseteq \mathcal{P}_{\leq d}$ for suitable d chosen based on Lemma 27 in Appendix 2.C. Therefore, we have $\bar{R}(\mathcal{P}_{(\leq c\cdot i^{-\alpha})}) \leq \bar{R}(\mathcal{P}_{\leq d}) \leq O(n^{\frac{1}{\alpha}}\log\frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}})$ and the theorem. □

Similarly for the class $\mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$ we have the following theorem.

Theorem 23. *For $n < k^{\frac{\alpha}{1.1}}$,*

$$\bar{R}(\mathcal{P}_{(zipf(\geq\alpha,k))}) = \Theta\left(n^{\frac{1}{\alpha}} \log \frac{\max\{k,n\}}{n^{\frac{1}{\alpha}}}\right).$$

Proof. The proof is similar to the one of Theorem 22 and results from the fact that $\mathcal{P}_{(\text{zipf}(\alpha,k))} \subseteq \mathcal{P}_{(\text{zipf}(\geq\alpha,k))}$ and $\mathcal{P}_{(\text{zipf}(\geq\alpha,k))} \subseteq \mathcal{P}_{\leq d}$ with d chosen as the expected number of distinct elements for $\mathcal{P}_{(\text{zipf}(\alpha,k))}$ based on Lemma 28 in Appendix 2.C. □

2.8 Acknowledgments

Chapter 2, in part, is a reprint of the material as it appears in *2015 IEEE International Symposium on Information Theory (ISIT)*. Falahatgar, Moein, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh, 2015. The dissertation author was the primary investigator and author of this paper. This chapter has also been submitted with the same authors for publication of the material as it may appear in the journal of IEEE Transactions on Information Theory.

2.A Proof of Lemma 10

[Lemma 10]For any class $\mathcal{P}$,

$$\bar{R}(\mathcal{P}^n) \geq \frac{1}{2}\bar{R}(\mathcal{P}^{\text{poi}(n)}).$$

Proof. By definition,

$$\bar{R}(\mathcal{P}^{\text{poi}(n)}) = \min_q \max_{p\in\mathcal{P}} \mathbb{E}_{\text{poi}(n)}\left[\log \frac{p_{\text{poi}(n)}(x^{n'})}{q(x^{n'})}\right], \tag{2.7}$$

where subscript poi(n) indicates that the probabilities are calculated under Poisson sampling. Similarly, for every n',

$$\bar{R}(\mathcal{P}^{n'}) = \min_q \max_{p\in\mathcal{P}} \mathbb{E}\left[\log \frac{p(x^{n'})}{q(x^{n'})}\right].$$

Let $q_{n'}$ denote the distribution that achieves the above minimum. We upper bound the right hand side of Equation (2.7) by constructing an explicit q. Let

$$q(x^{n'}) = e^{-n}\frac{n^{n'}}{n'!}q_{n'}(x^{n'}).$$

Clearly q is a distribution as it adds up to 1. Furthermore, since $p_{\text{poi}(n)}(x^{n'}) = e^{-n}\frac{n^{n'}}{n'!}p(x^{n'})$, we get

$$\begin{aligned}\bar{R}(\mathcal{P}^{\text{poi}(n)}) &\le \max_{p\in\mathcal{P}} \mathbb{E}_{\text{poi}(n)}\left[\log \frac{p_{\text{poi}(n)}(x^{n'})}{q(x^{n'})}\right]\\ &= \max_{p\in\mathcal{P}} \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\mathbb{E}\left[\log \frac{e^{-n}\frac{n^{n'}}{n'!}p(x^{n'})}{e^{-n}\frac{n^{n'}}{n'!}q_{n'}(x^{n'})}\right]\\ &\le \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\max_{p\in\mathcal{P}}\mathbb{E}\left[\log \frac{p(x^{n'})}{q_{n'}(x^{n'})}\right]\\ &= \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\bar{R}(\mathcal{P}^{n'}),\end{aligned}$$

where the last equality follows from definition of $q_{n'}$. By monotonicity and sub-additivity of $\bar{R}(\mathcal{P}^{n'})$ (see Lemma 5 in [ADO12]), it follows that

$$\begin{aligned}\bar{R}(\mathcal{P}^{n'}) &\le \bar{R}(\mathcal{P}^{n\lceil \frac{n'}{n}\rceil})\\ &\le \left\lceil \frac{n'}{n}\right\rceil \bar{R}(\mathcal{P}^{n})\\ &\le \left(\frac{n'}{n}+1\right)\bar{R}(\mathcal{P}^{n}).\end{aligned}$$

Substituting the above bound we get

$$\begin{aligned}\bar{R}(\mathcal{P}^{\mathrm{poi}(n)}) &\le \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\left(\frac{n'}{n}+1\right)\bar{R}(\mathcal{P}^{n})\\ &= 2\bar{R}(\mathcal{P}^{n}),\end{aligned}$$

where the last equality follows from the fact that expectation of n' is n. □

2.B Proof of Lemma 15

Recall that if distribution p is sampled *i.i.d.* $poi(n)$ times, then the number of times symbol i appears, μ_i is an independent Poisson random variable with mean $\lambda_i \overset{\text{def}}{=} np_i$. First we state a lemma, bounding the entropy of a Poisson random variable.

Lemma 24. *If* $X \sim poi(\lambda)$ *for* $\lambda < 1$*, then*

$$H(X) \le \lambda(1-\log\lambda) + e^{-\lambda}\frac{\lambda^2}{1-\lambda}.$$

Proof.

$$
\begin{aligned}
H(X) &= -\sum_{i=0}^{\infty} p_i \log p_i \\
&= -\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\log\frac{e^{-\lambda}\lambda^i}{i!} \\
&= -\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\left(\log e^{-\lambda} + i\log\lambda - \log(i!)\right) \\
&= \lambda\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} - \log\lambda\sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} + \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\log(i!) \\
&\overset{(a)}{=} \lambda - \lambda\log\lambda + e^{-\lambda}\sum_{i=2}^{\infty}\frac{\lambda^i\log(i!)}{i!} \\
&\leq \lambda - \lambda\log\lambda + e^{-\lambda}\sum_{i=0}^{\infty}\lambda^i \\
&\overset{(b)}{=} \lambda(1-\log\lambda) + e^{-\lambda}\frac{\lambda^2}{1-\lambda}
\end{aligned}
$$

where (a) is because the first two terms in the last summation is zero and for the rest of the terms, $\log(i!) < i!$. Also (b) follows from the geometric sum for $\lambda < 1$. □

To bound the type entropy we can write

$$
\begin{aligned}
H(\tau^k) &= \sum_{i=1}^{k} H(\mu_i) \\
&= \sum_{i=1}^{k} H(poi(\lambda_i)) \\
&= \sum_{\lambda_i<0.7} H(poi(\lambda_i)) + \sum_{\lambda_i\geq 0.7} H(poi(\lambda_i)) \\
&\overset{(a)}{=} \sum_{\lambda_i<0.7}\left(\lambda_i - \lambda_i\log\lambda_i + e^{-\lambda_i}\frac{\lambda_i^2}{1-\lambda_i}\right) + \sum_{\lambda_i\geq 0.7} H(poi(\lambda_i)) \\
&\overset{(b)}{\leq} \sum_{\lambda_i<0.7}(3\lambda_i - \lambda_i\log\lambda_i) + \sum_{\lambda_i\geq 0.7}\frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) \qquad (2.8)
\end{aligned}
$$

where (a) is due to Lemma 24 and (b) is by using Equation (1) in [ALY10] and the fact that $e^{-x}\frac{x^2}{1-x} < 2x$ for $x < 0.7$.

Next in Lemmas 25 and 26 we evaluate an upper bound on the two summations in (2.8) with substituting the exact value of p_i for the class $\mathcal{P}_{(\text{zipf}(\alpha,k))}$.

Lemma 25. *For $n < k^{\frac{\alpha}{1.1}}$*

$$\sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) \leq \frac{1}{2}(\frac{1}{0.7C_{k,\alpha}})^{\frac{1}{\alpha}} \log\left(5.6\pi e^{\alpha+1}\right) n^{\frac{1}{\alpha}}.$$

Proof.

$$\begin{aligned}\sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) &= \sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\left(2\pi e\lambda_i + \frac{2\pi e}{12}\right) \\ &\leq \sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\lambda_i + \sum_{\lambda_i \geq 0.7}\left(1 + \frac{1}{2}\log 2\pi e\right)\end{aligned}$$

Let d' be such that $\lambda_{d'} = 0.7$, namely $d' = \left(\frac{n}{0.7C_{k,\alpha}}\right)^{\frac{1}{\alpha}}$. For proof simplicity lets assume d' is an integer. Considering the first term above,

$$\begin{aligned}\sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\lambda_i &= \sum_{i=1}^{d'} \frac{1}{2}\log\left(\frac{n}{C_{k,\alpha} i^{\alpha}}\right) \\ &= \frac{d'}{2}\log\frac{n}{C_{k,\alpha}} - \frac{\alpha}{2}\sum_{i=1}^{d'}\log i \\ &= \frac{d'}{2}\log\frac{n}{C_{k,\alpha}} - \frac{\alpha}{2}d'\log d' + \frac{\alpha}{2}(d'-1)\log e.\end{aligned}$$

Therefore,

$$\sum_{\lambda_i \geq 0.7} \frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) \leq \frac{d'}{2}\log\frac{n}{C_{k,\alpha}} - \frac{\alpha}{2}d'\log d' + \frac{\alpha}{2}(d'-1)\log e + d'\left(1 + \frac{1}{2}\log 2\pi e\right).$$

By substituting the value of d' we have the lemma. □

Lemma 26. *For $n < k^{\frac{\alpha}{1.1}}$*

$$\sum_{\lambda_i < 0.7} (3\lambda_i - \lambda_i \log \lambda_i) \leq \left(\frac{32.2\alpha - 25.2}{10(\alpha-1)^2} \left(\frac{1}{0.7C_{k,\alpha}} \right)^{\frac{1}{\alpha}} \right) n^{\frac{1}{\alpha}} + 0.7 \log \frac{n}{0.7C_{k,\alpha}}.$$

Proof. In the below calculations "≈" means that the quantities are equal up-to a multiplicative factor of $1 + o_n(1)$.

$$\begin{aligned}
\sum_{\lambda_i < 0.7} \lambda_i = & \sum_{i = \lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor + 1}^{k} n \frac{i^{-\alpha}}{C_{k,\alpha}} \\
\approx & n \int_{\left(\frac{10n}{7C_{k,\alpha}}\right)^{1/\alpha}}^{k} \frac{i^{-\alpha}}{C_{k,\alpha}} di \\
\approx & \frac{n}{(\alpha-1)C_{k,\alpha}} \left(\left(\frac{10n}{7C_{k,\alpha}} \right)^{-(\alpha-1)/\alpha} - k^{-(\alpha-1)} \right) \\
\approx & \frac{n}{(\alpha-1)C_{k,\alpha}} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-(\alpha-1)/\alpha} \\
= & \frac{7}{10(\alpha-1)} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}}
\end{aligned}$$

Let $n^- \stackrel{\text{def}}{=} \sum_{\lambda_i < 0.7} \lambda_i$. Then,

$$
\begin{aligned}
\sum_{\lambda_i<0.7} -\lambda_i \log \lambda_i = & \sum_{i=\lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor+1}^{k} n\frac{i^{-\alpha}}{C_{k,\alpha}} \log(\frac{C_{k,\alpha} i^{\alpha}}{n}) \\
=& n^{-} \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \sum_{i=\lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor+1}^{k} i^{-\alpha} \log i \\
\leq& n^{-} \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \int_{\left(\frac{10n}{7C_{k,\alpha}}\right)^{1/\alpha}-1}^{k} i^{-\alpha} \log i di \\
\leq& n^{-} \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \left[\frac{x^{1-\alpha}((\alpha-1)\log x + 1)}{(\alpha-1)^2} \right]_{k}^{\left(\frac{2n}{C_{k,\alpha}}\right)^{1/\alpha}} + \\
& \frac{n\alpha}{C_{k,\alpha}} \frac{1}{\alpha} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-1} \log\left(\frac{10n}{7C_{k,\alpha}} \right) \\
\leq& n^{-} \log(\frac{C_{k,\alpha}}{n}) + \frac{n}{C_{k,\alpha}} \frac{1}{\alpha-1} (\frac{10n}{7C_{k,\alpha}})^{\frac{1-\alpha}{\alpha}} \log \frac{10n}{7C_{k,\alpha}} + \frac{n\alpha}{C_{k,\alpha}(\alpha-1)^2} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}-1} \\
& + \frac{n}{C_{k,\alpha}} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-1} \log\left(\frac{10n}{7C_{k,\alpha}} \right) \\
\leq& \frac{7}{10(\alpha-1)} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}} \log(\frac{C_{k,\alpha}}{n}) + \frac{7}{10(\alpha-1)} (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \log \frac{10n}{7C_{k,\alpha}} \\
& + \frac{7\alpha}{10(\alpha-1)^2} (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} + \frac{7}{10} \log \frac{10n}{7C_{k,\alpha}} \\
=& \frac{11.2\alpha - 4.2}{10(\alpha-1)^2} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}} + \frac{7}{10} \log \frac{10n}{7C_{k,\alpha}}
\end{aligned}
$$

□

2.C Expected number of distinct elements

Lemma 27. *For $\mathcal{P}_{(\leq c \cdot i^{-\alpha})}$ and for $n < k^{\frac{\alpha}{1.1}}$,*

$$\mathbb{E}[\varphi_{+}^{n}] \leq \left(\frac{\alpha}{\alpha-1} \right) c^{1/\alpha} n^{1/\alpha}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[\varphi_+^n] &= \sum_{i=1}^{k} \mathbb{E}[\mathbb{I}_{\mu_i>0}] \\
&= \sum_{i=1}^{k} 1-(1-p_{(i)})^n \\
&\leq \sum_{i=1}^{k} 1-(1-f(i))^n \\
&\leq \sum_{i:f(i)\geq 1/n} 1 + \sum_{i:f(i)<1/n} 1-(1-f(i))^n \\
&\leq \sum_{i:f(i)\geq 1/n} 1 + \sum_{i:f(i)<1/n} nf(i).
\end{aligned}$$

Thus we need to bound the number of elements with envelope $\geq 1/n$ and the sum of envelopes for elements that are less than $1/n$. For $\mathcal{P}_{(\leq c\cdot i^{-\alpha})}$, the first term is $\leq (nc)^{1/\alpha}$ and the second term is

$$\begin{aligned}
\leq \sum_{i=(nc)^{1/\alpha}}^{k} cni^{-\alpha} &\leq \frac{c}{\alpha-1} n(nc)^{\frac{1-\alpha}{\alpha}} \\
&\leq \frac{c^{1/\alpha}}{\alpha-1} n^{1/\alpha}.
\end{aligned}$$

Combining these, we get

$$\mathbb{E}[\varphi_+^n] \leq \left(\frac{\alpha}{\alpha-1}\right) c^{1/\alpha} n^{1/\alpha}.$$

□

Lemma 28. *For $\mathcal{P}_{(zipf(\alpha,k))}$ and for $n < k^{\frac{\alpha}{1.1}}$,*

$$\frac{2}{3}(\frac{1}{C_{k,\alpha}})^{1/\alpha} n^{1/\alpha} \leq \mathbb{E}[\varphi_+^n] \leq \left(\frac{\alpha}{\alpha-1}\right)(\frac{1}{C_{k,\alpha}})^{1/\alpha} n^{1/\alpha}.$$

Proof.

$$\begin{aligned}\mathbb{E}[\varphi_+^n] &= \sum_{i=1}^{k} \mathbb{E}[\mathbb{I}_{\mu_i>0}] \\ &= \sum_{i=1}^{k} 1-(1-p_{(i)})^n \\ &\overset{(a)}{\geq} \sum_{i:p_{(i)}\geq 1/n} \frac{1}{2} + \sum_{i:p_{(i)}<1/n} \frac{np_{(i)}}{3},\end{aligned}$$

where (a) is because for $x \geq \frac{1}{n}$, $1-(1-x)^n > 0.5$ and for $x < \frac{1}{n}$, $1-(1-x)^n > \frac{nx}{3}$. We need to lower bound the number of elements having probability $\geq 1/n$ and the sum of probabilities for the elements with probability $< 1/n$. For $\mathcal{P}_{(\text{zipf}(\alpha,k))}$, the first term is $(\frac{n}{C_{k,\alpha}})^{1/\alpha}$ and the second term is

$$\sum_{i=(\frac{n}{C_{k,\alpha}})^{1/\alpha}}^{k} \frac{ni^{-\alpha}}{3C_{k,\alpha}} \geq \frac{1}{6}(\frac{n}{C_{k,\alpha}})^{1/\alpha}.$$

Combining these two, we get

$$\mathbb{E}[\varphi_+^n] \geq \frac{2}{3}(\frac{1}{C_{k,\alpha}})^{1/\alpha} n^{1/\alpha}.$$

The proof for other side of the inequality is similar to Lemma 27 and thus omitted. □

Chapter 3

Learning Power-Law Distributions: Absolute Discounting is Optimal

3.1 Introduction

Many natural problems involve uncertainties about categorical objects. When modeling language, we reason about words, meanings, and queries. When inferring about mutations, we manipulate genes, SNPs, and phenotypes.

It is sometimes possible to embed these discrete objects into continuous spaces, which allows us to use the arsenal of the latest machine learning tools that often (though admittedly not always) need numerically meaningful data. But why not operate in the discrete space directly? One of the main obstacles to this is the dilution of data due to the high-dimensional aspect of the problem, where dimension in this case refers to the number k of categories.

The classical framework of categorical distribution estimation, studied at length by the information theory community, involves a fixed small k, [BS04]. Add-constant estimators are sufficient for this purpose. Some of the impetus to understanding the large k regime came from the neuroscience world, [Pan04]. But this extended the pessimistic worst-case perspective of the earlier framework, resulting in guarantees that left a lot to be desired. This is because high-dimension often also comes with additional structure. In particular, if a distribution produces only roughly d distinct categories in a sample of size n, then we ought to think of d (and not k) as the *effective* dimension of the problem. There are also some ubiquitous structures, like power-law distributions. Natural language is a flagship example of this, which was observed as early as by Zipf in [Zip35]. Species and genera, rainfall, terror incidents, to mention just a few all obey power-laws [SLE$^+$03, CSN09, ADW13].

Are there estimators that mold to both dimension *and* structure? It turns out we don't need to search far. In natural language processing (NLP) it was first discovered that an estimator proposed by Good and Turing worked very well [Goo53]. Only recently did we start understanding why and how [OSZ03, OD12a, AJOS13a, OS15]. And the best explanation thus far is that it implicitly *competes* with the best estimator in a very small neighborhood of the true distribution. But NLP researchers [NEK94, KN95a, CG99] have long realized that another simpler estimator, *absolute discounting*, is equally good. But why and how this is the case was never properly determined, save some mention in [OD12a] and in [FNT16], where the focus is primarily on form.

In this chapter, we first show that absolute discounting, defined in Section 3.3, recovers pessimistic minimax optimality in both the low- and high-dimensional regimes. This is an immediate consequence of an upper bound that we provide in Section 3.5. We then study lower bounds with classes defined by the number of distinct categories d and also power-law structure in Section 3.6. This reveals that absolute discounting in fact *adapts* to the family of these classes. We further unravel the relationship of absolute discounting with the Good–Turing estimator, for power-law distributions. Interestingly, this leads to a further refinement of the estimator's performance in terms of *competitivity*. Lastly, we give some synthetic experiments in Section 3.8 and then explore forecasting global terror incidents on real data [LDMN16], which showcases very well the "all-dimensional" learning power of absolute discounting. These contributions are summarized in more detail in Section 3.4. We start out in Section 3.2 with laying out what we mean by these notions of optimality.

3.2 Optimal distribution learning

In this section we concretely formulate the optimal distribution learning framework and take the opportunity to point out related work.

3.2.1 Problem setting

Let $p = (p_1, p_2, \ldots, p_k)$ be a distribution over $[k] \overset{\text{def}}{=} \{1, 2, \ldots, k\}$ categories. Let $[k]^*$ be the set of finite sequences over $[k]$. An estimator q is a mapping that assigns to every sequence $x^n \in [k]^*$ a distribution $q(x^n)$ over $[k]$. We model p as being the underlying distribution over the categories. We have access to data consisting of n samples $X^n = X_1, X_2, ..., X_n$ generated *i.i.d.* from p. Intuitively, our goal is to find a choice of q that is guaranteed to be as close as any other estimator can be to p, in average. We first need to quantify how performance is measured.

General notation: Let $(\mu_j \,:\, j = 1, \cdots, k)$ denote the empirical counts, i.e. the number of times symbol j appears in X^n and let D be the number of *distinct* categories appearing in X^n, i.e. $D = \sum_j \mathbb{1}\{\mu_j > 0\}$. We denote by $d \overset{\text{def}}{=} \mathbb{E}[D]$ its expectation. Let $(\Phi_\mu \,:\, \mu = 0, \cdots, n)$, be the total number of categories appearing exactly μ times, $\Phi_\mu \overset{\text{def}}{=} \sum_j \mathbb{1}\{\mu_j = \mu\}$. Note that $D = \sum_{\mu > 0} \Phi_\mu$. Also let $(S_\mu \,:\, \mu = 0, \cdots, n)$, be the total probability within each such group, $S_\mu \overset{\text{def}}{=} \sum_j p_j \mathbb{1}\{\mu_j = \mu\}$. Lastly, denote the empirical distribution by $q_j^{+0} \overset{\text{def}}{=} \mu_j / n$.

3.2.2 KL-Risk

We adopt the Kullback-Leibler (KL) divergence as a measure of loss between two distributions. When a distribution p is approximated by another q, the KL divergence is given by

$\mathsf{KL}(p||q) \overset{\text{def}}{=} \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}$. We can then measure the performance of an estimator q that depends on data in terms of the *KL-risk*, the expectation of the divergence with respect to the samples.

We use the following notation to express the KL-risk of q after observing n samples X^n:

$$r_n(p,q) \overset{\text{def}}{=} \underset{X^n \sim p^n}{\mathbb{E}}[\mathsf{KL}(p||q(X^n))].$$

An estimator that is identical to p regardless of the data is unbeatable, since $r_n(p,q) = 0$. Therefore it is important to model our ignorance of p and gauge the optimality of an estimator q accordingly. This can be done in various ways. We elaborate the three most relevant such perspectives: *minimax*, *adaptive*, and *competitive* distribution learning.

3.2.3 Minimax

In the *minimax* setting, p is only known to belong to some class of distributions $\mathcal{P}$, but we don't know which one. We would like to perform well, no matter which distribution it is. To each q corresponds a distribution $p \in \mathcal{P}$ (assuming the class is finite or closed) on which q has its *worst* performance:

$$r_n(\mathcal{P},q) \overset{\text{def}}{=} \max_{p \in \mathcal{P}} r_n(p,q).$$

The minimax risk is the *least* worst-case KL-risk achieved by *any* estimator q,

$$r_n(\mathcal{P}) \overset{\text{def}}{=} \min_{q} r_n(\mathcal{P},q).$$

The minimax risk depends only on the class $\mathcal{P}$. It is a *lower bound*: no estimator can beat it *for all p*, i.e. it's not possible that $r_n(p,q) < r_n(\mathcal{P})$ for all $p \in \mathcal{P}$. An estimator q that satisfies an *upper bound* of the form $r_n(\mathcal{P},q) = (1+o(1))r_n(\mathcal{P})$ is said to be minimax *optimal* "even to the constant" (an informal but informative expression that we adopt in this chapter). If instead $r_n(\mathcal{P},q) = O(1)r_n(\mathcal{P})$, we say that q is *rate optimal*. Near-optimality notions are also possible, but we don't dwell on them. As an aside, note that *universal compression* is minimax optimality using *cumulative* risk. See [FJO$^+$15] for such related work on universal compression for power laws.

3.2.4 Adaptive

The minimax perspective captures our ignorance of p in a pessimistic fashion. This is because $r_n(\mathcal{P})$ may be large, but for a specific $p \in \mathcal{P}$ we may have a much smaller $r_n(p,q)$. How can we go beyond this pessimism? Observe that when a class is smaller, then $r_n(\mathcal{P})$ is smaller. This is because we'd be maximizing on a smaller set. In the extreme case noted earlier, when $\mathcal{P}$ contains only a single distribution, we have $r_n(\mathcal{P}) = 0$. The *adaptive* learning setting finds an intermediate ground where we have a *family* of distribution classes $\mathcal{F} = \{\mathcal{P}_s : s \in \mathcal{S}\}$ indexed by a (not necessarily countable) index set $\mathcal{S}$. For each s, we have a corresponding $r_n(\mathcal{P}_s)$ which is often much smaller than $r_n\left(\bigcup_{s\in\mathcal{S}} \mathcal{P}_s\right)$, and we would like the estimator to achieve the risk bound corresponding to the smaller class. We say that an estimator q is *adaptive* to the family $\mathcal{F}$ if for all $s \in \mathcal{S}$:

$$r_n(p,q) \leq \ O_s(1)\ r_n(\mathcal{P}_s) \quad \forall p \in \mathcal{P}_s \quad \Longleftrightarrow \quad r_n(\mathcal{P}_s,q) \leq \ O_s(1)\ r_n(\mathcal{P}_s)$$

There often is a price to adaptivity, which is a function of the granularity of $\mathcal{F}$ and is paid in the form of varying/large leading constants per class. This framework has been particularly successful in density estimation with smoothness classes [Tsy09] and has been recently used in the discrete setting for universal compression [BGO15].

3.2.5 Competitive

The adaptive perspective can be tightened by demanding that, rather than a multiplicative constant, the KL-risk tracks the risk up to a vanishingly small *additive* term:

$$r_n(p,q) = r_n(\mathcal{P}_s) + \varepsilon_n(\mathcal{P}_s, q) \quad \forall p \in \mathcal{P}_s.$$

Ideally, we would like the *competitive loss* $\varepsilon_n(\mathcal{P}_s, q)$ to be negligible compared to the risk of each class $r_n(\mathcal{P}_s)$. If $\varepsilon_n(\mathcal{P}_s, q) = O_s(1) r_n(\mathcal{P}_s)$ for all s, then we recover adaptivity. And when $\varepsilon_n(\mathcal{P}_s, q) = o_s(1) r_n(\mathcal{P}_s)$ for all $s \in \mathcal{S}$, we have minimax optimality even to the constant within each class, which is a much stronger form of adaptivity. We then say that the estimator is *competitive* with respect to the family $\mathcal{F}$. We may also evaluate the *worst-case* competitive loss, over $\mathcal{S}$.

This formulation was recently introduced in [OS15] in the context of distribution learning. This work shows that the celebrated Good–Turing estimator [Goo53], combined with the empirical estimator, has small worst-case competitive loss over the family of classes defined by any given distribution and all its permutations. Most importantly, this loss was shown to stay bounded, even as the dimension increases. This provided a rigorous theoretical explanation for the performance of the Good–Turing estimator in high-dimensions. A similar framework is also studied for ℓ_1-loss in [VV15].

3.3 Absolute discounting

One of the first things to observe is that the empirical distribution is particularly ill-suited to handle KL-risk. This is most easily seen by the fact that we'd have infinite blow-up when any $\mu_j = 0$, which *will* happen with positive probability. Instead, one could resort to an add-constant estimator, which for a positive β is of the form $q_j^{+\beta} \overset{\text{def}}{=} (\mu_j + \beta)/(n + k\beta)$.

The most widely-studied class of distributions is the one that includes all of them:

the $k-$dimensional simplex, $\Delta_k \overset{\text{def}}{=} \{(p_1, p_2, \ldots, p_k),: \sum_i p_i = 1,\ p_i \geq 0\ \forall i \in [k]\}$. In the low-dimensional scaling, when $n/k \to \infty$ (the "dimension" here being the support size k), the minimax risk is

$$r_n(\Delta_k) = (1 + o(1))\ \frac{k-1}{2n},$$

In [BS04], a variant of the add-constant estimator is shown to achieve this risk even to the constant. Furthermore, any add-constant estimator is rate optimal when k is fixed. But in the very high-dimensional setting, when $k/n \to \infty$, [Pan04] showed that the minimax risk behaves as

$$r_n(\Delta_k) = (1 + o(1))\ \log \frac{k}{n},$$

achieved by an add-constant estimator, but with a constant that depends on the ratio of k and n.

Despite these classical results on minimax optimal estimators, in practice people often use other estimators that have better empirical performance. This was a long-running mystery in the language modeling community [CG99], where variants of the Good–Turing estimator were shown to perform the best [JM85, GS95]. The gap in performance was only understood recently, using the notion of competitivity [OS15]. In essence, the Good–Turing estimator works well in *both* low- and high-dimensional regimes, and in-between. Another estimator, *absolute discounting*, unlike add-constant estimators, simply *subtracts* a positive constant from the empirical counts and redistributes the subtracted amount to unseen categories. For a discount parameter $\delta \in [0, 1)$, it is defined as:

$$q_j^{-\delta} \overset{\text{def}}{=} \begin{cases} \frac{\mu_j - \delta}{n} & \text{if } \mu_j > 0, \\ \frac{D\delta}{n(k-D)} & \text{if } \mu_j = 0. \end{cases} \tag{3.1}$$

Starting with the work of [NEK94], absolute discounting soon supplanted the Good–Turing estimator, due to both its simplicity and comparable performance. Kneser-Ney smoothing [KN95a], which uses absolute discounting at its core was long held as the preferred way to train

N-gram models. Even to this day, the state-of-the-art language models are combined systems where one usually interpolates between recurrent neural networks and Kneser-Ney smoothing [JVS$^+$16]. Can this success be explained?

Kneser-Ney is for the most part a principled implementation of the notion of back-off, which we only touch upon in the conclusion. The use of absolute discounting is critical however, as performance deteriorates if we back-off with care but use a more QAïve add-constant or even Katz-style smoothing [Kat87], which switches from the Good–Turing to the empirical distribution at a fixed frequency point.

It is also important to mention the Bayesian approach of [Teh06a] that performs similarly to Kneser-Ney, called the Hierarchical Pitman-Yor language model. The hierarchies in this model reprise the role of back-off, while the two-parameter Poisson-Dirichlet prior proposed by Pitman and Yor [PY97] results in estimators that are very similar to absolute discounting. The latter is not a surprise because this prior almost surely generates a power law distribution, which is intimately related to absolute discounting as we study in this chapter. Though our theory applies more generally, it can in fact be straightforwardly adapted to give guarantees to estimators built upon this prior.

3.4 Contributions

We investigate the reason behind the auspicious behavior of the absolute discounting estimator. We achieve this by demonstrating the adaptivity and competitivity of this estimator for many relevant families of distribution classes. In summary:

- We analyze the performance of the absolute discounting estimator by upper bounding the KL-risk for each class in a family of distribution classes defined by the expected number of distinct categories. [Section 3.5, Theorem 29] This result implies that absolute discounting achieves classical minimax rate-optimality in *both* the low- and high-dimensional regimes over the whole simplex Δ_k, as outlined in Section 3.2.

- We provide a generic lower bound to the minimax risk of classes defined by a single distribution and all of its permutations. We then show that if the defining distribution is a truncated (possibly perturbed) power-law, then this lower bound matches the upper bound of absolute discounting, up to a constant factor. [Section 3.6, Corollaries 31 and 32]

- This implies that absolute discounting is adaptive to the family of classes defined by a truncated power-law distribution and its permutations. Also, since classes defined by the expected number of distinct categories necessarily includes a power-law, absolute discounting is also adaptive to this family. This is a strict refinement of classical minimax rate-optimality.

- We give an equivalence between the absolute discounting and Good–Turing estimators in the high-dimensional setting, whenever the distribution is a truncated power-law. This is a finite-sample guarantee, as compared to the asymptotic version of [OD12a].

  As a consequence, absolute-discounting becomes competitive with respect to the family of classes defined by permutations of power-laws, inheriting Good–Turing's behavior [OS15]. [Section 3.7, Lemma 33 and Theorem 34]

We corroborate the theoretical results with synthetic experiments that reproduce the theoretical minimax risk bounds. We also show that the prowess of absolute discounting on real data is not restricted only to language modeling. In particular, we explore a striking application to forecasting global terror incidents and show that, unlike naive estimators, absolute discounting gives accurate predictions simultaneously in all of low-, medium-, and high-activity zones. [Section 3.8]

3.5 Upper bound and classical minimax optimality

We now give an upper bound for the risk of the absolute discounting estimator and show that it recovers classical minimax rates in the low- and high-dimensional regimes. Recall that $d \stackrel{\text{def}}{=} \mathbb{E}[D]$ is the expected number of distinct categories in the samples. The upper bound that we derive can be written as function of only d, k, and n, and is non-decreasing in d. For a given n and k, let $\mathcal{P}_d$ be the set of all distributions for which $\mathbb{E}[D] \leq d$. The upper bound is thus also a worst-case bound over $\mathcal{P}_d$.

Theorem 29 (Upper bound)**.** *Consider the absolute discounting estimator $q = q^{-\delta}$, defined in* (3.1)*. Let p be such that $\mathbb{E}[D] = d$. Given a discount* $0 < \delta < 1$*, there exists a constant c that may depend on* δ *and only* δ*, such that*

$$r_n(p,q) \leq \begin{cases} \frac{d}{n} \log \frac{k - \frac{d}{2}}{\frac{d}{2}} + c\frac{d}{n} & \textit{if} \quad d \geq 10 \log\log k, \\ \frac{d}{n} \log k + c\frac{d}{n} & \textit{if} \quad d < 10 \log\log k. \end{cases} \tag{3.2}$$

The same bound holds for $r_n(\mathcal{P}_d, q)$.

We defer the proof the theorem to the supplementary material. Here are the immediate implications. For the low-dimensional regime $\frac{n}{k} \to \infty$ and the class Δ_k, the largest d can be once $n > k$ is k. The risk of absolute discounting is thus bounded by $c(1+o(1))\frac{k}{n} = O(1)\frac{k}{n}$. This is minimax rate-optimal [BS04]. For the high-dimensional regime $\frac{k}{n} \to \infty$ and the class Δ_k, the largest d can be when $k > n$ is n. The risk of absolute discounting is thus dominated by the first term, which reduces to $(1+o(1))\log\frac{k}{n}$. This is the optimal risk for the class Δ_k [Pan04], even to the constant.

Therefore on the two extreme ranges of k and n absolute discounting recovers the best performance, either as rate-optimal or optimal even to the constant. These results are for the entire $k-$dimensional simplex Δ_k. Furthermore, for smaller classes, it characterizes the worst-case risk of the class by the d, the expected number of distinct categories. Is this characterization tight?

3.6 Lower bounds and adaptivity

In order to lower bound the minimax risk of a given class $\mathcal{P}$, we use a finer granularity than the $\mathcal{P}_d$ classes described in Section 3.5. In particular, let $\mathcal{P}_p$ be the *permutation class* of distributions consisting of a single distribution p and all of its permutations. Note that the multiset of probabilities is the same for all distributions in $\mathcal{P}_p$, and since the expected number of distinct categories only depends on the multiset ($d = \sum_j [1-(1-p_j)^n]$) it follows that $\mathcal{P}_p \subset \mathcal{P}_d$[1]. To find a good lower bound for $\mathcal{P}_d$, we need a p that is "worst case". In what follows, we start by giving a lower bound for $\mathcal{P}_p$, and then specialize it for $\mathcal{P}_d$.

We also assume that an oracle specifies the *true* probability of all observed categories. With this side-information, the best estimator *has* to use the true probabilities for the observed categories. For the unobserved categories, it needs to redistribute all the missing mass (the total probability of unobserved categories). Since the multiset of probabilities is fixed and any permutation of the remaining categories is equally probable, by symmetry there is no advantage in favoring one over the other. Therefore the best oracle-aided estimator is uniquely specified: exact probabilities for seen categories and uniform redistribution of the missing mass (S_0) over the unobserved categories. This argument can be proven formally via the maximin trick: substitute the maximum with a mean against an arbitrary prior over p, at which point the optimal q is the posterior, and then optimize over priors. It then suffices to use the convexity of $p \log \frac{p}{q}$ with respect to q. We first give the following generic lower bound.

Theorem 30 (Generic lower bound). *Let $\mathcal{P}_p$ be a permutation class defined by a distribution p and let $\gamma > 1$. Then for $k > \gamma d$, the minimax risk is bounded by:*

$$r_n(\mathcal{P}_p) \geq \left(1-\frac{1}{\gamma}\right)\left(\sum_{j=\gamma d}^{k} p_j\right)\log\frac{k-\gamma d}{\sum_{j=\gamma d}^{k} p_j} + \sum_{i=\gamma d} p_j \log p_j \tag{3.3}$$

[1] We abuse notation by distinguishing the classes by the letter used, while at the same time using the letters to denote actual quantities. From the context we understand that d is the expected number of distinct categories for p, at the given n.

Equation (3.3) can be used as a starting point for more concrete lower bounds on various distribution classes. We illustrate this for two cases. First, let us choose p to be a truncated power-law distribution with power α: $p_j \propto j^{-\alpha}, \quad \text{for } j = 1, \cdots, k$. We always assume $\alpha \geq \alpha_0 > 1$. This leads to the following lower bound.

Corollary 31. *Let $\mathcal{P}$ be all permutations of a single power-law distribution with power α truncated over k categories. Then there exists a constant $c > 0$ and large enough n_0 such that when $n > n_0$ and $k > \max\{n, 1.2^{\frac{1}{\alpha-1}} n^{\frac{1}{\alpha}}\}$,*

$$r_n(\mathcal{P}) \geq c\frac{d}{n}\log\frac{k-2d}{2d}.$$

Next, we use a different choice of p for $\mathcal{P}_p$ to provide a lower bound whenever d grows linearly with n. This essentially closes the gap of the previous corollary when α approaches 1.

Corollary 32. *Let $\rho \in (1, 1.75)$ and let $\mathcal{P}$ be all permutations of a single uniform distribution over a subset $k' = \frac{n}{\rho}$ out of k categories. Then $d \sim (1 - e^{-\rho})n/\rho$ and there exists a constant $c > 0$ and large enough n_0 such that when $n > n_0$ and $k > n^5$,*

$$r_n(\mathcal{P}) \geq c\frac{d}{n}\log\frac{k-1.2d}{d} .$$

We defer the proofs of the theorem and its corollaries to the supplementary material. The upper bound of Theorem 29 and the lower bounds of Corollaries 31 and 32 are within constant factors of each other. The immediate consequence is that absolute discounting is adaptive with respect to the families of classes of the Corollaries. Furthermore, over the family of classes $\mathcal{P}_d$ where we can write d as $n^{\frac{1}{\alpha}}$ for some $\alpha > 1$ or $d \propto n$, we can select a distribution from the Corollaries among each class and use the corresponding lower bound to match the upper bound of Theorem 29 up to a constant factor. Therefore absolute discounting is adaptive to this family of classes.

Intuitively, adaptivity to these classes establishes optimality in the intermediate range between low- and high-dimensional settings in a distribution-dependent fashion and governed by the expected number of distinct categories d, which we may regard as the *effective* dimension of the problem.

3.7 Relationship to Good–Turing and competitivity

We now establish a relationship between the absolute discounting and Good–Turing estimators and refine the adaptivity results of the previous section into competitivity results.

When [OS15] introduced the notion of competitive optimality, they showed that a variation of the Good–Turing estimator is worst-case competitive with respect to the family of distribution classes defined by any given probability distribution and its permutations. In light of the results of Sections 3.5 and 3.6, it is natural to ask whether absolute discounting enjoys the same kind of competitive properties. Not only that, but it was observed empirically by [NEK94] and shown theoretically in [OD12a] that *asymptotically* Good–Turing behaves exactly like absolute discounting, when the underlying distribution is a (possibly perturbed) power-law. We therefore choose this family of classes to prove competitivity for. We first make the aforementioned equivalence concrete by establishing a *finite sample* version. We use the following *idealized* version of the Good–Turing estimator [Goo53]:

$$q_j^{\mathsf{GT}} \stackrel{\text{def}}{=} \begin{cases} \frac{\mu_j+1}{n}\frac{\mathbb{E}[\Phi_{\mu_j+1}]}{\mathbb{E}[\Phi_{\mu_j}]} & \text{if } \mu_j > 0, \\ \frac{\mathbb{E}[\Phi_1]}{n(k-D)} & \text{if } \mu_j = 0. \end{cases} \tag{3.4}$$

Lemma 33. *Let p be a power law with power α truncated over k categories. Then for $k > \max\{n, n^{\frac{1}{\alpha-1}}\}$, we have the equivalence:*

$$q_j^{\mathsf{GT}} = \frac{\mu_j - \frac{1}{\alpha}}{n}\left(1 + O\left(n^{-\frac{1}{2}\frac{3}{2\alpha+1}}\right)\right) \sim \frac{\mu_j - \frac{1}{\alpha}}{n} \qquad \forall\, \mu_j \in \left\{1, \cdots, n^{\frac{1}{2\alpha+1}}\right\}.$$

An interesting outcome of the equivalence of Lemma 33 is that it suggests a choice of the discount δ in terms of the power, $1/\alpha$. To give a data-driven version of $1/\alpha$, we will use a robust version of the ratio Φ_1/D proposed in [OD12a, BBO17], which is a strongly consistent estimator when $k = \infty$.

Theorem 34. *Let $\mathcal{P}$ be all permutations of a truncated power law p with power α. Let q be the absolute discounting estimator with $\delta = \min\left\{\frac{\max\{\Phi_1,1\}}{D}, \delta_{\max}\right\}$, for a suitable choice of $\delta_{\max}$. Then for $k > \max\{n, n^{\frac{1}{\alpha-1}}\}$, the competitive loss is*

$$\varepsilon_n(\mathcal{P}_p, q) = \mathcal{O}\left(n^{-\frac{2\alpha-1}{2\alpha+1}}\right) .$$

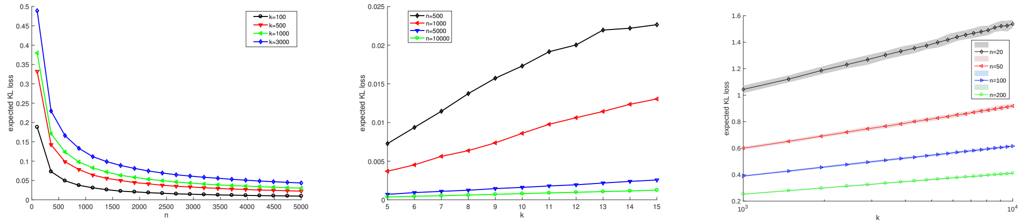
The implications are as follows. For the union of all such classes above a given α, we find that we beat the $n^{-1/3}$ rate of the worst-case competitive loss obtained for the estimator in [OS15].

Theorem 34 and the bounds of Sections 3.5 and 3.6, together imply that absolute discounting is not only worst-case competitive, but also *class-by-class* competitive with respect to the power-law permutation family. In other words, it in fact achieves minimax optimality even to the constant.

One of the advantages of absolute discounting is that it gradually transitions between values that are close to the empirical distribution for abundant categories (since μ then dominates the discount δ), to a behavior that imitates the Good–Turing estimator for rare categories (as established by Lemma 33). In contrast, the estimator proposed in [OS15], and its antecedents starting from [Kat87], have to carefully choose a threshold where they switch abruptly from one estimator to the other.

3.8 Experiments

We now illustrate the theory with some experimental results. Our purpose is to (1) validate the functional form of the risk as given by our lower and upper bounds and (2) compare absolute discounting on both synthetic and real data to estimators that have various optimality guarantees. In all synthetic experiments, we use 500 Monte Carlo iterations. Also, we set the discount value based on data, $\delta = \min\{\frac{\max(\Phi_1,1)}{D}, 0.9\}$. This is as suggested in Section 3.7, assuming $\delta_{\max} = 0.9$ is sufficient.



(a) k fixed (b) n fixed, $k << n$ (c) n fixed, $k >> n$

Figure 3.1: Risk of absolute discounting in different ranges of k and n for a power-law with $\alpha = 2$.

3.8.1 Validation

For our first goal, we consider absolute discounting in isolation. Figure 3.1(a) shows the decay of KL-risk with the number of samples n for a power-law distribution. The dependence of the risk on the number of categories k is captured in Figures 3.1(b) (linear x-axis) and 4.1(c) (logarithmic x-axis). Note the linear growth when k is small and the logarithmic growth when k is large. For the last plot we give 95% confidence intervals for the simulations, by performing 100 restarts.

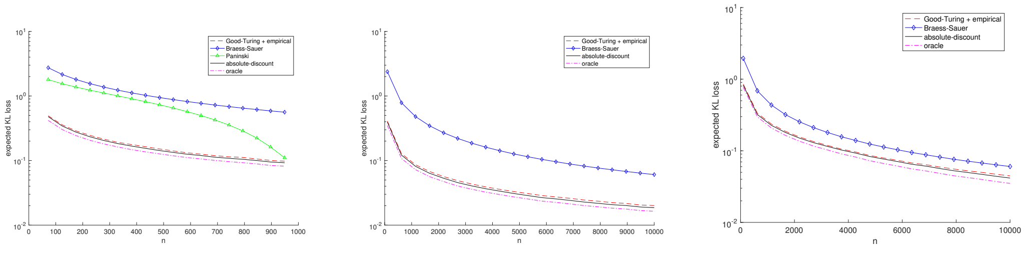
3.8.2 Synthetic data

For our second goal, we start with synthetic data. In Figure 3.2, we pit absolute discounting against a number of distributions related to power-laws. The estimators used for our comparisons are: empirical $q^{+0}(x) = \frac{\mu_x}{n}$, add-beta $q^{+\beta}(x) = \frac{\mu_x + \beta_{\mu_x}}{N}$, and its two variants:

- Braess and Sauer, q^{BS} [BS04] $q^{+\beta}$ with $\beta_0 = 0.5$, $\beta_1 = 1$, and $\beta_i = 0.75\ \forall i \geq 2$
- Paninski, q^{Pan} [Pan04] $q^{+\beta}$ with $\beta_i = \frac{n}{k} \log \frac{k}{n}\ \forall i$,

absolute discounting, $q^{-\delta}$, described in 3.1, Good–Turing + empirical q^{GT} in [OS15], and an oracle-aided estimator where S_μ is known.

In Figures 3.2(a) and 3.2(b), samples are generated according to a power-law distribution with power $\alpha = 2$ over $k = 1,000$ categories. However, the underlying distribution in Figure 3.2(c) is a piece-wise power-law. It consists of three equal-length pieces, with powers 1.3, 2, and 1.5. Paninski's estimator is not shown in Figures 3.2(b) and 3.2(c) since it is not well-defined in this range (it is designed for the case $k > n$ only). Unsurprisingly, absolute discounting dominates these experiments. What is more interesting is that it does not seem to need a pure power-law (similar results hold for other kinds of perturbations, such as mixtures and noise). Also Good–Turing is a tight second.



(a) pure power-law (b) pure power-law (c) piece-wise power-law

Figure 3.2: Comparing estimators for power-law variants with power $\alpha = 2$ and $k = 1000$.

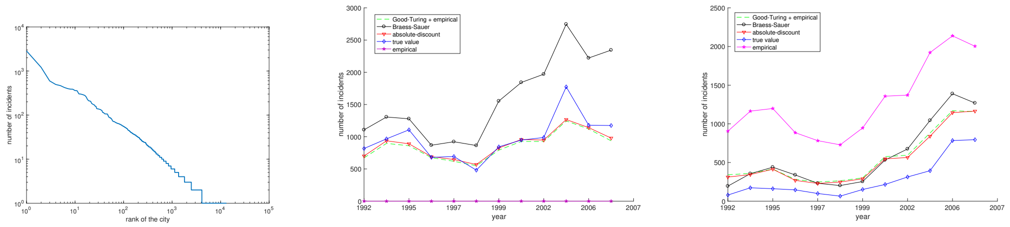
3.8.3 Real data

One of the chief motivations to investigate absolute discounting is natural language modeling. But there have been such extensive empirical studies that have verified over and over the power of absolute discounting (see the classical survey of [CG99]) that we chose to use this space for something new. We use the START *Global terrorism database* from the University of Maryland [LDMN16] and explore how well we can forecast the number of terrorist incidents in different cities. The data contains the record of more than $50,000$ terror incidents between the years 1992 and 2010, in more than $12,000$ different cities around the world. First, we display in Figure 3.3(a) the frequency of incidents across the entire dataset versus the activity rank of the city in log-log scale, showing a striking adherence to a power-law (see [CSN09] for more on this).

The forecasting problem that we solve is to estimate the number of total incidents in a subset of the cities over the coming year, using the current year's data from all cities. In order to emulate the various dimension regimes, we look at three subsets: (1) low-activity cities with *no* incidents in the current year and less than 20 incidents in the whole data, (2) medium-activity cities, with *some* incidents in the current year and less than 20 incidents in the whole data, and (3) high-activity individual cities with a large number of overall incidents.

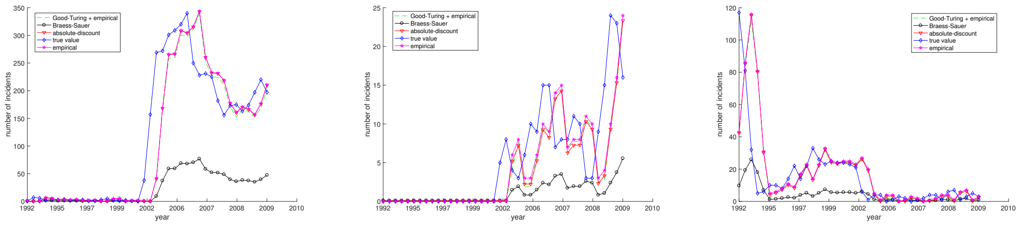
The results for (1) are in Figure 3.3(b). The frequency estimator trivially estimates zero. Braess-Sauer does something meaningful. But absolute discounting and Good–Turing estimators, indistinguishable from each other, are remarkably on spot. And this, without having observed any of the cities! This nicely captures the importance of using structure when dimensionality is so high and data is so scarce. The results for (2) are in Figure 3.3(c). The frequency estimator markedly overestimates. But now absolute discounting, Good–Turing, and Braess-Sauer, perform similarly. This is a lower dimensional regime than in (1), but still not adequate for simply using frequencies. This changes in case (3), illustrated in Figure 3.4. To take advantage of the abundance of data, in this case at each time point we used the previous $2,000$ incidents for learning, and predicted the share of each city for the next $2,000$ incidents. In fact, incidents are so abundant that we can simply rely on the previous window's count. Note how Braess-Sauer over-penalizes such abundant categories and suffers, whereas absolute discounting and Good–Turing continue to hold their own, mimicking the performance of the empirical counts. This is a very low-dimensional regime.

The closeness of the Good–Turing estimator to absolute discounting in all of our experiments validates the equivalence result of Lemma 33. The robustness in various regimes and the improvement in performance over such minimax optimal estimators as Braess-Sauer's and Paninski's are evidence that absolute discounting truly molds to both the raw dimension and effective dimension / structure.



(a) frequency vs rank (b) unobserved cities (c) observed cities

Figure 3.3: (a) power-law behavior of frequency vs rank in terror incidents, (b), and (c) comparing forecasts of the number of incidents in unobserved cities and observed ones, respectively.



(a) Baghdad (b) Fallujah (c) Belfast

Figure 3.4: Estimating the number of incidents based on previous data for different cities.

3.9 Acknowledgment

Chapter 3, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati, 2017. The dissertation author was one of the primary investigators and authors of this paper.

3.A Proof of Theorem 29, upper bound

We start with a technical note. Though we presented the framework for a fixed sample size n, the entirety of the chapter analyzes the "Poissonized" version. In the Poisson sampling model, the number of samples is in fact $N \sim \text{POI}(n)$, a Poisson random variable with mean n. This is often the more natural model when data is collected within a fixed time window, in contrast to until a certain number of samples are collected. Or we can think of Poisson sampling as a convenience because it makes all counts independent and distributed according to $\mu_j \sim \text{POI}(np_j)$. It is possible to "de-Poissonize" the results, but we omit this for brevity.

In proof of the theorem, we show a more general upper bound. We upper bound the instantaneous risk of a class of distributions based on d, $\mathbb{E}[\Phi_1]$, and $\mathbb{E}[\Phi_2]$, the expected number of distinct categories, categories that appeared once, and twice respectively. Namely, we show for some constant c,

$$\max_{p \in \mathcal{P}_d} \mathbb{E}_{x^n}\left[\mathsf{KL}(p||q(x^n))\right] \leq \frac{\mathbb{E}[\Phi_1]}{n} \log \frac{2k-d}{d\delta} + \frac{\mathbb{E}[\Phi_1]}{n} + \frac{2\mathbb{E}[\Phi_2]}{n} \log \frac{1}{1-\delta} + \frac{c \cdot d}{n}.$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{X^n\sim p^n}\left[\mathsf{KL}(p||q(X^n))\right]\\
&=\mathbb{E}_{X^n\sim p^n}\left[\sum_{j=1}^{k} p_j\log\frac{p_j}{q_j(X^n)}\right]\\
&=\mathbb{E}\left[\sum_{j=1}^{k}\mathbb{1}_j^0 p_j\log\frac{np_j(k-D)}{D\delta}+\sum_{j=1}^{k}\sum_{i=1}^{\infty}\mathbb{1}_j^i p_j\log\frac{np_j}{i-\delta}\right]\\
&=\mathbb{E}\left[\sum_{j=1}^{k}\mathbb{1}_j^0 p_j\log np_j+\mathbb{1}_j^0 p_j\log\frac{(k-D)}{D\delta}+\sum_{j=1}^{k}\sum_{i=1}^{\infty}\mathbb{1}_j^i p_j\log\frac{np_j}{i-\delta}\right]\\
&\overset{(a)}{=}\frac{1}{n}\sum_{j=1}^{k}e^{-\lambda_j}\lambda_j\log\lambda_j+\frac{1}{n}\sum_{j=1}^{k}\lambda_j\mathbb{E}\left[\mathbb{1}_j^0\log\frac{k-D}{D\delta}\right]+\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{\infty}\lambda_j\log\frac{\lambda_j}{i-\delta}\mathrm{poi}(np_j,i)\\
&\overset{(b)}{=}\frac{1}{n}\sum_{j=1}^{k}\lambda_j\mathbb{E}\left[\mathbb{1}_j^0\log\frac{k-D}{D\delta}\right]+\frac{1}{n}\sum_{j=1}^{k}\left(\lambda_j\log\lambda_j+\sum_{i=1}^{\infty}\lambda_j\log\frac{1}{i-\delta}\mathrm{poi}(\lambda_j,i)\right) \qquad (3.5)
\end{aligned}
$$

where (a) is by Poisson sampling and replacing $\lambda_j \stackrel{\text{def}}{=} np_j$, and (b) is by combining the first and last expressions. Now we state two lemmas that are helpful in bounding each of the two terms in (3.5).

Lemma 35. *For all $p\in\mathcal{P}_d$ and with the assumption of $D>2$, for $d>10\log\log k$,*

$$\mathop{\mathbb{E}}_{X^n\sim p^n}\left[\mathbb{1}_j^0\log\frac{k-D}{D}\right]\le e^{-\lambda_j}\left(1+\log\frac{k-\frac{d}{2}}{\frac{d}{2}}\right)$$

and for $d<10\log\log k$,

$$\mathop{\mathbb{E}}_{X^n\sim p^n}\left[\mathbb{1}_j^0\log\frac{k-D}{D}\right]\le e^{-\lambda_j}\log k.$$

Lemma 36. *For $x>0$ and for $0<\delta<1$, $x\log x+\sum_{i=2}^{\infty}x\log\frac{1}{i-\delta}\mathrm{poi}(x,i)<c'$ for some constant c'.*

We can write the second part in (3.5) as

$$\frac{1}{n}\sum_{j=1}^{k}\lambda_j\log\frac{1}{1-\delta}\mathrm{poi}(\lambda_j,1)+\frac{1}{n}\sum_{j=1}^{k}\left(\lambda_j\log\lambda_j+\sum_{i=2}^{\infty}\lambda_j\log\frac{1}{i-\delta}\mathrm{poi}(\lambda_j,i)\right), \qquad (3.6)$$

and since the second term in (3.6) is negative for all $\lambda_j < 1$, (3.6) is upper bounded by

$$\frac{1}{n}\sum_{j=1}^{k}\lambda_j^2 e^{-\lambda_j}\log\frac{1}{1-\delta}+\frac{1}{n}\sum_{\lambda_j\geq 1}\left(\lambda_j\log\lambda_j+\sum_{i=2}^{\infty}\lambda_j\log\frac{1}{i-\delta}\mathrm{poi}(\lambda_j,i)\right).$$

Continuing from (3.5), using Lemmas 35, 36 and the definitions of $\mathbb{E}[\Phi_1]=\sum_{j=1}^{k}e^{-\lambda_j}\lambda_j$ and $\mathbb{E}[\Phi_2]=\sum_{j=1}^{k}e^{-\lambda_j}\frac{\lambda_j^2}{2}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{X^n\sim p^n}[\mathsf{KL}(p||q(X^n))] &\leq \frac{\mathbb{E}[\Phi_1]}{n}\left(\log\frac{2k-d}{d\delta}+1\right)+\frac{2\mathbb{E}[\Phi_2]}{n}\log\frac{1}{1-\delta}+\frac{1}{n}\sum_{j:\lambda_j\geq 1}c'\\ &\leq \frac{\mathbb{E}[\Phi_1]}{n}\left(\log\frac{2k-d}{d\delta}+1\right)+\frac{2\mathbb{E}[\Phi_2]}{n}\log\frac{1}{1-\delta}+\frac{c'\cdot d}{n(1-e^{-1})},\end{aligned}$$

where the last line is because $d=\sum_j(1-e^{\lambda_j})\geq\sum_{\lambda_j\geq 1}(1-e^{\lambda_j})\geq|\{j:\lambda_j\geq 1\}|(1-e^{-0.7})$ □

3.A.1 Proof of Lemma 35

Proof. Using Lemma 45,

$$\begin{aligned}\mathop{\mathbb{E}}_{X^n\sim p^n}[\mathbb{1}_j^0\log\frac{k-D}{D}] &= e^{-\lambda_j}\mathbb{E}\left[\log\frac{k-D}{D}\middle|D<d-\sqrt{2vs},\mu_j=0\right]\Pr(D<d-\sqrt{2vs})\\ &\quad+e^{-\lambda_j}\mathbb{E}\left[\log\frac{k-D}{D}\middle|D>d-\sqrt{2vs},\mu_j=0\right]\Pr(D>d-\sqrt{2vs})\\ &\leq e^{-\lambda_j}\left(e^{-s}\log(k-1)+\log\frac{k-d+\sqrt{2vs}}{d-\sqrt{2vs}}\right).\end{aligned}$$

Choosing $s=\log\log k$ and assuming $D\geq 2$ and $d>10\log\log k$ yield the results. Note that if $\mu_j=0$, it can change D by at most one and its effect can be ignored. Also when $d<10\log\log k$ we can use the naive bound of $\log k$, since $\log\frac{k-D}{D}<\log k$ for $D>1$. □

3.A.2 Proof of Lemma 36

Proof. We first assume $x > 100$ and prove the lemma.

$$\begin{align}
&\sum_{i=2}^{\infty} \mathrm{poi}(x,i)\log(i-\delta) \tag{3.7}\\
&\geq \sum_{i=x-x_0}^{x+x_0} \mathrm{poi}(x,i)\log(i-\delta) \notag\\
&= \mathrm{poi}(x,x)\log(x-\delta) + \sum_{a=1}^{x_0} \mathrm{poi}(x,x-a)\log(x-a-\delta) + \mathrm{poi}(x,x+a)\log(x+a-\delta) \notag\\
&\geq \mathrm{poi}(x,x)\log(x-\delta) + \sum_{a=1}^{x_0} \mathrm{poi}(x,x-a)\Big[\log(x-a-\delta) + \log(x+a-\delta)\Big] \notag\\
&= \sum_{a=0}^{x_0} \frac{\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta) + \log(x+a-\delta)\Big] \notag\\
&+ \sum_{a=1}^{x_0} \frac{\mathrm{poi}(x,x-a)-\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta) + \log(\ x+a-\delta)\Big] \tag{3.8}
\end{align}$$

By Lemma 46,

$$\begin{aligned}
\sum_{a=0}^{x_0} \mathrm{poi}(x,x-a) + \mathrm{poi}(x,x+a) &= \mathrm{poi}(x,x) + 1 - \Pr(\mathrm{POI}(x) > x+x_0) - \Pr(\mathrm{POI}(x) < x-x_0)\\
&\geq \frac{1}{e\sqrt{x}} + 1 - 2\cdot e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})},
\end{aligned}$$

Also we can lower bound the bracket in (3.8) as

$$\begin{aligned}
\log(x+a-\delta) + \log(x-a+\delta) &= \log\left((x-\delta)^2 - a^2\right)\\
&= \log(x^2 - 2x\delta + \delta^2 - a^2)\\
&= \log\left(x^2(1 - \frac{2\delta}{x} + \frac{\delta^2 - a^2}{x^2})\right)\\
&= 2\log x + \log(1 - \frac{2\delta}{x} + \frac{\delta^2 - a^2}{x^2})\\
&\geq 2\log x - \frac{4\delta}{x} - \frac{2(a^2-\delta^2)}{x^2}.
\end{aligned}$$

Thus for some constant c_1 and $x_0 = x^{0.8}$,

$$\begin{aligned}
&\sum_{a=0}^{x_0} \frac{\operatorname{poi}(x,x-a)+\operatorname{poi}(x,x+a)}{2}\Big[\log(x-a-\delta)+\log(x+a-\delta)\Big] \\
&\geq \Big(1+\frac{1}{e\sqrt{x}} - 2e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})}\Big)\Big(\log x - \frac{2\delta}{x}\Big) \\
&-\sum_{a=0}^{x_0}\Big(\operatorname{poi}(x,x-a)+\operatorname{poi}(x,x+a)\Big)\Big(\frac{a^2}{x^2}\Big) \\
&= \log x - \frac{2\delta}{x} - 2e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})}(\log x - \frac{2\delta}{x}) \\
&-\sum_{a=0}^{x_0}\Big(\operatorname{poi}(x,x-a)+\operatorname{poi}(x,x+a)\Big)\Big(\frac{a^2}{x^2}\Big) \\
&\geq \log x - \frac{c_1}{x}. \qquad (3.9)
\end{aligned}$$

where the last line is due to the following lemma.

Lemma 37. *For $x_0 = x^{0.8}$ there exists a constant c_1 such that*

$$\sum_{a=0}^{x_0}\Big[\operatorname{poi}(x,x-a)+\operatorname{poi}(x,x+a)\Big](\frac{a^2}{x^2}) \leq \frac{c_1}{x}.$$

The difference in probabilities of two equidistant points from the mean of a Poisson distribution is bounded by

$$\begin{aligned}
\operatorname{poi}(x,x+a)-\operatorname{poi}(x,x-a) &= \frac{e^{-x}x^{x-a}}{(x-a)!}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&= \frac{e^{-x}x^{x-a}e^{x-a}}{(x-a)^{x-a}\sqrt{2\pi(x-a)}}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&= \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&\approx \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\frac{4}{x},
\end{aligned}$$

and therefore for $x_0 = x^{0.8}$ and some constant c_5,

$$
\begin{aligned}
&\sum_{a=1}^{x_0} \frac{\mathrm{poi}(x,x-a)-\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta)+\log(x+a-\delta)\Big] \\
&\geq -\sum_{a=1}^{x_0} \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\frac{2}{x}\log\left((x-\delta)^2-a^2\right) \\
&\geq -\sum_{a=1}^{x_0} \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\frac{2}{x}\log x^2 \\
&\geq -\frac{\sum_{a=1}^{\infty} e^{-a}}{\sqrt{2\pi(x-x_0)}}\frac{4}{x}\log x \\
&\geq -\frac{4\log x}{x\sqrt{2\pi(x-x_0)}} \\
&\geq -\frac{c_5}{x}. \qquad (3.10)
\end{aligned}
$$

Selecting $c > c_1 + c_5$ leads to the Lemma. It can be shown that the lemma is valid for $x < 100$ by plotting the function. □

3.A.3 Proof of Lemma 37

Proof.

$$\begin{aligned}
&\sum_{a=0}^{x_0}\Big[\text{poi}(x,x-a)+\text{poi}(x,x+a)\Big](\frac{a^2}{x^2})\\
&\leq \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\text{poi}(x,x-a)\\
&= \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\frac{e^{-x}x^{x-a}}{(x-a)!}\\
&\overset{(a)}{\leq} \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-x+x-a}x^{x-a}}{(x-a)^{x-a}\sqrt{2\pi(x-a)}}\Big]\\
&= \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-a}}{\sqrt{2\pi(x-a)}}\left(1+\frac{a}{x-a}\right)^{x-a}\Big]\\
&= \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-a}}{\sqrt{2\pi(x-a)}}e^{(x-a)\ln(1+\frac{a}{x-a})}\Big]\\
&\overset{(b)}{\leq} \sum_{a=0}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big],
\end{aligned}$$

where (a) is by Stirling's approximation and (b) is because $\ln(1+x) < x-\frac{x^2}{4}$ for $x<1$. We can decompose the last summation to three different summations as

$$\begin{aligned}
&\sum_{a=0}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big]\\
&= \sum_{a=0}^{\sqrt{x}}\frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big]+\sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x}\frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big]+\sum_{a=\sqrt{x}\ln x}^{x_0}\frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] \qquad (3.11)
\end{aligned}$$

Now we bound each term in (3.11). For the first term and for some constant c_2:

$$\begin{aligned}
\sum_{a=0}^{\sqrt{x}} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] &\le 2\sqrt{x}\frac{x}{x^2}\frac{1}{\sqrt{2\pi(x-\sqrt{x})}} \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x}\frac{1}{\sqrt{1-\frac{\sqrt{x}}{x}}} \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x}(1+\frac{\sqrt{x}}{2x}) \le \frac{c_2}{x}.
\end{aligned}$$

Also for the middle term in (3.11) and some constant c_4:

$$\begin{aligned}
&\sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}}\sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x} a^2 e^{-\frac{a^2}{4x}} \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}}\int_{\sqrt{x}}^{\sqrt{x}\ln x} a^2 e^{-\frac{a^2}{4x}}da \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}}2\Big[x\sqrt{x}e^{-\frac{1}{4}} - x\sqrt{x}e^{-\frac{x\ln^2 x}{4x}}\ln x + 2\sqrt{\pi}\left(\mathrm{Erf}(\frac{\ln x}{2}) - \mathrm{Erf}(\frac{1}{2})\right)\Big] \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2\sqrt{x}}\frac{1}{\sqrt{1-\frac{\sqrt{x}\ln x}{x}}}(4x^{\frac{3}{2}}e^{-\frac{1}{4}}) \\
&\le \frac{c_4}{x}.
\end{aligned}$$

Similarly for the third term in (3.11) and for some constant c_3, we can write

$$\begin{aligned}
\sum_{a=\sqrt{x}\ln x}^{x_0} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] &\le (x_0-\sqrt{x}\ln x)\frac{2x_0^2}{x^2}\Big[\frac{e^{-\frac{(\sqrt{x}\ln x)^2}{4x}}}{\sqrt{2\pi(x-x_0)}}\Big] \\
&\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{x_0^3}{\sqrt{x-x_0}}e^{-\frac{\ln^2 x}{4}} \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{x_0^3}{\sqrt{x-x_0}}\frac{1}{x^{\frac{\ln x}{4}}} \le \frac{c_3}{x}.
\end{aligned}$$

Choosing $c_1 \geq c_2 + c_3 + c_4$ leads to the lemma. □

3.B Proofs of lower bound

In this part we provide the proofs of Theorem 30 as well as Corollaries 31 and 32. In order to lower bound the minimax risk of a given class $\mathcal{P}$, we can resort to two simplifications. First, we consider classes at a much a finer granularity than the $\mathcal{P}_d$ classes described in Section 3.5. In particular, let $\mathcal{P}_p$ be the *permutation class* of distributions consisting of a single distribution p and all of its permutations. Note that the multiset of probabilities is the same for all distributions in $\mathcal{P}_p$, and since the expected number of distinct categories only depends on the multiset ($d = \sum_j [1-(1-p_j)^n]$) it follows that $\mathcal{P}_p \subset \mathcal{P}_d$. [2]. To find a good lower bound for $\mathcal{P}_d$, we need a p that is "worst case" among all those who have the same value of d and then use the corresponding lower bound for $\mathcal{P}_p$. In what follows, we start by giving a lower bound for $\mathcal{P}_p$, and then specialize it for $\mathcal{P}_d$.

We also assume that an oracle specifies the *true* probability of all observed categories. With this side-information, the best estimator *has* to use the true probabilities for the observed categories. For the unobserved categories, it needs to redistribute all the missing mass (the total probability of unobserved categories). Since the multiset of probabilities is fixed and any permutation of the remaining categories is equally probable, by symmetry there is no advantage in favoring one over the other. Therefore the best oracle-aided estimator is uniquely specified: exact probabilities for seen categories and uniform redistribution of the missing mass (S_0) over the unobserved categories. This argument can be proven formally via the maximin trick: substitute the maximum with a mean against an arbitrary prior over p, at which point the optimal q is the posterior, and then optimize over priors. It then suffices to use the convexity of $p \log \frac{p}{q}$ with

[2] We abuse notation by distinguishing the classes by the letter used, while at the same time using the letters to denote actual quantities. From the context we understand that d is the expected number of distinct categories for p, at the given n.

respect to q.

3.B.1 Proof of Theorem 30

Proof. Without loss of generality assume that $p_1 \geq p_2 \geq p_3 \geq \ldots \geq p_k$. Let $\gamma > 1$, we have:

$$
\begin{aligned}
r_n(\mathcal{P}_p) = \min_q \max_{p \in \mathcal{P}_p} \mathbb{E}\left[\sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}\right] &\geq \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j}{\frac{\sum_{j=D+1}^{k} p_j}{k-D}}\right] \\
&= \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j(k-D)}{n\sum_{j=D+1}^{k} p_j} \,\middle|\, D < \gamma d\right] \Pr(D < \gamma d) \\
&+ \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j(k-D)}{\sum_{j=D+1}^{k} p_j} \,\middle|\, D \geq \gamma d\right] \Pr(D \geq \gamma d) \\
&\overset{(a)}{\geq} \left(1 - \frac{1}{\gamma}\right) \sum_{j=\gamma d}^{k} p_j \log \frac{p_j(k-\gamma d)}{\sum_{j=\gamma d}^{k} p_j}
\end{aligned}
$$

where (a) is by the following arguments: By Markov's inequality we have $\Pr(D \geq \gamma d) \leq \frac{1}{\gamma}$. Also, $\sum_{j=D+1}^{k} p_j \log \frac{np_j(k-D)}{n\sum_{j=D+1}^{k} p_j}$ is positive and decreasing in D (in the extreme case, when $D = k$ is zero). Therefore,

$$
r_n(\mathcal{P}_p) \geq \left(1 - \frac{1}{\gamma}\right)\left(\sum_{j=\gamma d}^{k} p_j\right) \log \frac{k-\gamma d}{\sum_{j=\gamma d}^{k} p_j} + \sum_{i=\gamma d} p_j \log p_j
$$

This completes the proof. For any specific classes of distributions, we can find a lower bound by calculating d, $\sum_{j=\gamma d}^{k} p_j$, and $\sum_{j=\gamma d}^{k} p_j \log p_j$ for some $\gamma > 1$. □

3.B.2 Proof of Corollary 31

Proof. To use Theorem 30, we first calculate d, $\sum_{j>L} p_j$, and $\sum_{j>L} p_j \log p_j$ and then let $L = \gamma d$ for $\gamma = 2$.

$$\begin{aligned}
\sum_{j=L+1}^{k} p_j &= \sum_{j=L+1}^{k} \frac{c}{j^{\alpha}} \\
&\overset{(a)}{\geq} \int_{L+1}^{k+1} \frac{c}{x^{\alpha}} dx \\
&= \frac{c}{\alpha - 1} \left[(L+1)^{1-\alpha} - (k+1)^{1-\alpha} \right],
\end{aligned}$$

where (a) is by integration bound for monotone series. Similarly, we can show:

$$\sum_{j=L+1}^{k} p_j \leq \int_{L}^{k} \frac{c}{x^{\alpha}} dx = \frac{c}{\alpha - 1} \left[L^{1-\alpha} - k^{1-\alpha} \right].$$

For the last summation in the lower bound of Theorem 30 we have:

$$\begin{aligned}
&\sum_{j=L+1}^{k} p_j \log p_j \\
&= \sum_{j=L+1}^{k} \frac{c}{j^{\alpha}} \log \frac{c}{j^{\alpha}} \\
&= c \sum_{j=L+1}^{k} \frac{1}{j^{\alpha}} \log \frac{1}{j^{\alpha}} + \log c \sum_{j=L+1}^{k} \frac{c}{j^{\alpha}} \\
&\overset{(a)}{\geq} c \int_{L+1}^{k+1} \frac{1}{j^{\alpha}} \log \frac{1}{j^{\alpha}} dj + \log c \sum_{j=L+1}^{k} p_j \\
&\overset{(b)}{=} \frac{c}{\alpha} \int_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} x^{-\frac{1}{\alpha}} \log x \, dx + \log c \sum_{j=L+1}^{k} p_j \\
&\geq \frac{c}{\alpha - 1} \left[x^{1-\frac{1}{\alpha}} \log x \right]_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} - \frac{c}{\alpha - 1} \int_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} x^{-\frac{1}{\alpha}} + \frac{c \log c}{\alpha - 1} \left[(L+1)^{1-\alpha} - (k+1)^{1-\alpha} \right]
\end{aligned}$$

Using Theorem 30, if $k > \max\{n, \left(\frac{10}{9}^{\frac{1}{\alpha-1}}\right) n^{\frac{1}{\alpha}}\}$ we have,

$$
\begin{aligned}
& r_n(\mathcal{P}) \\
& \geq \frac{c}{\alpha-1}(L+1)^{1-\alpha} \log \frac{k-2d}{\frac{c}{\alpha-1}(L+1)^{1-\alpha}} + (L+1)^{1-\alpha}\left[\frac{c}{\alpha-1}\log(L+1)^{-\alpha} - \frac{c\alpha}{(\alpha-1)^2} + \frac{c\log c}{\alpha-1}\right] \\
& \geq \frac{c}{10(\alpha-1)}(L+1)^{1-\alpha} \log \frac{k-2d}{(L+1)} + \frac{c}{\alpha-1}(L+1)^{1-\alpha}\left[\frac{1}{10}\log(\alpha-1) - \frac{\alpha}{\alpha-1}\right]
\end{aligned}
$$

where (a) is by integration bound for monotone series, and (b) is by change of variable $x = \frac{1}{j^\alpha}$. Using Equation (3.3) , choosing $L = 2d$,

$$
r_n(\mathcal{P}) \geq \frac{2^{1-\alpha}c}{\alpha-1} d^{1-\alpha} \log \frac{k-2d}{2d} + \frac{2^{1-\alpha}c}{\alpha-1} d^{1-\alpha}\left[\log(\alpha-1) - \frac{\alpha}{\alpha-1}\right],
$$

and since for power-law distributions, d grows proportionally to $n^{\frac{1}{\alpha}}$, we can write

$$
r_n(\mathcal{P}) \geq c_1 \frac{d}{n} \log \frac{k-2d}{2d}(1-o(1)),
$$

for some constants c_1 and c_2. To compare this with the upper bound in the proof of Theorem 29, note that we always have $\mathbb{E}[\Phi_1] \leq d$, but for power law distributions both expressions grow proportionally to $n^{\frac{1}{\alpha}}$ and furthermore $\mathbb{E}[\Phi_1]/d$ converges to a constant, $\frac{1}{\alpha}$. This shows that the upper and lower bounds for power-law distributions are tight in the first order term, when k is large. □

3.B.3 proof of Corollary 32

Proof. To use Theorem 30, we first calculate d, $\sum_{j>\gamma d} p_j$, and $\sum_{j>\gamma d} p_j \log p_j$. For the expected number of distinct categories,

$$d = \sum_{j=1}^{k'} 1 - e^{-np_j} = k'(1 - e^{-\frac{n}{k'}}) = \frac{1-e^{-\rho}}{\rho} n.$$

For the sum of probabilities of unobserved categories,

$$\sum_{j>\gamma d} p_j = \frac{k' - \gamma d}{k'} = 1 - \frac{\gamma n(1-e^{-\rho})}{\rho k'} = 1 - \gamma(1-e^{-\rho}),$$

and for the last summation in (3.3),

$$\sum_{j=\gamma d+1}^{k} p_j \log p_j = \frac{k' - \gamma d}{k'} \log(\frac{1}{k'}) = \left(1 - \gamma(1-e^{-\rho})\right) \log\left(\frac{\rho}{n}\right).$$

Therefore, by (3.3) we have

$$r_n(\mathcal{P}) \geq \left(1 - \frac{1}{\gamma}\right) \left(1 - \gamma(1-e^{-\rho})\right) \log \frac{k - \gamma d}{1 - \gamma(1-e^{-\rho})} + \left(1 - \gamma(1-e^{-\rho})\right) \log\left(\frac{\rho}{n}\right),$$

which can also be written as

$$r_n(\mathcal{P}) \geq \left(1 - \frac{1}{\gamma}\right) \frac{\rho\left(1 - \gamma(1-e^{-\rho})\right)}{1-e^{-\rho}} \frac{d}{n} \log \frac{k - \gamma d}{d} + \left(1 - \gamma(1-e^{-\rho})\right) \log \frac{1-e^{-\rho}}{(1 - \gamma(1-e^{-\rho}))} + \\ \frac{1}{\gamma}\left(1 - \gamma(1-e^{-\rho})\right) \log \frac{1-e^{-\rho}}{d}.$$

Choosing $\gamma = 1.2$ and having $k > n^5$, the corollary follows for $\rho \leq 1.75$. □

3.C Proofs of Good–Turing and absolute-discount relationship

3.C.1 Proof of Lemma 33

Proof. For notational simplicity we define $C(\mu) \overset{\text{def}}{=} \frac{c^{\frac{1}{\alpha}}\Gamma(\mu-\frac{1}{\alpha})}{\alpha\mu!}n^{\frac{1}{\alpha}}$. Using Lemma 43,

$$\begin{aligned}
\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_{\mu}]} &\overset{(a)}{\leq} \frac{C(\mu+1)+O(\mu^{-\frac{1}{2}})}{C(\mu)\left(1-O(\mu^{-1}n^{-\frac{1}{\alpha}})\right)-O(\mu^{-\frac{1}{2}})} \\
&\leq \frac{C(\mu+1)}{C(\mu)}\left(\frac{1+O(\mu^{-\frac{1}{2}}C^{-1}(\mu+1))}{1-O(\mu^{-1}n^{-\frac{1}{\alpha}})-O(\mu^{-\frac{1}{2}}C^{-1}(\mu))}\right) \\
&\leq \frac{C(\mu+1)}{C(\mu)}\left(1+O(\mu^{-\frac{1}{2}}C^{-1}(\mu+1))+O(\mu^{-1}n^{-\frac{1}{\alpha}})\right) \\
&\overset{(b)}{\leq} \frac{C(\mu+1)}{C(\mu)}\left(1+O(\mu^{-\frac{1}{2}+1+\frac{1}{\alpha}}n^{\frac{-1}{\alpha}})\right) \\
&\overset{(c)}{=} \frac{\mu-\frac{1}{\alpha}}{\mu+1}\left(1+O(n^{\frac{-3}{2(2\alpha+1)}})\right)
\end{aligned}$$

The inequality in (a) and (b) are by Lemma 43 and (c) is by the fact that $\mu < n^{\frac{1}{2\alpha+1}}$. □

3.C.2 Proof of Theorem 34

Recall that S_μ denotes the total probability of symbols appearing μ times, and let $\hat{S}_\mu$ be the probability assigned to those symbols by an estimator. Note that, given the samples, we may think of S and $\hat{S}$ as legitimate probability distributions on the set $\mu = 0, 1, \cdots, n$. In [OS15], it was shown that the competitive loss of an estimator over a class defined by a single distribution p and its permutations can be bounded by:

$$\varepsilon_n(\mathcal{P}_p, q) = r_n(p,q) - r_n(\mathcal{P}_p) \leq \mathbb{E}[\mathsf{KL}(S||\hat{S})].$$

This is well defined, since S and $\hat{S}$ only refer to the multiset probabilities, which stays invariant over all distributions in the class. Using this bound and the equivalence of Lemma 33, we can proceed with the proof. In the proof, we analyze the absolute-discount estimator with discount $\delta = \min\{\frac{\max\{\Phi_1,1\}}{D}, \delta_{\max}\}$.

Proof. We have:

$$
\begin{aligned}
&\mathsf{KL}(S||\hat{S}) \\
&= \sum_{\mu=0}^{\infty} S_\mu \log \frac{S_\mu}{\hat{S}_\mu} \\
&\overset{(a)}{\leq} \sum_{\mu=0}^{\infty} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} \\
&= \frac{(S_0 - \hat{S}_0)^2}{\hat{S}_0} + \sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} \\
&= \frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}} + \sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu + \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} \\
&\overset{(b)}{\leq} \frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}} + 2\sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} + 2\sum_{\mu=1}^{\mu_0} \frac{(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n})^2\Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}
\end{aligned}
\tag{3.12}
$$

where (a) is by Lemma 42 and (b) is by $(a+b)^2 \leq 2a^2 + 2b^2$. We choose $\mu_0 = n^{\frac{1}{2\alpha+1}}$ and show the proof for the case when $n^{\frac{1}{2\alpha+1}} \geq 20\log n$, namely, $\alpha \leq \frac{\log n}{2(\log\log n + \log 20)} - \frac{1}{2}$. For $\alpha > \frac{\log n}{2(\log\log n + \log 20)} - \frac{1}{2}$, the proof follows the same lines, but by a different choice of μ_0. Lemmas 38, 39, 40, and 41 bound each term in Equation (5.1) separately, and hence

$$\mathbb{E}[\mathsf{KL}(S||\hat{S})] = O\left(\frac{1}{n^{\frac{2\alpha-1}{2\alpha+1}}}\right). \qquad \square$$

Lemma 38. *For a power-law distribution with exponent* $\alpha > \alpha_0 > 1$*, and the choice of* $\delta = \min\{\frac{\max\{\Phi_1,1\}}{D}, \delta_{\max}\}$*,*

$$\mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\Big] = O\left(\frac{1}{n}\right).$$

Proof. To upper bound the first term of the KL loss in Equation (5.1), namely the loss of proposed estimator for the missing mass, let A be the event $(1-t)\mathbb{E}[\Phi_1] \le \Phi_1 \le (1+t)\mathbb{E}[\Phi_1]$ and $\frac{1-t}{1+t}\frac{\mathbb{E}[\Phi_1]}{d} \le \frac{\Phi_1}{D} \le \frac{1+t}{1-t}\frac{\mathbb{E}[\Phi_1]}{d}$ for some $0 < t < 1$,

$$\begin{aligned}
\mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\Big] &= \mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A\Big]\Pr(A) + \mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A^c\Big]\Pr(A^c) \\
&\overset{(a)}{\le} \mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A\Big]\Pr(A) + 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)n^2 \\
&\overset{(b)}{\le} \mathbb{E}\Big[\frac{2(S_0 - \frac{\mathbb{E}[\Phi_1]}{n})^2 + 2(\frac{\mathbb{E}[\Phi_1]}{n} - \frac{\Phi_1}{n})^2}{\frac{\Phi_1}{n}}\ \Big|\ A\Big]\Pr(A) + 4n^2\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right) \\
&\overset{(c)}{\le} \frac{\mathbb{E}\Big[2(S_0 - \frac{\mathbb{E}[\Phi_1]}{n})^2 + 2(\frac{\mathbb{E}[\Phi_1]}{n} - \frac{\Phi_1}{n})^2\ \Big|\ A\Big]\Pr(A)}{\frac{(1-t)\mathbb{E}[\Phi_1]}{n}} + 4n^2\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right) \\
&\overset{(d)}{\le} \frac{2\mathrm{Var}(S_0) + \frac{2}{n^2}\mathrm{Var}(\Phi_1)}{\frac{\mathbb{E}[\Phi_1]}{2n}} + o\left(\frac{1}{n}\right) \\
&\overset{(e)}{\le} \frac{\frac{4}{n^2}\mathbb{E}[\Phi_2] + \frac{2}{n^2}\mathbb{E}[\Phi_1]}{\frac{\mathbb{E}[\Phi_1]}{2n}} + o\left(\frac{1}{n}\right) \\
&= O\left(\frac{1}{n}\right),
\end{aligned}$$

where (b) is by choosing t such that $\frac{1+t}{1-t}\frac{1}{\alpha_0} < \delta_{\max}$ and therefore conditioned on A, $\delta = \frac{\Phi_1}{D}$. Also, (c) is by concentration of Φ_1 (see Lemma 44), (d) is by choosing $t = n^{-\frac{1}{4\alpha}}$, and (e) is because $\mathrm{Var}(\Phi_1) \le \mathbb{E}[\Phi_1]$ and $\mathrm{Var}(S_0) \le \frac{2}{n^2}\mathbb{E}[\Phi_2]$ (see Lemma 49). $\square$

Lemma 39. *For a power-law distribution with exponent* α *and choice of* $\mu_0 = O(n^{\frac{1}{2\alpha+1}})$*,*

$$\mathbb{E}\Big[\sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Big] = O\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right)$$

Proof. Using Lemma 33 and $(a+b)^2 \leq 2a^2 + 2b^2$, we bound the second term in (5.1):

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Big] \\
&\leq 2\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu O(n^{\frac{-3}{2(2\alpha+1)}}))^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] + 2\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right]
\end{aligned} \tag{3.13}$$

For the first term in right hand side of Equation (3.13),

$$\begin{aligned}
\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu O(n^{\frac{-3}{2(2\alpha+1)}}))^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] &\leq n^{-1-\frac{3}{2\alpha+1}} \sum_{\mu=1}^{\mu_0} \frac{(\mu+1)^2\left(\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\right)^2 \mathbb{E}[\Phi_\mu]}{\mu-\delta} \\
&\leq \frac{4}{1-\delta_{\max}} n^{\frac{1}{\alpha}-1-\frac{3}{2\alpha+1}} \sum_{\mu=1}^{\mu_0} \mu^{-\frac{1}{\alpha}} \\
&\leq \frac{4}{1-\delta_{\max}} n^{\frac{1}{\alpha}-1-\frac{3}{2\alpha+1}} \mu_0^{1-\frac{1}{\alpha}} = O\left(\frac{1}{n}\right),
\end{aligned}$$

where the last line is by choosing $\mu_0 = n^{\frac{1}{2\alpha+1}}$. For the second term in Equation 3.13, using $(a+b)^2 \leq 2a^2 + 2b^2$ we have,

$$\begin{aligned}
\left(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\right)^2 &= \left[S_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}] + \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}] - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\right]^2 \\
&\leq 2\left(S_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2 + 2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2,
\end{aligned}$$

and therefore:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right]\\
&=\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right)+\\
&\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu<\frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu<\frac{\mathbb{E}[\Phi_\mu]}{2}\right)\\
&\overset{(a)}{\leq}\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right)+n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)\\
&\leq\left(\sum_{\mu=1}^{\mu_0}\frac{\mathbb{E}[(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2]}{\frac{\mu-\delta}{2n}\mathbb{E}[\Phi_\mu]}\;\Big|\;\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right)\Pr\left(\Phi_\mu\geq\frac{\mathbb{E}[\Phi_\mu]}{2}\right)+n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)\\
&\overset{(b)}{\leq}\sum_{\mu=1}^{\mu_0}\frac{\mathbb{E}\left[2\left(S_\mu-\frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2+2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu-\frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2\right]}{\frac{\mu-\delta}{2n}\mathbb{E}[\Phi_\mu]}+n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)\\
&\leq\sum_{\mu=1}^{\mu_0}\frac{2\mathrm{Var}(S_\mu)+2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\right)^2\mathrm{Var}(\Phi_\mu)}{\frac{\mu-\delta_{\max}}{2n}\mathbb{E}[\Phi_\mu]}+n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)\\
&\overset{(d)}{\leq}\sum_{\mu=1}^{\mu_0}\frac{2\frac{(\mu+2)^2}{n^2}\mathbb{E}[\Phi_{\mu+2}]+2\frac{(\mu+1)^2}{n^2}\frac{\mathbb{E}^2[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}}{\frac{\mu-\delta_{\max}}{2n}\mathbb{E}[\Phi_\mu]}+n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)\\
&\overset{(e)}{\leq}\sum_{\mu=1}^{\mu_0}\frac{3}{1-\delta_{\max}}\frac{\mu}{n}+o\left(\frac{1}{n}\right)\\
&\overset{(f)}{\leq}\frac{3}{n(1-\delta_{\max})}\left(\frac{\mu_0^2}{2}+2\mu_0\right)+o\left(\frac{1}{n}\right)=O\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right).
\end{aligned}
$$

Note that (a) follows from Lemma 44, (b) from $(x+y)^2\leq 2x^2+2y^2$, and (c) from $\mathbb{E}[S_\mu]=$

$\frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]$ and $\delta < \delta_{\max}$. Also, (d) results from $\mathrm{Var}[\Phi_\mu] \le \mathbb{E}[\Phi_\mu]$ and $\mathrm{Var}[S_\mu] \le \frac{(\mu+2)^2}{n^2}\mathbb{E}[\Phi_{\mu+2}]$ (see Lemma 49), (e) is by Lemma 43, and (f) results from the choice of $\mu_0 = n^{\frac{1}{2\alpha+1}}$. □

Lemma 40. *For a power-law distribution with exponent* α *and the choice of* $\mu_0 = n^{\frac{1}{2\alpha+1}}$,

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] = O\left(\frac{n^{\frac{1}{2\alpha}}}{n}\right).$$

Proof.

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \le \frac{1}{n}\sum_{\mu=1}^{\mu_0} \frac{\mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu\right]}{\mu-\delta_{\max}}.$$

Similar to the proof of Lemma 38, let A be the event $(1-t)\mathbb{E}[\Phi_1] \le \Phi_1 \le (1+t)\mathbb{E}[\Phi_1]$ and $\frac{1-t}{1+t}\frac{\mathbb{E}[\Phi_1]}{d} \le \frac{\Phi_1}{D} \le \frac{1+t}{1-t}\frac{\mathbb{E}[\Phi_1]}{d}$ for some $0 < t < 1$. Thus,

$$\begin{aligned}\mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu\right] &= \mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu \,\middle|\, A\right]\Pr(A) + \mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu \,\middle|\, A^c\right]\Pr(A^c)\\ &\le t^2\Pr(A)\,\mathbb{E}\left[\Phi_\mu \,\middle|\, A\right] + \Pr(A^c)\,\mathbb{E}\left[\Phi_\mu \,\middle|\, A^c\right]\\ &\le n^{-\frac{1}{2\alpha}}\mathbb{E}\left[\Phi_\mu\right]\end{aligned}$$

where the last line is by choosing $t = n^{-\frac{1}{4\alpha}}$ and using Lemma 50. Hence, we have

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \le \frac{n^{-\frac{1}{2\alpha}}}{n}\sum_{\mu=1}^{\mu_0} \frac{\mathbb{E}\left[\Phi_\mu\right]}{\mu-\delta_{\max}} = O\left(\frac{n^{\frac{1}{2\alpha}}}{n}\right),$$

where the constant depends on $\delta_{\max}$ and therefore on α_0. □

Lemma 41. *For a power-law distribution with exponent* α*, and* $\mu_0 = n^{\frac{1-2\alpha}{2\alpha+1}}$*,*

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] = O\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right)$$

Proof. For the last part in Equation 5.1:

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \le \mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{2(S_\mu - \frac{\mu}{n}\Phi_\mu)^2 + 2(\frac{\delta\Phi_\mu}{n})^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right].$$

We bound both terms in the above expression separately. For the second term, we have:

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(\frac{\delta\Phi_\mu}{n})^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \le \frac{1}{n}\sum_{\mu=\mu_0+1}^{\infty} \frac{\mathbb{E}[\Phi_\mu]}{\mu-\delta_{\max}} \le \frac{2c^{\frac{1}{\alpha}}\Gamma\left(1-\frac{1}{\alpha}\right)}{\alpha n}\frac{1}{\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}} + \frac{2}{n\sqrt{\mu_0}} = O\left(n^{-\frac{2\alpha}{2\alpha+1}}\right),$$

and for the first part, we have:

$$\begin{aligned}
&\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} \\
&\overset{(a)}{\le} \sum_{\mu=\mu_0+1}^{\infty}\sum_{x} \mathbb{1}_x^{\mu}\frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu-1}{n}} \\
&\le 2\sum_{x:\, np_x \ge \frac{\mu_0}{2}}\ \sum_{\mu=\mu_0+1}^{\infty} \mathbb{1}_x^{\mu}\frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} + 2\sum_{x:\, np_x < \frac{\mu_0}{2}}\ \sum_{\mu=\mu_0+1}^{\infty} \mathbb{1}_x^{\mu}\frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} \\
&\le 2\sum_{x:\, np_x \ge \frac{\mu_0}{2}}\ \sum_{\mu=1}^{\infty} \mathbb{1}_x^{\mu}\frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} + 2\sum_{x:\, np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0}\mathbb{1}_x^{>\mu_0},
\end{aligned}$$

where (a) follows from $(\sum_{i=1}^n a_i)^2 \leq n(\sum_{i=1}^n a_i^2)$. Taking expectations of both sides:

$$
\begin{aligned}
\mathbb{E}\Bigg[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Bigg] &\overset{(a)}{\leq} 2\left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{1}{n}\mathbb{E}\Big[\frac{(np_x)^2 - 2\mu n p_x + \mu^2}{\mu}\Big] + 2\sum_{x\colon np_x<\frac{\mu_0}{2}} \frac{n}{\mu_0}\mathbb{E}[\mathbb{1}_x^{>\mu_0}] \\
&\overset{(b)}{\leq} 2\left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{3}{n} + 2\sum_{x\colon np_x<\frac{\mu_0}{2}} \frac{n}{\mu_0}\mathbb{E}[\mathbb{1}_x^{>\mu_0}] \\
&\overset{(c)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{6}{n} + 2\sum_{x\colon np_x<\frac{\mu_0}{2}} \frac{n}{\mu_0}\exp\left(\mu_0 - np_x - \mu_0\ln\left(\frac{\mu_0}{np_x}\right)\right) \\
&\overset{(d)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{6}{n} + 2\sum_{x\colon np_x<\frac{\mu_0}{2}} \frac{n}{\mu_0}\exp\left(\frac{np_x - \mu_0}{3}\right) \\
&\overset{(e)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{4}{n} + 2e^{-\frac{\mu_0}{6}}\left(\frac{n}{\mu_0}\right)^2 \\
&= O(n^{\frac{1-2\alpha}{2\alpha+1}}),
\end{aligned}
$$

where (a) is by bounding the number of elements with probability greater than $\mu_0/2n$, (b) follows from the fact that $\mathbb{E}[\frac{1}{\mu}]$ when μ is a Poisson distribution with mean λ, is bounded by $\frac{1}{\lambda} + \frac{3}{\lambda^2}$(note that $\mu = 0$ is excluded). Also, (c) follows from Lemma 46, (d) follows from $3(x-1-x\ln x) \leq 1-x$ for $x > 2$, and (e) is by convexity of the exponential term in p_x and the fact that a convex function is maximized at the boundaries. □

3.D Tools

This section provides a summary of tools used in the proofs throughout the chapter.

Lemma 42. *For two distributions p and q,*

$$\mathsf{KL}(p||q) \overset{\text{def}}{=} \sum_i p_i \log \frac{p_i}{q_i} \leq \sum_i \frac{(p_i - q_i)^2}{q_i}$$

Lemma 43. *For a power-law distribution with power $\alpha > \alpha_0 > 1$ and normalization factor c, for $\mu \geq 1$*

$$\mathbb{E}[\Phi_\mu] \leq \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha \mu!} n^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}} \leq \frac{c^{\frac{1}{\alpha}} \Gamma(1 - \frac{1}{\alpha})}{\mu\alpha} \left(\frac{n}{\mu}\right)^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}}.$$

Also, for $1 \leq \mu < n^{\frac{1}{\alpha+1}}$, and $k > n^{\frac{1}{\alpha-1}}$,

$$\mathbb{E}[\Phi_\mu] \geq \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha \mu!} n^{\frac{1}{\alpha}} - \frac{1}{\sqrt{2\pi\mu}}.$$

Proof. For the upper bound on the expected number of elements that appeared μ times:

$$\begin{aligned}
\mathbb{E}[\Phi_\mu] &= \mathbb{E}\left[\sum_{x=1}^{k} \mathbb{1}_x^\mu\right] \\
&= \sum_{x=1}^{k} e^{-np_x} \frac{(np_x)^\mu}{\mu!} \\
&= \sum_{x=1}^{k} e^{-\frac{nc}{x^\alpha}} \frac{\left(\frac{nc}{x^\alpha}\right)^\mu}{\mu!} \\
&\overset{(a)}{\leq} \int_1^k e^{-\frac{nc}{x^\alpha}} \frac{\left(\frac{nc}{x^\alpha}\right)^\mu}{\mu!} dx + \max_x \{ e^{-\frac{nc}{x^\alpha}} \frac{\left(\frac{nc}{x^\alpha}\right)^\mu}{\mu!} \} \\
&\overset{(b)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!} \int_{\frac{nc}{k^\alpha}}^{nc} e^{-y} y^{\mu - 1 - \frac{1}{\alpha}} dy + \max_x \{ e^{-\frac{nc}{x^\alpha}} \frac{\left(\frac{nc}{x^\alpha}\right)^\mu}{\mu!} \} \\
&\overset{(c)}{\leq} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!} \left[\Gamma\left(\mu - \frac{1}{\alpha}, \frac{nc}{k^\alpha}\right) - \Gamma\left(\mu - \frac{1}{\alpha}, nc\right) \right] + \frac{1}{\sqrt{2\pi\mu}} \\
&= \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha\mu!} n^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}}, \qquad (3.14)
\end{aligned}$$

where (a) is followed by the integration bound for a uni-modal series, (b) is by changing of variables $\frac{nc}{x^\alpha} = y$. Also (c) is by the definition of Gamma function and the fact that $e^{-t}t^\mu$ is maximized at $t = \mu$ followed by Stirling's approximation. By further simplifying the Gamma function term:

$$\begin{aligned}
\frac{\Gamma(\mu - \frac{1}{\alpha})}{\mu!} &= \frac{(\mu - 1 - \frac{1}{\alpha})(\mu - 2 - \frac{1}{\alpha})\dots(1 - \frac{1}{\alpha})\Gamma(1 - \frac{1}{\alpha})}{\mu!} \\
&= \frac{1}{\mu}\prod_{j=1}^{\mu-1}\left(1 - \frac{1}{j\alpha}\right)\Gamma\left(1 - \frac{1}{\alpha}\right) \\
&= \frac{1}{\mu}\exp\left(\sum_{j=1}^{\mu-1}\log\left(1 - \frac{1}{j\alpha}\right)\right)\Gamma\left(1 - \frac{1}{\alpha}\right) \\
&\overset{(a)}{\leq} \frac{1}{\mu}\exp\left(-\frac{1}{\alpha}\sum_{j=1}^{\mu-1}\frac{1}{j}\right)\Gamma\left(1 - \frac{1}{\alpha}\right) \\
&\overset{(b)}{\leq} \frac{1}{\mu}\mu^{-\frac{1}{\alpha}}\Gamma\left(1 - \frac{1}{\alpha}\right).
\end{aligned}$$

where (a) is by $\log(1 - x) \leq -x$ for $0 < x < 1$, and (b) is because $\sum_{j=1}^{t}\frac{1}{j} \geq \log(t + 1)$. Similarly for the lower bound we have:

$$
\begin{aligned}
\mathbb{E}[\Phi_\mu] &= \mathbb{E}\left[\sum_{x=1}^{k} \mathbb{1}_x^\mu\right] \\
&= \sum_{x=1}^{k} e^{-np_x}\frac{(np_x)^\mu}{\mu!} \\
&= \sum_{x=1}^{k} e^{-\frac{nc}{x^\alpha}}\frac{(\frac{nc}{x^\alpha})^\mu}{\mu!} \\
&\overset{(a)}{\geq} \int_1^k e^{-\frac{nc}{x^\alpha}}\frac{(\frac{nc}{x^\alpha})^\mu}{\mu!}dx - \max_x\{e^{-\frac{nc}{x^\alpha}}\frac{(\frac{nc}{x^\alpha})^\mu}{\mu!}\} \\
&\overset{(b)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!}\int_{\frac{nc}{k^\alpha}}^{nc} e^{-y}y^{\mu-1-\frac{1}{\alpha}}dy - \frac{1}{\sqrt{2\pi\mu}} \\
&\overset{(c)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!}\left[\Gamma\left(\mu-\frac{1}{\alpha}\right) - \gamma\left(\mu-\frac{1}{\alpha},\frac{nc}{k^\alpha}\right) - \Gamma\left(\mu-\frac{1}{\alpha},nc\right)\right] - \frac{1}{\sqrt{2\pi\mu}} \\
&\overset{(d)}{\geq} \frac{c^{\frac{1}{\alpha}}\Gamma\left(\mu-\frac{1}{\alpha}\right)}{\alpha\mu!}n^{\frac{1}{\alpha}}\left(1-O\left(\mu^{-1}n^{-\frac{1}{\alpha}}\right)\right) - \frac{1}{\sqrt{2\pi\mu}}, \qquad (3.15)
\end{aligned}
$$

By Lemma 48 we have $\gamma\left(\mu-\frac{1}{\alpha},\frac{nc}{k^{\,\alpha}}\right) \leq \frac{1}{\mu+1-\frac{1}{\alpha}}\left(1+(\mu-\frac{1}{\alpha})e^{-\frac{nc}{k^\alpha}}\right)\frac{\left(\frac{nc}{k^\alpha}\right)^{\mu-\frac{1}{\alpha}}}{\mu-\frac{1}{\alpha}}$ or $\gamma\left(\mu-\frac{1}{\alpha},\frac{nc}{k^{\,\alpha}}\right) = O(\mu^{-1}n^{-\frac{1}{\alpha}})$ when $k > n^{\frac{1}{\alpha-1}}$. Lemma 47 implies that $\Gamma\left(\mu-\frac{1}{\alpha},nc\right) \leq B(nc)^{\mu-\frac{1}{\alpha}}e^{-nc}$ for some constant B and for every $1 < \mu < n^{\frac{1}{\alpha+1}}$. This and the recursion $\Gamma(s+1,t) = s\Gamma(s,x) + x^s e^{-x}$ lead to $\Gamma\left(\mu-\frac{1}{\alpha},nc\right) = O(\frac{1}{n})$ for $1 \leq \mu < n^{\frac{1}{\alpha+1}}$ and therefore (d). Also, (a) is followed by the integration bound for a uni-modal series, (b) is by changing of variables $\frac{nc}{x^\alpha} = y$, and (c) is by the definition of Gamma function and the fact that $e^{-t}t^\mu$ is maximized at $t = \mu$ followed by Stirling's approximation. □

Lemma 44. *For a power-law distribution with power* α *and* $\mu < n^{\frac{1}{\alpha+1}}$,

$$
\Pr\left[\Phi_\mu < \frac{\mathbb{E}[\Phi_\mu]}{2}\right] \leq \exp\left(-\frac{1}{6\mu}\left(\frac{n}{\mu}\right)^{\frac{1}{\alpha}}\right)
$$

Proof. $\Phi_\mu = \sum_x \mathbb{1}_x^\mu$, and therefore is a sum of independent random variables $\mathbb{1}_x^\mu$. By Bernstein's inequality

$$\Pr\left[\left|\Phi_\mu - \mathbb{E}[\Phi_\mu]\right| > t\right] \le 2\exp\left(-\frac{t^2/2}{\mathrm{Var}(\Phi_\mu)+t/3}\right)$$

Substituting $\mathbb{E}[\Phi_\mu]$ from Lemma 43 and using $\mathrm{Var}(\Phi_\mu) \le \mathbb{E}[\Phi_\mu]$, for $t = \frac{\mathbb{E}[\Phi_\mu]}{2}$ we have the lemma. □

Lemma 45 ([OD12a]). *Let D be the number of distinct categories and $d = \mathbb{E}[D]$. Also let $v = \mathbb{E}[\Phi_1]$ be the expected number of categories that appeared once. Then,*

$$\Pr[D < d - \sqrt{2vs}] \le e^{-s}$$

Lemma 46. *Let $X \sim POI(x)$, then for $x_0 > 0$, $\Pr(X \ge x + x_0) \le e^{x_0 - (x+x_0)\ln(1+\frac{x_0}{x})}$, and Also $\Pr(X \le x - x_0) \le e^{x_0 - (x+x_0)\ln(1+\frac{x_0}{x})}$.*

Proof. Chernoff bound suggests that for every $t > 0$

$$\Pr(X \ge a) \le \frac{\mathbb{E}[e^{tX}]}{e^{t\cdot a}},$$

and similarly for every $t < 0$,

$$\Pr(X \le a) \le \frac{\mathbb{E}[e^{tX}]}{e^{t\cdot a}}.$$

Moment generating function, $\mathbb{E}[e^{tX}]$ for X distributed according to POI(x) is $e^{x(e^t-1)}$. Therefore,

$$\begin{aligned}\Pr(X \ge a) &\le \inf_{t>0} \frac{e^{x(e^t-1)}}{e^{t\cdot a}}\\ &= \inf_{t>0} e^{x(e^t-1)-t\cdot a}\\ &= e^{a-x-a\ln\frac{a}{x}}.\end{aligned}$$

Substituting a by $x + x_0$ leads to the lemma. □

Lemma 47 ([NP00])**.** *For $a > 1$, $B > 1$, and $x > \frac{B}{B-1}(a-1)$, we have*

$$x^{a-1}e^{-x} < |\Gamma(a,x)| < Bx^{a-1}e^{-x}.$$

Lemma 48 (Theorem 4.1 in [Neu13])**.** *For $a > 0$ and $x > 0$, we have*

$$\exp\left(-\frac{ax}{a+1}\right) \leq \frac{a}{x^a}\gamma(a,x) \leq \frac{1}{a+1}(1+ae^{-x}).$$

Lemma 49. *For every distribution, $\mu \geq 1$, and in the presence of Poisson sampling,*

$$\mathrm{Var}(\Phi_\mu) \leq \mathbb{E}[\Phi_\mu], \quad \mathrm{Var}(S_\mu) \leq \frac{(\mu+1)(\mu+2)}{n^2}\mathbb{E}[\Phi_{\mu+2}], \quad \mathbb{E}[S_\mu] = \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]$$

Proof. We use the property of the Poisson sampling that the counts are independent. For the variance of Φ_μ we can write:

$$\begin{aligned}\mathrm{Var}(\Phi_\mu) &= \mathrm{Var}\left(\sum_j \mathbb{1}_j^\mu\right)\\ &= \sum_j \mathrm{Var}(\mathbb{1}_j^\mu)\\ &\leq \sum_j \mathbb{E}[\mathbb{1}_j^\mu]\\ &= \mathbb{E}[\Phi_\mu].\end{aligned}$$

Also, for the expected value of the sum of probabilities that appeared μ times, we have:

$$\begin{aligned}
\mathbb{E}[S_\mu] &= \mathbb{E}\Big[\sum_j p_j \mathbb{1}_j^\mu\Big] \\
&= \sum_j p_j e^{-np_j} \frac{(np_j)^\mu}{\mu!} \\
&= \frac{\mu+1}{n} \sum_j e^{-np_j} \frac{(np_j)^{\mu+1}}{(\mu+1)!} \\
&= \frac{\mu+1}{n} \sum_j \mathbb{E}[\mathbb{1}_j^{\mu+1}] \\
&= \frac{\mu+1}{n} \mathbb{E}[\Phi_{\mu+1}],
\end{aligned}$$

and for their variance, we can write:

$$\begin{aligned}
\mathrm{Var}[S_\mu] &= \mathrm{Var}\Big[\sum_j p_j \mathbb{1}_j^\mu\Big] \\
&= \sum_j p_j^2 \mathrm{Var}(\mathbb{1}_j^\mu) \\
&= \sum_j p_j^2 \mathbb{E}[\mathbb{1}_j^\mu] \\
&= \sum_j p_j^2 e^{-np_j} \frac{(np_j)^\mu}{\mu!} \\
&= \frac{(\mu+1)(\mu+2)}{n^2} \sum_j e^{-np_j} \frac{(np_j)^{\mu+2}}{(\mu+2)!} \\
&= \frac{(\mu+1)(\mu+2)}{n^2} \sum_j \mathbb{E}[\mathbb{1}_j^{\mu+2}] \\
&= \frac{(\mu+1)(\mu+2)}{n} \mathbb{E}[\Phi_{\mu+2}].
\end{aligned}$$

□

Lemma 50. *Let Φ_1 be number of categories appeared once. Also let D be the number of distinct* *categories observed and* $d = \mathbb{E}[D]$*, then for* $0 < t < 1$*,*

$$\Pr\left(\left|\frac{\Phi_1}{D} - \frac{\mathbb{E}[\Phi_1]}{d}\right| > \frac{2t}{1-t}\right) \le 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$$

Proof. Using Lemma 44 we have

$$\begin{aligned}\Pr\left(\left|\frac{\Phi_1}{\mathbb{E}[\Phi_1]} - 1\right| > t\right) &= \Pr\left(\left|\Phi_1 - \mathbb{E}[\Phi_1]\right| > t\mathbb{E}[\Phi_1]\right)\\ &\le \exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)\end{aligned}$$

Similarly for number of distinct elements we have

$$\begin{aligned}\Pr\left(\left|\frac{D}{d} - 1\right| > t\right) &= \Pr\left(|D-d| > td\right)\\ &\overset{(a)}{\le} \Pr\left(|D-d| > t\mathbb{E}[\Phi_1]\right)\\ &\overset{(b)}{\le} \exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)\end{aligned}$$

where (a) is because $\mathbb{E}[\Phi_1] \le d$ and (b) is because $\mathrm{Var}(D) = \mathbb{E}[\Phi_1]$. Hence, with probability at least $1 - 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$ we have $1 - t \le \frac{\Phi_1}{\mathbb{E}[\Phi_1]} \le 1+t$ and $1 - t \le \frac{D}{d} \le 1+t$. With probability $\ge 1 - 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$, $\frac{1-t}{1+t} \le \frac{\Phi_1}{D}\frac{d}{\mathbb{E}[\Phi_1]} \le \frac{1+t}{1-t}$, namely $\left|\frac{\Phi_1}{D} - \frac{\mathbb{E}[\Phi_1]}{d}\right| \le \max\left(\frac{1+t}{1-t} - 1, 1 - \frac{1-t}{1+t}\right) = \frac{2t}{1-t}$. □

Chapter 4

Learning Low Rank Conditional Probability Matrices

4.1 Introduction

One of the fundamental tasks in statistical learning is probability estimation. When the possible outcomes can be divided into k discrete categories, e.g. types of words or bacterial species, the task of interest is to use data to estimate the probability masses $p_1, \cdots, p_k$, where p_j is the probability of observing category j. More often than not, it is not a single distribution that is to be estimated, but multiple *related* distributions, e.g. frequencies of words within various contexts or species in different samples. We can group these into a conditional probability (row-stochastic) matrix $P_{i,1}, \cdots, P_{i,k}$ as i varies over c contexts, and P_{ij} represents the probability of observing category j in context i. Learning these distributions individually would cause the data to be unnecessarily diluted. Instead, the *structure* of the relationship between the contexts should be harnessed.

A number of models have been proposed to address this structured learning task. One of the wildly successful approaches consists of positing that P, despite being a $c \times k$ matrix, is in fact of much lower rank m. Effectively, this means that there exists a latent context space of size $m \ll c, k$ into which the original context maps probabilistically via a $c \times m$ stochastic matrix A, then this latent context in turn determines the outcome via an $m \times k$ stochastic matrix B. Since this structural model means that P factorizes as $P = AB$, this problem falls within the framework of low-rank (non-negative) matrix factorization. Many topic models, such as the original work on probabilistic latent semantic analysis PLSA, also map to this framework. We narrow our attention here to such low-rank models, but note that more generally these efforts fall under the areas of structured and transfer learning. Other examples include: manifold learning, multi-task learning, and hierarchical models.

In natural language modeling, low-rank models are motivated by the inherent semantics of language: context first maps into meaning which then maps to a new word prediction. An alternative form of such latent structure, word embeddings derived from recurrent neural networks (or LSTMs) are the state-of-the-art of current language models, [MKB$^+$11, SPC14, WPM$^+$15]. A first chief motivation for the present work is to establish a theoretical underpinning of the success of such representations. We restrict the exposition to *bigram* models. The traditional definition of the bigram is that language is modeled as a sequence of words generated by a first order Markov-chain. Therefore the 'context' of a new word is simply its preceding word, and we have $c = k$. Since the focus here is not the dependencies induced by such memory, but rather the ramifications of the structural assumptions on P, we take bigrams to model word-pairs *independently* sampled by first choosing the contextual word with probability π and then choosing the second word according to the conditional probability P, thus resulting in a joint distribution over word-pairs $(\pi_i P_{ij})$.

What is the natural measure of performance for a probability matrix estimator? Since ultimately such estimators are used to accurately characterize the likelihood of test data, the measure of choice used in empirical studies is the perplexity, or alternatively its logarithm, the cross entropy. For data consisting of n word-pairs, if C_{ij} is the number of times pair (i, j)

appears, then the cross entropy of an estimator Q is $\frac{1}{n}\sum_{ij} C_{ij} \log \frac{1}{Q_{ij}}$. The population quantity that corresponds to this empirical performance measure is the (row-by-row weighted) KL-divergence $D(P\|Q) = \sum_{ij} \pi_i P_{ij} \log \frac{P_{ij}}{Q_{ij}}$.

Note that this is indeed the expectation of the cross entropy modulo the true entropy, an additive term that does not depend on Q. This is the natural notion of *risk* for the learning task, since we wish to infer the likelihood of future data, and our goal can now be more concretely stated as using the data to produce an estimator Q_n with a 'small' value of $D(P\|Q_n)$. The choice of KL-divergence introduces a peculiar but important problem: the necessity to handle small frequencies appropriately. In particular, using the empirical conditional probability is not viable, since a zero in Q implies infinite risk. This is the problem of *smoothing*, which has received a great amount of attention by the NLP community. Our second salient motivation for the present work is to propose principled methods of integrating well-established smoothing techniques, such as *add-*$\frac{1}{2}$ and *absolute discounting*, into the framework of structured probability matrix estimation.

Our contributions are as follows, we provide:

- A general framework for integrating smoothing and structured probability matrix estimation, as an alternating-minimization that converges to a stationary point of a penalized empirical risk.
- A sample complexity upper bound of $O(km \log^2(2n+k)/n)$ for the expected KL-risk, for the global minimizer of this penalized empirical risk.
- A lower bound that matches this upper bound up to the logarithmic term, showing near-optimality.

This chapter is organized as follows. Section 4.2 reviews related work. Section 4.3 states the problem and Section 4.4 highlights our main results. Section 4.5 proposes our central algorithm and Section 4.6 analyzes its idealized variant. Section 4.7 provides some experiments and Section 4.8 concludes.

4.2 Related Work

Latent variable models, and in particular non-negative matrix factorization and topic models, have been such an active area of research in the past two decades that the space here cannot possibly do justice to the many remarkable contributions. We list here some of the most relevant to place our work in context. We start by mentioning the seminal papers [Hof99, LS01] which proposed the alternating minimization algorithm that forms the basis of the current work. This has appeared in many forms in the literature, including the multiplicative updates [ZYO13]. Some of the earliest work is reviewed in [PTRV98]. These may be generally interpreted as discrete analogs to PCA (and even ICA) [BJ04].

An influential Bayesian generative topic model, the Latent Dirichlet Allocation, [BNJ03] is very closely related to what we propose. In fact, add-half smoothing effectively corresponds to a Dirichlet$(1/2)$ (Jeffreys) prior. Our exposition differs primarily in adopting a minimax sample complexity perspective which is often not found in the otherwise elegant Bayesian framework. Furthermore, exact Bayesian inference remains a challenge and a lot of effort has been expended lately toward simple iterative algorithms with provable guarantees, e.g. [AAJN13, AGMM15]. Besides, a rich array of efficient smoothing techniques exists for probability vector estimation [AJOS13b, KOPS15, OS15, VV15], of which one could directly avail in the methodology that is presented here.

A direction that is very related to ours was recently proposed in [HKKV16]. There, the primary goal is to recover the rows of A and B in ℓ_1-risk. This is done at the expense of additional separation conditions on these rows. This makes the performance measure not easily comparable to our context, though with the proper weighted combination it is easy to see that the implied ℓ_1-risk result on P is subsumed by our KL-risk result (via Pinsker's inequality), up to logarithmic factors, while the reverse isn't true. Furthermore, the framework of [HKKV16] is restricted to symmetric joint probability matrices, and uses an SVD-based algorithm that is difficult to scale beyond very small latent ranks m. Apart from this recent paper for the ℓ_1-risk, sample complexity bounds for related (not fully latent) models have been proposed for the KL-risk, e.g. [AWT91]. But these remain partial, and far from optimal. It is also worth noting that information geometry gives conditions under which KL-risk behaves close to ℓ_2-risk [BZ08], thus leading to a Frobenius-type risk in the matrix case.

Although the core optimization problem itself is not our focus, we note that despite being a non-convex problem, many instances of matrix factorization admit efficient solutions. Our own heuristic initialization method is evidence of this. Recent work, in the ℓ_2 context, shows that even simple gradient descent, appropriately initialized, could often provably converge to the global optimum [BKS15].

Concerning whether such low-rank models are appropriate for language modeling, there has been evidence that some of the abovementioned word embeddings [MKB$^+$11] can be interpreted as implicit matrix factorization [LG14]. Some of the traditional bigram smoothing techniques, such as the Kneser-Ney algorithm [KN95b, CG99], are also reminiscent of rank reduction [HOF11, PSDX13, HOF15].

4.3 Problem Statement

Data $\mathcal{D}_n$ consists of n pairs (X_s,Y_s), $s = 1,\cdots,n$, where X_s is a context and Y_s is the corresponding outcome. In the spirit of a bigram language model, we assume that the context and outcome spaces have the same cardinality, namely k. Thus (X_s,Y_s) takes values in $[k]^2$. We denote the count of pairs (i,j) by C_{ij}. As a shortcut, we also write the row-sums as $C_i = \sum_j C_{ij}$.

We assume the underlying generative model of the data to be i.i.d., where each pair is drawn by first sampling the context X_s according to a probability distribution $\pi = (\pi_i)$ over $[k]$ and then sampling Y_s conditionally on X_s according to a $k \times k$ conditional probability (stochastic) matrix $P = (P_{ij})$, a non-negative matrix where each row sums to 1. We also assume that P has non-negative rank m. We denote the set of all such matrices by $\mathcal{P}_m$. They can all be factorized (non-uniquely) as $P = AB$, where both A and B are stochastic matrices in turn, of size $k \times m$ and $m \times k$ respectively.

A conditional probability matrix estimator is an algorithm that maps the data into a stochastic matrix $Q_n(X_1, \cdots, X_n)$ that well-approximates P, in the absence of any knowledge about the underlying model. We generally drop the explicit notation showing dependence on the data, and use instead the implicit n-subscript notation. The performance, or how well any given stochastic matrix Q approximates P, is measured according to the KL-risk:

$$R(Q) = \sum_{ij} \pi_i P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{4.1}$$

Note that this corresponds to an expected loss, with the log-loss $L(Q, i, j) = \log P_{ij}/Q_{ij}$. Although we do seek out PAC-style (in-probability) bounds for $R(Q_n)$, in order to give a concise definition of optimality, we consider the average-case performance $\mathbb{E}[R(Q_n)]$. The expectation here is with respect to the data. Since the underlying model is completely unknown, we would like to do well against adversarial choices of π and P, and thus we are interested in a uniform upper bound of the form:

$$r(Q_n) = \max_{\pi, P \in \mathcal{P}_m} \mathbb{E}[R(Q_n)].$$

The optimal estimator, in the minimax sense, and the minimax risk of the class $\mathcal{P}_m$ are thus given by:

$$Q_n^\star = \arg\min_{Q_n} r(Q_n) = \arg\min_{Q_n} \max_{\pi, P \in \mathcal{P}_m} \mathbb{E}[R(Q_n)] \tag{4.2}$$

$$r^\star(\mathcal{P}_m) = \min_{Q_n} \max_{\pi, P \in \mathcal{P}_m} \mathbb{E}[R(Q_n)].$$

Explicitly obtaining minimax optimal estimators is a daunting task, and instead we would like to exhibit estimators that compare well.

Definition 51 (Optimality)**.** *If an estimator satisfies* $\mathbb{E}[R(Q_n)] \leq \varphi \cdot \mathbb{E}[R(Q_n^\star)],\ \forall \pi,$ *(called an* oracle inequality*), then if* φ *is a constant (of n, k, and m), we say that the estimator is (order)* optimal*. If* φ *is not constant, but its growth is negligible with respect to the decay of* $r^\star(\mathcal{P}_m)$ *with n or the growth of* $r^\star(\mathcal{P}_m)$ *with k or m, then we can call the estimator* near-optimal*. In particular, we reserve this terminology for a logarithmic gap in growth, that is an estimator is near-optimal if* $\log\varphi/\log r^\star(\mathcal{P}_m) \to 0$ *asymptotically in any of n, k, or m. Finally, if* φ *does not depend on P we have* strong *optimality, and* $r(Q_n) \leq \varphi \cdot r^\star(\mathcal{P}_m)$*. If* φ *does depend on P, we have* weak *optimality.*

As a proxy to the true risk (4.1), we define the empirical risk:

$$R_n(Q) = \frac{1}{n}\sum_{ij} C_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{4.3}$$

The conditional probability matrix that minimizes this empirical risk is the empirical conditional probability $\hat{P}_{n,ij} = C_{ij}/C_i$. Not only is $\hat{P}_{n,ij}$ not optimal, but since there always is a positive (even if slim) probability that some $C_{ij} = 0$ even if $P_{ij} \neq 0$, it follows that $\mathbb{E}[R_n(\hat{P}_n)] = \infty$. This shows the importance of smoothing. The simplest benchmark smoothing that we consider is *add-*$\frac{1}{2}$ smoothing $\hat{P}_{ij}^{\mathsf{Add}\text{-}\frac{1}{2}} = \big(C_{ij}+1/2\big)\,/\,(C_i+k/2)$, where we give an additional “phantom” half-sample to each word-pair, to avoid zeros. This simple method has optimal minimax performance when estimating probability vectors. However, in the present matrix case it is possible to show that this can be a factor of k/m away from optimal, which is significant (cf. Figure 4.1(a) in Section 4.7). Of course, since we have not used the low-rank structure of P, we may be tempted to “smooth by factoring”, by performing a low-rank approximation of $\hat{P}_n$. However, this will not eliminate the zero problem, since a whole column may be zero. These facts highlight the importance of principled smoothing. The problem is therefore to construct (possibly weakly) optimal or near-optimal smoothed estimators.

4.4 Main Results

In Section 4.5 we introduce the ADD-$\frac{1}{2}$-SMOOTHED LOW-RANK algorithm, which essentially consists of EM-style alternating minimizations, with the addition of smoothing at each stage. Here we state the main results. The first is a characterization of the implicit risk function that the algorithm targets.

Theorem 52 (Algorithm)**.** $Q^{\mathsf{Add}\text{-}\frac{1}{2}\text{-}\mathsf{LR}}$ *converges to a stationary point of the* penalized *empirical risk*

$$R_{\mathsf{n,penalized}}(W,H) = R_n(Q) + \frac{1}{2n}\sum_{i,\ell}\log\frac{1}{W_{i\ell}} + \frac{1}{2n}\sum_{\ell,j}\log\frac{1}{H_{\ell j}}, \quad \textit{where } Q = WH. \tag{4.4}$$

Conversely, any stationary point of (4.4) *is a stable point of* ADD-$\frac{1}{2}$-SMOOTHED LOW-RANK.

The proof of Theorem 52 follows closely that of [LS01]. We now consider the global minimum of this implicit risk, and give a sample complexity bound. By doing so, we intentionally decouple the algorithmic and statistical aspects of the problem and focus on the latter.

Theorem 53 (Sample Complexity)**.** *Let $Q_n \in \mathcal{P}_m$ achieve the global minimum of Equation 4.4. Then for all $P \in \mathcal{P}_m$ such that $P_{ij} > \frac{km}{n}$ $\forall i,j$ and $n > 3$,*

$$\mathbb{E}[R(Q_n)] \leq \bar{c}\frac{km}{n}\log^2(2n+k), \qquad \text{with } \bar{c} = 4{,}300.$$

We outline the proof in Section 4.6. The basic ingredients are: showing the problem is near-realizable, a quantization argument to describe the complexity of $\mathcal{P}_m$, and a PAC-style [Vap98] relative uniform convergence which uses a sub-Poisson concentration for the sums of log likelihood ratios and uniform variance and scale bounds. Finer analysis based on VC theory may be possible, but it would need to handle the challenge of the log-loss being possibly unbounded and negative. The following result shows that Theorem 53 gives weak near-optimality for n large, as it is tight up to the logarithmic factor.

Theorem 54 (Lower Bound)**.** *For $n > k$, the minimax rate of $\mathcal{P}_m$ satisfies:*

$$r^\star(\mathcal{P}_m) \geq \underline{c}\frac{km}{n}, \qquad \text{with } \underline{c} = 0.06.$$

This is based on the vector case lower bound and providing the oracle with additional information: instead of only (X_s, Y_s) it observes (X_s, Z_s, Y_s), where Z_s is sampled from X_s using A and Y_s is sampled from Z_s using B. This effectively allows the oracle to estimate A and B directly.

4.5 Algorithm

Our main algorithm is a direct modification of the classical alternating minimization algorithm for non-negative matrix factorization [Hof99, LS01]. This classical algorithm (with a slight variation) can be shown to essentially solve the following mathematical program:

$$Q^{\mathsf{NNMF}}(\Phi) = \arg\min_{Q=WH} \sum_i \sum_j \Phi_{ij} \log \frac{1}{Q_{ij}}.$$

The analysis is a simple extension of the original analysis of [Hof99, LS01]. By "essentially solves", we mean that each of the update steps can be identified as a coordinate descent, reducing the cost function and ultimately converging as $T \to \infty$ to a stationary (zero gradient) point of this function. Conversely, all stationary points of the function are stable points of the algorithm. In particular, since the problem is convex in W and H individually, but not jointly in both, the algorithm can be thought of as taking exact steps toward minimizing over W (as H is held fixed) and then minimizing over H (as W is held fixed), whence the *alternating-minimization* name.

Before we incorporate smoothing, note that there are two ingredients missing from this algorithm. First, the cost function is the sum of row-by-row KL-divergences, but each row is *not weighted*, as compared to Equation (4.1). If we think of Φ_{ij} as $\hat{P}_{ij} = C_{ij}/C_i$, then the natural weight of row i is π_i or its proxy C_i/n. For this, the algorithm can easily be patched. Similarly to the analysis of the original algorithm, one finds that this change essentially minimizes the *weighted* KL-risks of the empirical conditional probability matrix, or equivalently the empirical risk as defined in Equation (4.3):

$$Q^{\mathsf{LR}}(C) = \arg\min_{Q=WH} R_n(Q) = \arg\min_{Q=WH} \sum_i \frac{C_i}{n} \sum_j \frac{C_{ij}}{C_i} \log \frac{1}{Q_{ij}}.$$

Of course, this is nothing but the maximum likelihood estimator of P under the low-rank constraint. Just like the empirical conditional probability matrix, it suffers from lack of smoothing. For instance, if a whole column of C is zero, then so will be the corresponding column of $Q^{\mathsf{ERM}}(C)$. The first naive attempt at smoothing would be to add-$\frac{1}{2}$ to C and then apply the algorithm:

$$Q^{\mathsf{Naive\ Add}\text{-}\frac{1}{2}\text{-}\mathsf{LR}}(C) = Q^{\mathsf{LR}}(C + \tfrac{1}{2})$$

However, this would result in excessive smoothing, especially when m is small. The intuitive reason is this: in the extreme case of $m = 1$ all rows need to be combined, and thus instead of adding $\frac{1}{2}$ to each category, $Q^{\mathsf{Naiveadd}-\frac{1}{2}\mathsf{LR}}$ would add $k/2$, leading to the the uniform distribution overwhelming the original distribution. We may be tempted to mitigate this by adding instead $1/2k$, but this doesn't generalize well to other smoothing methods. A more principled approach should perform smoothing directly *inside* the factorization, and this is exactly what we propose here. Our main algorithm is:

Algorithm: ADD-$\frac{1}{2}$-SMOOTHED LOW-RANK

- Input: $k \times k$ matrix (C_{ij}); Initial W^0 and H^0; Number of iterations T
- Iterations: Start at $t = 1$, increment and repeat while $t \leq T$

  - For all $i \in [k], \ell \in [m]$, update $W^t_{i\ell} \leftarrow W^{t-1}_{i\ell} \sum_j \frac{C_{ij}}{(WH)^{t-1}_{ij}} H^{t-1}_{\ell j}$
  - For all $\ell \in [m], j \in [k]$, update $H^t_{\ell j} \leftarrow H^{t-1}_{\ell j} \sum_i \frac{C_{ij}}{(WH)^{t-1}_{ij}} W^{t-1}_{i\ell}$
  - Add-$1/2$ to each element of W^t and H^t, then normalize each row.

- Output: $Q^{\mathsf{Add}\text{-}\frac{1}{2}\text{-}\mathsf{LR}}(C) = W^T H^T$

The intuition here is that, prior to normalization, the updated W and H can be interpreted as *soft counts*. One way to see this is to sum each row i of (pre-normalized) W, which would give C_i. As for H, the sums of its (pre-normalized) columns reproduce the sums of the columns of C. Next, we are naturally led to ask: is $Q^{\mathsf{Add}\text{-}\frac{1}{2}\mathsf{LR}}(C)$ implicitly minimizing a risk, just as $Q^{\mathsf{LR}}(C)$ minimizes $R_n(Q)$? Theorem 52 shows that indeed $Q^{\mathsf{Add}\text{-}\frac{1}{2}\mathsf{LR}}(C)$ essentially minimizes a *penalized* empirical risk.

More interestingly, ADD-$\frac{1}{2}$-SMOOTHED LOW-RANK lends itself to a host of generalizations. In particular, an important smoothing technique, *absolute discounting*, is very well suited for heavy-tailed data such as natural language [CG99, OD12b, BBO17]. We can generalize it to fractional counts as follows. Let C_i indicate counts in traditional (vector) probability estimation, and let D be the total number of distinct observed categories, i.e. $D = \sum_i \mathbb{I}\{C_i \geq 1\}$. Let the number of *fractional* distinct categories d be defined as $\mathrm{d} = \sum_i C_i \mathbb{I}\{C_i < 1\}$. We have the following *soft absolute discounting* smoothing:

$$\hat{P}_i^{\mathsf{Soft\text{-}AD}}(C,\alpha) = \begin{cases} \frac{C_i - \alpha}{\sum C} & \text{if } C_i \geq 1, \\ \frac{1-\alpha}{\sum C} C_i + \frac{\alpha(D+\mathrm{d})}{(k-D-\mathrm{d})\sum C}(1 - C_i) & \text{if } C_i < 1. \end{cases}$$

This gives us the following patched algorithm, which we do not place under the lens of theory currently, but we strongly support it with our experimental results of Section 4.7.

Algorithm: ABSOLUTE-DISCOUNTING-SMOOTHED LOW-RANK

- Input: Specify $\alpha \in (0,1)$

- Iteration:
  - Add-$1/2$ to each element of W^t, then normalize.
  - Apply soft absolute discounting to $H^t_{\ell j} \leftarrow \hat{P}^{\textsf{Soft-AD}}_j(H^t_{\ell,\cdot},\alpha)$
- Output: $Q^{\textsf{AD-LR}}(C,\alpha) = W^T H^T$

4.6 Analysis

We now outline the proof of the sample complexity upper bound of Theorem 53. Thus for the remainder of this section we have:

$$Q_n(C) = \underset{Q=WH}{\arg\min} R_n(Q) + \frac{1}{2n}\sum_{i,\ell}\log\frac{1}{W_{i\ell}} + \frac{1}{2n}\sum_{\ell,j}\log\frac{1}{H_{\ell j}},$$

that is $Q_n \in \mathcal{P}_m$ achieves the global minimum of Equation 4.4. Since we have a penalized empirical risk minimization at hand, we can study it within the classical PAC-learning framework. However, rates of order $\frac{1}{n}$ are often associated withe the *realizable* case, where $R_n(Q_n)$ is exactly zero [Vap98]. The following Lemma shows that we are *near* the realizable regime.

Lemma 55 (Near-realizability)**.** *We have*

$$\mathbb{E}[R_n(Q_n)] \leq \frac{k}{n} + \frac{km}{n}\log(2n+k).$$

We characterize the complexity of the class $\mathcal{P}_m$ by *quantizing* probabilities, as follows. Given a positive integer L, define Δ_L to be the subset of the appropriate simplex Δ consisting of L-empirical distributions (or "types" in information theory): Δ_L consists exactly of those distributions p that can be written as $p_i = L_i/L$, where L_i are non-negative integers that sum to L.

Definition 56 (Quantization)**.** *Given a positive integer L, define the L-quantization operation as mapping a probability vector p to the closest (in ℓ_1-distance) element of* Δ_L*,* $\tilde{p} = \arg\min_{q\in\Delta_L} \|p -$

$q\|_1$. *For a matrix $P \in \mathcal{P}_m$, define an L-quantization for any given factorization choice $P = AB$ as $\tilde{P} = \tilde{A}\tilde{B}$, where each row of $\tilde{A}$ is the mL-quantization of the respective row of A and each row of $\tilde{B}$ is the kL-quantization of the respective row of A. Lastly, define $\mathcal{P}_{m,L}$ to be the set of all quantized probability matrices derived from $\mathcal{P}_m$.*

Lemma 57. *For a probability vector p over k elements, L-quantization satisfies $|p_i - \tilde{p}_i| \leq \frac{1}{L}$ for all i, and $\|p - \tilde{p}\|_1 \leq \frac{k}{L}$, and $|\Delta_L| \leq \left(\frac{(L+k-1)e}{k-1}\right)^{k-1}$.*

Proof. Since the distance between two consecutive quantization points in an L-quantization is $\frac{1}{L}$, each coordinate in the probability vector p can change by at most $\frac{1}{L}$ when quantized. For the same reason, the at most ℓ_1-distance between a probability vector and its quantized version is bounded by $\frac{k}{L}$. To bound the size of Δ_L, we should count the number of probability vectors with all coordinates in the form $\frac{L_i}{L}$, where L_is are non-negative integers that sum to L. The number of such vectors is $\binom{L+k-1}{k-1}$ and by Stirling's approximation we have $|\Delta_L| \leq \left(\frac{(L+k-1)e}{k-1}\right)^{k-1}$. □ □

Lemma 58. *The cardinality of $P_{m,L}$ is bounded by $|P_{m,L}| \leq (2Le+e)^{2km}$.*

Proof. Based on Lemma 57, total number of possible quantized vectors for each row of B is $\leq \left(\frac{(kL+k-1)e}{k-1}\right)^{k-1} \leq (2Le+e)^{k-1}$. Similarly, for each row of A, size of the possible quantized vectors is $\leq (2Le+e)^{m-1}$, hence the Lemma. □ □

Lemma 59 (De-quantization)**.** *For a conditional probability matrix $Q \in \mathcal{P}_m$, any $L-$quantization $\tilde{Q}$ satisfies $|Q_{ij} - \tilde{Q}_{ij}| \leq \frac{2}{L}$ for all i. Furthermore, if $Q > \varepsilon$ per entry and $L > \frac{4}{\varepsilon}$, then:*

$$|R(Q) - R(\tilde{Q})| \leq \frac{4}{L\varepsilon} \quad and \quad |R_n(Q) - R_n(\tilde{Q})| \leq \frac{4}{L\varepsilon}.$$

Proof. Recall the quantization of probability vectors given by Definition 56. Let $Q = AB$ be a factorization of Q, and let $\tilde{Q} = \tilde{A}\tilde{B}$ be a quantization, row-by-row as described in Definition 56. For rows of A we have $L_A = mL$ and for rows of B, we have $L_B = kL$. Now let's write

$\delta_{ij} \overset{\text{def}}{=} \tilde{A}_{ij} - A_{ij}$ and similarly, $\delta'_{ij} \overset{\text{def}}{=} \tilde{B}_{ij} - B_{ij}$. Based on the Lemma 57, we then have

$$\sum_{\ell} \delta_{i\ell} \leq \frac{1}{L}, \quad \sum_{j} \delta'_{\ell j} \leq \frac{1}{L},$$

and the difference of each coordinate in the original matrix and the quantized one is much less than $\frac{1}{L}$, namely

$$|\delta_{ij}| \leq \frac{1}{L}, \quad |\delta'_{ij}| \leq \frac{1}{L}.$$

Moreover, we have

$$\begin{aligned}
\tilde{Q}_{ij} &= \sum_{\ell=1}^{m} \tilde{A}_{i\ell} \tilde{B}_{\ell j} \\
&= \sum_{\ell=1}^{m} (A_{i\ell} + \delta_{i\ell})(B_{\ell j} + \delta'_{\ell j}) \\
&= \sum_{\ell=1}^{m} A_{i\ell} B_{\ell j} + \sum_{\ell} A_{i\ell} \delta'_{\ell j} + \sum_{\ell} \delta_{i\ell} \tilde{B}_{\ell j} \\
&\overset{(a)}{\leq} Q_{ij} + \sum_{\ell} A_{i\ell} \frac{1}{L} + \sum_{\ell} |\delta_{i\ell}| \\
&\overset{(b)}{\leq} Q_{ij} + \frac{1}{L} + \frac{1}{L}.
\end{aligned}$$

where (a) is because $\delta'_{\ell j} \leq \frac{1}{L}$ and $\tilde{B}_{\ell j} \leq 1$, (b) is by A being row stochastic and $\sum_{\ell} \delta_{i\ell} \leq \frac{1}{L}$. Since the same derivation holds also by swapping Q and $\tilde{Q}$, we have $|\tilde{Q}_{ij} - Q_{ij}| \leq \frac{2}{L}$. We can then write:

$$\begin{aligned}
R(Q) - R(\tilde{Q}) &= \sum_{ij} \pi_i P_{ij} \left[\log \frac{P_{ij}}{Q_{ij}} - \log \frac{P_{ij}}{\tilde{Q}_{ij}} \right] \\
&= \sum_{ij} \pi_i P_{ij} \log \frac{\tilde{Q}_{ij}}{Q_{ij}} \\
&\leq \sum_{ij} \pi_i P_{ij} \log \frac{Q_{ij} + \frac{2}{L}}{Q_{ij}} \\
&\overset{(a)}{\leq} \sum_{ij} \pi_i P_{ij} \frac{2}{LQ_{ij}} \\
&\overset{(b)}{\leq} \frac{2}{L\varepsilon}
\end{aligned}$$

where (a) is by $\log(1+x) \leq x$ and (b) is by $Q_{ij} \geq \varepsilon$. Using the same arguments, we can bound the difference in the other direction. Namely,

$$\begin{aligned}
R(\tilde{Q}) - R(Q) &= \sum_{ij} \pi_i P_{ij} \left[\log \frac{P_{ij}}{\tilde{Q}_{ij}} - \log \frac{P_{ij}}{Q_{ij}} \right] \\
&\overset{(a)}{\leq} \sum_{ij} \pi_i P_{ij} \log \frac{Q_{ij}}{Q_{ij} - \frac{2}{L}} \\
&\overset{(b)}{\leq} \sum_{ij} \pi_i P_{ij} \frac{\frac{2}{LQ_{ij}}}{1 - \frac{2}{LQ_{ij}}} \\
&\overset{(c)}{\leq} \sum_{ij} \pi_i P_{ij} \frac{4}{LQ_{ij}} \\
&\leq \frac{4}{L\varepsilon}.
\end{aligned}$$

Where (a) is by the fact that $\tilde{Q}_{ij} \geq Q_{ij} - \frac{2}{L}$, (b) follows from $\log \frac{1}{1-x} \leq \frac{x}{1-x}$, and (c) uses the fact that $\frac{x}{1-x} \leq 2x$ for $x \leq \frac{1}{2}$ and also $\frac{2}{LQ_{ij}} < \frac{1}{2}$, which holds if $L > \frac{4}{\varepsilon}$. Lastly, the proof for the empirical risk follows exactly the same lines. □ □

We now give a PAC-style relative uniform convergence bound on the empirical risk

[Vap98].

Theorem 60 (Relative uniform convergence)**.** *Assume lower-bounded $P > \delta$ and choose any $\tau > 0$. We then have the following uniform bound over all lower-bounded $\tilde{Q} > \varepsilon$ in $\mathcal{P}_{m,L}$ (Definition 56):*

$$\Pr\left\{ \sup_{\tilde{Q}\in\mathcal{P}_{m,L},\tilde{Q}>\varepsilon} \frac{R(\tilde{Q}) - R_n(\tilde{Q})}{\sqrt{R(\tilde{Q})}} > \tau \right\} \leq \mathrm{e}^{-\frac{n\tau^2}{20\log\frac{1}{\varepsilon}+3\tau\sqrt{\frac{1}{\delta}\log\frac{1}{\varepsilon}}}+2km\log(2Le+e)}. \tag{4.5}$$

The proof of this Theorem consists, for fixed $\tilde{Q}$, of showing a sub-Poisson concentration of the sum of the log likelihood ratios. This needs care, as a simple Bennett or Bernstein inequality is not enough, because we need to eventually self-normalize. A critical component is to relate the variance and scale of the concentration to the KL-risk and its square root, respectively. The theorem then follows from uniformly bounding the normalized variance and scale over $\mathcal{P}_{m,L}$ and a union bound.

To put the pieces together, first note that thanks to the fact that the optimum is also a stable point of the ADD-$\frac{1}{2}$-SMOOTHED LOW-RANK, the add-$\frac{1}{2}$ nature of the updates implies that all of the elements of Q_n are lower-bounded by $\frac{1}{2n+k}$. By Lemma 59 and a proper choice of L of the order of $(2n+k)^2$, the quantized version won't be much smaller. We can thus choose $\varepsilon = \frac{1}{2n+k}$ in Theorem 60 and use our assumption of $\delta = \frac{km}{n}\log(2n+k)$. Using Lemmas 55 and 59 to bound the contribution of the empirical risk, we can then integrate the probability bound of (4.5) similarly to the realizable case. This gives a bound on the expected risk of the quantized version of Q_n of order $\frac{km}{n}\log\frac{1}{\varepsilon}\log L$ or effectively $\frac{km}{n}\log^2(2n+k)$. We then complete the proof by de-quantizing using Lemma 59.

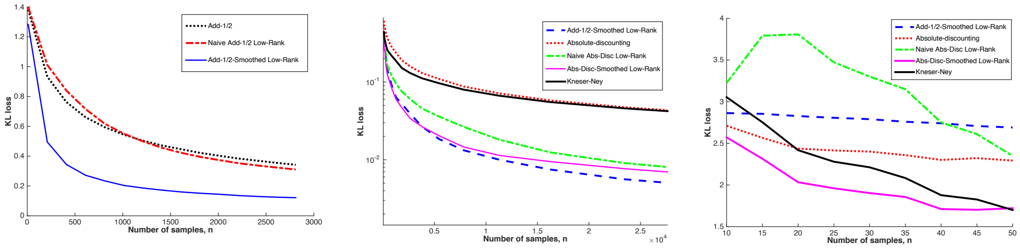
4.7 Experiments

Having expounded the theoretical merit of properly smoothing structured conditional probability matrices, we give a brief empirical study of its practical impact. We use both synthetic and real data. The various methods compared are as follows:

- Add-$\frac{1}{2}$, directly on the bigram counts: $\hat{P}_{n,ij}^{\mathsf{Add}\text{-}\frac{1}{2}} = (C_{ij} + \frac{1}{2})/(C_i + \frac{1}{2})$
- Absolute-discounting, directly on the bigram counts: $\hat{P}_n^{\mathsf{AD}}(C, \alpha)$ (see Section 4.5)
- Naive Add-$\frac{1}{2}$ Low-Rank, smoothing the counts then factorizing: $Q^{\mathsf{Naive\ Add}\text{-}\frac{1}{2}\text{-}\mathsf{LR}} = Q^{\mathsf{LR}}(C + \frac{1}{2})$
- Naive Absolute-Discounting Low-Rank: $Q^{\mathsf{Naive\ AD\text{-}LR}} = Q^{\mathsf{LR}}(n\hat{P}_n^{\mathsf{AD}}(C, \alpha))$
- Stupid backoff (SB) of Google, a very simple algorithm proposed in [BPX$^+$07]
- Kneser-Ney (KN), a widely successful algorithm proposed in [KN95b]
- Add-$\frac{1}{2}$-Smoothed Low-Rank, our proposed algorithm with provable guarantees: $Q^{\mathsf{Add}\text{-}\frac{1}{2}\text{-}\mathsf{LR}}$
- Absolute-Discounting-Smoothed Low-Rank, heuristic generalization of our algorithm: $Q^{\mathsf{AD\text{-}LR}}$

The synthetic model is determined randomly. π is uniformly sampled from the k-simplex. The matrix $P = AB$ is generated as follows. The rows of A are uniformly sampled from the k-simplex. The rows of B are generated in one of two ways: either sampled uniformly from the simplex or randomly permuted power law distributions, to imitate natural language. The discount parameter is then fixed to 0.75. Figure 4.1(a) uses uniformly sampled rows of B, and shows that, despite attempting to harness the low-rank structure of P, not only does Naive Add-$\frac{1}{2}$ fall short, but it may even perform worse than Add-$\frac{1}{2}$, which is oblivious to structure. Add-$\frac{1}{2}$-Smoothed Low-Rank, on the other hand, reaps the benefits of both smoothing *and* structure.

Figure 4.1(b) expands this setting to compare against other methods. Both of the proposed algorithms have an edge on all other methods. Note that Kneser-Ney is not expected to perform well in this regime (rows of B uniformly sampled), because uniformly sampled rows of B do not behave like natural language. On the other hand, for power law rows, even if $k \gg n$, Kneser-Ney does well, and it is only superseded by Absolute-Discounting-Smoothed Low-Rank. The consistent good performance of Absolute-Discounting-Smoothed Low-Rank may be explained by the fact that absolute-discounting seems to enjoy some of the competitive-optimality of Good-Turing estimation, as recently demonstrated by [OS15]. This is why we chose to illustrate the flexibility of our framework by heuristically using absolute-discounting as the smoothing component.



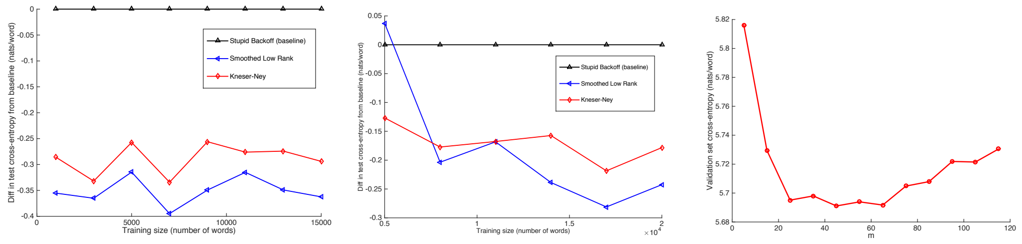
(a) $k = 100, m = 5$ (b) $k = 50, m = 3$ (c) $k = 1000, m = 10$

Figure 4.1: Performance of selected algorithms over synthetic data.

Before moving on to experiments on real data, we give a short description of the data sets. All but the first one are readily available through the Python NLTK:

- tartuffe, a French text, train and test size: 9.3k words, vocabulary size: 2.8k words.
- genesis, English version, train and test size: 19k words, vocabulary size: 4.4k words
- brown, shortened Brown corpus, train and test size: 20k words, vocabulary size: 10.5k words

For natural language, using absolute-discounting is imperative, and we restrict ourselves to Absolute-Discounting-Smoothed Low-Rank. The results of the performance of various algorithms are listed in Table 4.1. For all these experiments, $m = 50$ and 200 iterations were performed. Note that the proposed method has less cross-entropy per word across the board.



(a) Performance on tartuffe (b) Performance on genesis (c) rank selection for tartuffe

Figure 4.2: Experiments on real data.

Table 4.1: Cross-entropy results for different methods on several small corpora

Dataset	Add-$\frac{1}{2}$	AD	SB	KN	AD-LR
tartuffe	7.1808	6.268	6.0426	5.7555	**5.6923**
genesis	7.3039	6.041	5.9058	5.7341	**5.6673**
brown	8.847	7.9819	7.973	7.7001	**7.609**

We also illustrate the performance of different algorithms as the training size increases. Figure 4.2 shows the relative performance of selected algorithms with Stupid Backoff chosen as the baseline. As Figure 4.2(a) suggests, the amount of improvement in cross-entropy at $n = 15$k is around 0.1 nats/word. This improvement is comparable, even more significant, than that reported in the celebrated work of Chen and Goodman [CG99] for Kneser-Ney over the best algorithms at the time.

Even though our algorithm is given the rank m as a parameter, the internal dimension is not revealed, if ever known. Therefore, we could choose the best m using model selection. Figure 4.2(c) shows one way of doing this, by using a simple cross-validation for the tartuffe data set. In particular, half of the data was held out as a validation set, and for a range of different choices for m, the model was trained and its cross-entropy on the validation set was calculated. The figure shows that there exists a good choice of $m \ll k$. A similar behavior is observed for all data sets. Most interestingly, the ratio of the best m to the vocabulary size corpus is reminiscent of the choice of internal dimension in [MKB$^+$11].

4.8 Conclusion

Despite the theoretical impetus of our results, the resulting algorithms considerably improve over several benchmarks. There is more work ahead, however. Many possible theoretical refinements are in order, such as eliminating the logarithmic term in the sample complexity and dependence on P (strong optimality).

This framework naturally extends to tensors, such as for higher-order N-gram language models. It is also worth bringing back the Markov assumption and understanding how various mixing conditions influence the sample complexity. A more challenging extension, and one we suspect may be necessary to truly be competitive with RNNs/LSTMs, is to parallel this contribution in the context of generative models with long memory. The reason we hope to not only be competitive with, but in fact surpass, these models is that they do not use distributional properties of language, such as its quintessentially power-law nature. We expect smoothing methods such as absolute-discounting, which do account for this, to lead to considerable improvement.

4.9 Acknowledgment

Chapter 4, in full, is a reprint of the material as it appears in *Advances in Neural Information Processing Systems*. Falahatgar, Moein, Mesrob I. Ohannessian, and Alon Orlitsky, 2016. The dissertation author was one of the primary investigators and authors of this paper.

4.A Algorithms

The following is a slight variant of the original non-negagtive matrix factorization algorithm [LS01].

Algorithm: STOCHASTIC MATRIX FACTORIZATION

- Input:
  - Non-negative stochastic $k \times k$ matrix Φ
  - Initialization $k \times m$ matrix W^0 and $m \times k$ matrix H^0
  - Number of iterations T
- Iterations: Start at $t = 0$, increment and repeat while $t < T$

  - For all $i \in [k], \ell \in [m]$, update $W^t_{i\ell} \leftarrow W^{t-1}_{i\ell} \sum_j \frac{\Phi_{ij}}{(WH)^{t-1}_{ij}} H^{t-1}_{\ell j}$
  - For all $\ell \in [m], j \in [k]$, update $H^t_{\ell j} \leftarrow H^{t-1}_{\ell i} \sum_i \frac{\Phi_{ij}}{(WH)^{t-1}_{ij}} W^{t-1}_{i\ell}$
  - Normalize each row of H^t (W^t remains normalized)

- Output: $Q^{\mathsf{NNMF}}(\Phi) = (WH)^T$

We patch the algorithm as follows, to account for counts/weights.

Algorithm: Low-Rank ERM

- Input: Count $k \times k$ matrix (C_{ij}), instead of the stochastic matrix Φ
- Iterations:
  - For all $i \in [k], \ell \in [m]$, update $W^t_{i\ell} \leftarrow W^{t-1}_{i\ell} \sum_j \frac{C_{ij}}{(WH)^{t-1}_{ij}} H^{t-1}_{\ell j}$
  - For all $\ell \in [m], j \in [k]$, update $H^t_{\ell j} \leftarrow H^{t-1}_{\ell i} \sum_i \frac{C_{ij}}{(WH)^{t-1}_{ij}} W^{t-1}_{i\ell}$
  - Normalize each row of W^t and H^t
- Output: $Q^{\mathsf{LR}}(C) = (WH)^T$

4.A.1 Generalizations

The chapter's main algorithm, Add-$\frac{1}{2}$-Smoothed Low-Rank, lends itself to a host of generalization, which we do not place under the lens of theory currently, but which we illustrate in some of our experimental results in Section 4.7. Here we give more detailed about he *absolute discounting* smoothing outlined in the main text. Let C_i indicate counts in traditional (vector) probability estimation, and let D be the total number of distinct observed categories, i.e. $D = \sum_i \mathbb{I}\{C_i \geq 1\}$:

$$\hat{P}^{\mathsf{AD}}_n(C, \alpha) = \begin{cases} \frac{C_i - \alpha}{n} & \text{if } C_i \geq 1 \\ \frac{\alpha D}{n(k-D)} & \text{if } C_i = 0 \end{cases}$$

The discount parameter is either fixed or learned from data, via cross-validation or closed-form formulas that relate the discount to power-law type properties of the underlying distribution [CG99, OD12b]. When C represents a soft count, however, since it may have fractional values between 0 and α, we cannot outright subtract α. We ought to treat fractional counts in $[0,1]$ more like unseen symbols when close to 0 and more like seen symbols when close to 1. We suggest the following *soft absolute discounting* smoothing. Let the number of *fractional* distinct symbols Δ be defined as $\Delta = \sum_i C_i \mathbb{I}\{C_i < 1\}$:

$$\hat{P}_n^{\text{Soft-AD}}(C,\alpha) = \begin{cases} \frac{C_i - \alpha}{n} & \text{if } C_i \geq 1 \\ \frac{1-\alpha}{n} C_i + \frac{\alpha(D+\Delta)}{n(k-D-\Delta)}(1 - C_i) & \text{if } C_i < 1 \end{cases}$$

Note that for hard counts, when $C_i < 1$ implies $C_i = 0$, then $\Delta = 0$ and this reduces back to traditional absolute discounting. Otherwise, for fractional soft counts we interpolate between the behavior for a 1-count and that of a 0-count. It is easy to verify that this gives a valid probability distribution.

The suggested generalization of Add-$\frac{1}{2}$-Smoothed Low-Rank is to perform soft absolute discounting on each raw of H (or W or both, but if the goal is to capture power law behavior, only H is sufficient), instead of add-$\frac{1}{2}$. This gives us the following patched algorithm:

Algorithm: Absolute-Discounting-Smoothed Low-Rank

- Input: Specify $\alpha \in (0,1)$
- Iteration:
  - Add-$1/2$ to each element of W^t, then normalize.
  - Apply soft absolute discounting to $H^t_{\ell j} \leftarrow \hat{P}_n^{\text{Soft-AD}}(H^t_{\ell,\cdot},\alpha)$
- Output: $Q^{\text{AD-LR}}(C,\alpha) = (WH)^T$

4.A.2 Initialization

We have thus far stepped over the details of the initialization of the algorithm (the choice of W_0 and H_0). The characterization of Theorem 52 shows that we continue to be dealing with a non-convex optimization task. Multiple random restarts is always a viable option, but can be slow to latch onto the best region, especially for large values of k. Instead, we propose and implement a simple heuristic: identify the convex hull of the rows of the empirical conditional probability matrix $\hat{P}$. This is motivated by the fact the rows of B span the same subspace as those of P. Any convex hull algorithm could be used, but performance improves considerably if rows are incorporated from least to most noisy, by descending values of C_i. Such a smart initialization – in addition to a few random restarts – performs generally remarkably well, warranting further investigation.

4.B Proof of Theorem 52

The approach here parallels that of [LS01]. The relevant cost function, which we would like to show that we are implicitly descending, can be written as:

$$J(W,H) = \sum_{i,j} \frac{C_{ij}}{n} \log \frac{1}{(WH)_{ij}} + \frac{m}{2n} \sum_{i,\ell} \frac{1}{m} \log \frac{1}{W_{i\ell}} + \frac{k}{2n} \sum_{\ell,j} \frac{1}{k} \log \frac{1}{H_{\ell j}}.$$

Of course, the descent should be such that W and H are row-stochastic matrices. The simplex-projected gradients can be determined as:

$$\left.\frac{\partial J}{\partial W_{i\ell}}\right|_{\Delta} = -\sum_{j} \frac{C_{ij}}{n} \frac{H_{\ell j}}{(WH)_{ij}} - \frac{1}{2n} \frac{1}{W_{i\ell}} + \left(\frac{C_i}{n} + \frac{m}{2n} \right) \tag{4.6}$$

$$\left.\frac{\partial J}{\partial H_{\ell j}}\right|_{\Delta} = -\sum_{i} \frac{C_{ij}}{n} \frac{W_{i\ell}}{(WH)_{ij}} - \frac{1}{2n} \frac{1}{H_{\ell j}} + \left(\sum_{i,j} \frac{C_{ij}}{n} \frac{W_{i\ell} H_{\ell j}}{(WH)_{ij}} + \frac{k}{2n} \right) \tag{4.7}$$

These expressions can be readily obtained by computing the gradients of J and then for each row of W and H removing its component that is orthogonal to the simplex, i.e. parallel to the all-ones vector. The result are the terms in parenthesis, which effectively impose the constraints on W and H, by preventing each of their rows from leaving the simplex.

Next, we derive the multiplicative update rules for H and W and show that if $\nabla J = 0$, update rules will not change W and H. Let's denote $W_{i\ell}^{+}$ and $H_{\ell j}^{+}$ be the values of $W_{i\ell}$ and $H_{\ell j}$ after one update. Then,

$$W_{i\ell}^{+} = W_{i\ell} - \eta_i^W W_{i\ell}\left[-\sum_j \frac{C_{ij}}{n}\frac{H_{\ell j}}{(WH)_{ij}} - \frac{1}{2n}\frac{1}{W_{i\ell}} + \left(\frac{C_i}{n} + \frac{m}{2n}\right)\right],$$
$$H_{\ell j}^{+} = H_{\ell j} - \eta_\ell^H H_{\ell j}\left[-\sum_i \frac{C_{ij}}{n}\frac{W_{i\ell}}{(WH)_{ij}} - \frac{1}{2n}\frac{1}{H_{\ell j}} + \left(\sum_{i,j}\frac{C_{ij}}{n}\frac{W_{i\ell}H_{\ell j}}{(WH)_{ij}} + \frac{k}{2n}\right)\right].$$

Choosing

$$\eta_i^W = \frac{1}{\frac{C_i}{n} + \frac{m}{2n}},$$
$$\eta_\ell^H = \frac{1}{\sum_{i,j}\frac{C_{ij}}{n}\frac{W_{i\ell}H_{\ell j}}{(WH)_{ij}} + \frac{k}{2n}},$$

results to the following update rules,

$$W_{i\ell}^{+} = \frac{\sum_j C_{ij}\frac{H_{\ell j}W_{i\ell}}{(WH)_{ij}} + \frac{1}{2}}{C_i + \frac{m}{2}},$$
$$H_{\ell j}^{+} = \frac{\sum_i C_{ij}\frac{W_{i\ell}H_{\ell j}}{(WH)_{ij}} + \frac{1}{2}}{\sum_{i,j} C_{ij}\frac{W_{i\ell}H_{\ell j}}{(WH)_{ij}} + \frac{k}{2}}.$$

The correspondence between stationarity points of the function and the stable points of the algorithm are now evident. Moreover, each of these steps is an 'optimal' step, similar to a Newton step for a quadratic function: in the vector case, it would get to the solution in a single step. The fact that the cost function decreases at every step then follows from standard arguments, which completes the proof.

4.C Proof of Theorem 53

The quantization of Definition 56 and the de-quantization Lemma 59 are clearly a coarse characterization of the complexity of the class $\mathcal{P}_m$, since our choice of L depends on the probability lower bound ε and thus n and contributes a logarithmic factor in the exponent of the probability bounds and in the expectation bounds for the risk. A much finer characterization may be obtained via ℓ_1 covering numbers (see, for example, [SAJ04]). On the other hand, quantization gives us a finite class to work with, so that we can avail ourselves of a simple union bound. We first restate Theorem 60 more generally for any finite class, as follows.

Theorem 61 (Relative uniform convergence)**.** *Let* $\tau > 0$*. If* P *is lower bounded by* $\delta > 0$*, we then have the following relative uniform convergence of the empirical risk over lower-bounded* $Q > \varepsilon$ *in any finite class* $\mathcal{Q}$*:*

$$\Pr\left\{ \sup_{Q \in \mathcal{Q}, Q > \varepsilon} \frac{R(Q) - R_n(Q)}{\sqrt{R(Q)}} > \tau \right\} \leq \exp\left(-\frac{n\tau^2}{20 \log \frac{1}{\varepsilon} + 3\tau \sqrt{\frac{1}{\delta} \log \frac{1}{\varepsilon}}} + \log |\mathcal{Q}| \right).$$

Let us now prove this theorem. Let $Q > \varepsilon$ be any given lower-bounded conditional probability matrix. Let $\mathsf{KL} = \sum_{ij} \pi P_{ij} \log \frac{P_{ij}}{Q_{ij}}$ and define $b(x,y) = \frac{1}{\sqrt{\mathsf{KL}}} \log \frac{P_{xy}}{Q_{xy}}$. Let $B = b(X,Y)$ and $B_s = b(X_s, Y_s)$. Then the main quantity of interest is:

$$S_n = \mathbb{E}[B] - \frac{1}{n} \sum_{s=1}^{n} B_s$$

We have:

$$\Psi(\lambda) := \log \mathbb{E}\left[e^{\lambda S_n}\right] = \lambda \mathbb{E}[B] + n \log \mathbb{E}\left[e^{-\lambda B/n}\right] \tag{4.8}$$

$$\leq \lambda \mathbb{E}[B] + n\left(\mathbb{E}\left[e^{-\lambda B/n}\right] - 1\right) \tag{4.9}$$

$$= n\, \mathbb{E}\left[e^{-\lambda B/n} + \lambda B/n - 1\right] \tag{4.10}$$

$$= n \sum_{t\geq 2} \frac{(\lambda/n)^t}{t!} \mathbb{E}[(-B)^t]$$

Assume $\pi_i P_{ij} > \delta$. For $t \geq 2$, we have

$$\mathbb{E}[(-B)^t] = (-1)^t \sum_{ij} \pi_i P_{ij} \log^t \frac{P_{ij}}{Q_{ij}} \tag{4.11}$$

$$\leq \sum_{ij} \pi_i P_{ij} \left|\log \frac{P_{ij}}{Q_{ij}}\right|^t = \sum_{ij} \left[(\pi_i P_{ij})^{1/t} \left|\log \frac{P_{ij}}{Q_{ij}}\right|\right]^t \tag{4.12}$$

$$\leq \left[\sum_{ij} \left(\pi_i P_{ij}\right)^{2/t} \log^2 \frac{P_{ij}}{Q_{ij}}\right]^{t/2} \tag{4.13}$$

$$\leq \delta^{1-t/2} \left[\sum_{ij} \pi_i P_{ij} \log^2 \frac{P_{ij}}{Q_{ij}}\right]^{t/2} = \delta^{1-t/2} \mathbb{E}[B^2]^{t/2}$$

Going back, define $\phi(u) = e^u - u - 1$, $c = \frac{1}{n}\sqrt{\frac{1}{\delta} E[B^2]}$, and $v = \frac{1}{n}\mathbb{E}[B^2]$. We then have:

$$\Psi(\lambda) \leq n\delta \sum_{t\geq 2} \frac{(\lambda/n)^t}{t!} \left(\frac{1}{\delta}\mathbb{E}[B^2]\right)^{t/2} \tag{4.14}$$

$$= n\delta\phi\left(\frac{\lambda}{n}\sqrt{\frac{1}{\delta}\mathbb{E}[B^2]}\right) = \frac{v}{c^2}\phi(c\lambda).$$

This is of the Poisson form, see for example Theorem 2.9 of the concentration book, and implies in particular the following Bernstein-type inequality:

$$\Pr\{S_n > \tau\} \leq \exp\left[-\frac{\tau^2}{2(v + c\tau/3)}\right] \tag{4.15}$$

The following result bounds the second moment of the log-loss by its mean.

Lemma 62 (log-Loss Second Moment)**.** *Let*

$$v_n(Q) = \frac{1}{n} \frac{\sum_{i,j} \pi_i P_{ij} \log^2 \frac{P_{ij}}{Q_{ij}}}{\sum_{i,j} \pi_i P_{ij} \log \frac{P_{ij}}{Q_{ij}}}.$$

Then if $\varepsilon \leq \mathrm{e}^{-2}$ *we have that:*

$$\max_{Q:Q>\varepsilon} v_n(Q) \leq \frac{10}{n} \log \frac{1}{\varepsilon}.$$

The proof of this result may be found in the following section, along with the proofs of the near-realizability Lemma 55 and de-quantization Lemma 59.

Using the bound $E[B^2] < 10 \log \frac{1}{\varepsilon}$ of Lemma 62 in Equation (4.15), we then have:

$$\Pr\{S_n > \tau\} \leq \exp\left(- \frac{n\tau^2}{20 \log \frac{1}{\varepsilon} + 3\tau \sqrt{\frac{1}{\delta} \log \frac{1}{\varepsilon}}} \right)$$

Theorem 61 then follows by applying a union bound over all lower-bounded Q in $\mathcal{P}$. To obtain the stated form of Theorem 60, we simply apply to the class $\mathcal{P}_{m,L}$ given by Definition 56 which has cardinality at most $(2Le+e)^{2km}$, and use the bound $\bar{v}_n \leq \frac{10}{n} \log \frac{1}{\varepsilon}$ given by Lemma 62.

Recall that Q_n denotes the minimizer of the penalized risk of Equation (4.4). Let $\tilde{Q}_n$ be a quantization of Q_n. If Q_n is lower-bounded by ε, then by the de-quantization Lemma 59 $\tilde{Q}_n$ is lower-bounded by $\tilde{\varepsilon} = \varepsilon - \frac{2}{L}$. Let us assume that $L > \frac{4}{\varepsilon}$, and thus we have $\tilde{\varepsilon} > \varepsilon/2$.

Theorem 60 implies that for any $\tilde{Q}$ in $\mathcal{P}_{m,L}$, and in particular for $\tilde{Q}_n$ we have:

$$\Pr\left\{ \frac{R(\tilde{Q}_n) - R_n(\tilde{Q}_n)}{\sqrt{R(\tilde{Q}_n)}} \geq \tau \right\} \leq \exp\left(- \frac{n\tau^2}{20 \log \frac{1}{\varepsilon} + 3\tau \sqrt{\frac{1}{\delta} \log \frac{1}{\varepsilon}}} + 2km \log(2Le+e) \right). \quad (4.16)$$

We now would like to go from this probability bound to an expectation bound. We do so via the following integration lemma. For this, observe that we can always trivially substitute the bound of Equation (4.16) with 1, whenever it is greater than 1.

Lemma 63 (Integration)**.** *If X and Y are two random variables, X is non-negative, and for all $\tau > 0$*

$$\Pr\left\{\frac{X-Y}{\sqrt{X}} \geq \tau\right\} \leq \min\left\{\mathrm{e}^{-\frac{\tau^2}{2v+c\tau}+H}, 1\right\}$$

then for $H \geq 5$

$$\mathbb{E}[X] \leq 4\mathbb{E}[Y] + 24vH + 6c^2H^2.$$

This result is a generalization of the realizable-case integration, where Y is identically 0, and one recovers faster rates. By applying Lemma 63 to Equation (4.16), we get:

$$\mathbb{E}[R(\tilde{Q}_n)] \leq 4\mathbb{E}[R_n(\tilde{Q}_n)] + 24vH + 6c^2H^2,$$

with $H = 2km\log(2Le+e)$, $v = \frac{10}{n}\log\frac{2}{\varepsilon}$, and $c = \frac{3}{n}\sqrt{\frac{1}{\delta}\log\frac{2}{\varepsilon}}$.

Next, continuing to assume that $L > \frac{4}{\varepsilon}$, we can apply the de-quantization Lemma 59 twice to bound:

$$\begin{aligned}
\mathbb{E}[R(Q_n)] &\leq \mathbb{E}[R(\tilde{Q}_n)] + \frac{4}{\varepsilon L} && (4.17)\\
&\leq 4\mathbb{E}[R_n(\tilde{Q}_n)] + 24vH + 6c^2H^2 + \frac{4}{\varepsilon L} && (4.18)\\
&\leq 4\left(\mathbb{E}[R_n(Q_n)] + \frac{4}{\varepsilon L}\right) + \frac{4}{\varepsilon L} + 24vH + 6c^2H^2 && (4.19)\\
&= 4\mathbb{E}[R_n(Q_n)] + \frac{20}{\varepsilon L} + 24vH + 6c^2H^2
\end{aligned}$$

Using the near-realizability Lemma, we know that for any $\alpha \in (0,1)$

$$\mathbb{E}[R_n(Q_n)] \leq 2k\alpha + \frac{km}{n}\log\left(\frac{1+k\alpha}{\alpha}\right).$$

In particular, choose $\alpha = 1/2n$:

$$4\mathbb{E}[R_n(Q_n)] \leq 4\frac{k}{n} + 4\frac{km}{n}\log(2n+k).$$

Since the optimal solution Q_n is a stable point of the algorithm by Theorem 52, we can deduce that we can choose $\varepsilon = 1/(2n+k)$ as a lower bound. Also, since $k \geq 2$ then for $n \geq 3$ we have that $\varepsilon < \mathrm{e}^{-2}$ and the choice $L = (2n+k)^2 - 1$ handily satisfies $L > \frac{4}{\varepsilon}$ and $H \geq 5$, thus all the assumptions of scale that we have made. Also we can simplify $H < 8km\log(4n+2k)$, and using $\delta > \frac{km}{n}$ we get:

$$24vH \leq 3840\frac{km}{n}\log^2(4n+2k),$$

$$6c^2H^2 \leq 432\frac{km}{n}\log^2(4n+2k),$$

and

$$\frac{20}{\varepsilon L} \leq \frac{11}{n}.$$

Generously combining all these bounds, we obtain:

$$\mathbb{E}[R(Q_n)] \leq 4291\frac{km}{n}\log^2(4n+2k).$$

4.D Proofs of Lemmas

Proof. **(Lemma 55, Near-realizability).** Define

$$f(C) \overset{\text{def}}{=} \min_{W,H}\left(\sum_{i,j}\frac{C_{ij}}{n}\log\frac{1}{(WH)_{ij}} + \frac{1}{2n}\sum_{i,\ell}\log\frac{1}{W_{i\ell}} + \frac{1}{2n}\sum_{\ell,j}\log\frac{1}{H_{\ell j}}\right)$$

and let $Q^A(C) = W^A(C) * H^A(C)$ be the minimizer of $f(C)$. Here we show that the expected empirical risk for the estimator Q^A is small, namely,

$$\begin{aligned} \mathbb{E}[R_n(Q^A)] &= \mathbb{E}\Big[\sum_{i,j} \frac{C_{ij}}{n} \log \frac{P_{ij}}{Q^A_{ij}(C)}\Big] \\ &= \sum_{i,j} \pi_i P_{ij} \log P_{ij} + \mathbb{E}\Big[\sum_{i,j} \frac{C_{ij}}{n} \log \frac{1}{Q^A_{ij}(C)}\Big] \end{aligned} \tag{4.20}$$

For notational simplicity we drop the dependence of Q^A, W^A, and H^A on C. Since $f(C)$ is a point-wise minimum of linear functions, it is concave and therefore

$$\begin{aligned} \mathbb{E}[f(C)] &= \mathbb{E}\left[\sum_{i,j} \frac{C_{ij}}{n} \log \frac{1}{Q^A_{ij}} + \frac{1}{2n}\sum_{i,\ell} \log \frac{1}{W^A_{i\ell}} + \frac{1}{2n}\sum_{\ell,j} \log \frac{1}{H^A_{\ell j}}\right] \\ &\leq f(\mathbb{E}[C]) \\ &= f(n\pi P) \\ &= \min_{W,H} \sum_{i,j} \pi_i P_{ij} \log \frac{1}{(WH)_{ij}} + \frac{1}{2n}\sum_{i,\ell} \log \frac{1}{W_{i\ell}} + \frac{1}{2n}\sum_{\ell,j} \log \frac{1}{H_{\ell j}}. \\ &\leq \sum_{ij} \pi_i P_{ij} \log \frac{1}{(\tilde{A}\tilde{B})_{ij}} + \frac{1}{2n}\sum_{i,\ell} \log \frac{1}{\tilde{A}_{i\ell}} + \frac{1}{2n}\sum_{\ell,j} \log \frac{1}{\tilde{B}_{\ell j}}. \end{aligned}$$

Let $\tilde{A}_{i\ell} \overset{\text{def}}{=} \frac{A_{i\ell}+\delta}{1+m\delta}$ and $\tilde{B}_{\ell j} \overset{\text{def}}{=} \frac{B_{\ell j}+\delta}{1+k\delta}$, continuing from (4.20),

$$\mathbb{E}[R_n(Q^A)] \leq \sum_{i,j} \pi_i P_{ij} \log \frac{P_{ij}}{\sum_l \left(\frac{A_{i\ell}+\delta}{1+m\delta}\right)\left(\frac{B_{\ell j}+\delta}{1+k\delta}\right)} + \frac{1}{2n}\sum_{i,\ell} \log \frac{1}{\tilde{A}_{i\ell}} + \frac{1}{2n}\sum_{\ell,j} \log \frac{1}{\tilde{B}_{\ell j}}. \tag{4.21}$$

For the first expression in the right hand side (4.21), we have

$$\begin{aligned}
\sum_{i,j}\pi_i P_{ij}\log\frac{P_{ij}}{\sum_\ell\left(\frac{A_{i\ell}+\delta}{1+m\delta}\right)\left(\frac{B_{\ell j}+\delta}{1+k\delta}\right)} &= \sum_{i,j}\pi_i P_{ij}\log\frac{P_{ij}(1+m\delta)(1+k\delta)}{\sum_\ell A_{i\ell}B_{\ell j}+\sum_\ell\delta[A_{i\ell}+B_{\ell j}]+\sum_\ell\delta^2} \\
&\overset{(a)}{\leq} \sum_{i,j}\pi_i P_{ij}\log\frac{P_{ij}(1+m\delta)(1+k\delta)}{P_{ij}+\delta+\delta P_{ij}+m\delta^2} \\
&= \sum_{i,j}\pi_i P_{ij}\log\left(\frac{(1+m\delta)(1+k\delta)}{1+\delta}\left(\frac{P_{ij}}{P_{ij}+\frac{\delta+m\delta^2}{1+\delta}}\right)\right) \\
&\leq \sum_{i,j}\pi_i P_{ij}\Big(\log(1+m\delta)+\log(1+k\delta)\Big) \\
&\overset{(b)}{\leq} (m+k)\delta \\
&\leq 2k\delta.
\end{aligned}$$

Where (a) is because $\sum_\ell A_{i\ell}=1$ and $P_{ij}=\sum_\ell A_{i\ell}B_{\ell j}\leq\sum_\ell B_{\ell j}$. Also (b) follows from $\log(1+x)\leq x$.

For the second and the third terms in (4.21), we have

$$\begin{aligned}
\frac{1}{2n}\sum_{i,\ell}\log\frac{1}{\tilde{A}_{i\ell}}+\frac{1}{2n}\sum_{\ell,j}\log\frac{1}{\tilde{B}_{\ell j}} &\leq \frac{km}{2n}\log\frac{1+m\delta}{\delta}+\frac{km}{2n}\log\frac{1+k\delta}{\delta} \\
&\leq \frac{km}{n}\log\frac{1+k\delta}{\delta},
\end{aligned}$$

and therefore

$$\mathbb{E}[R_n(Q^A)]\leq 2k\delta+\frac{km}{n}\log\frac{1+k\delta}{\delta}.$$

Choosing $\delta=\frac{1}{2n}$ leads to the lemma. □ □

Proof. **(Lemma 62,** log**-Loss Second Moment).** For clarity, we write out the proof for the vector case, i.e. where p and q are single distributions. In particular, we show that for all k, p and

$q > \varepsilon$ with $\varepsilon \leq \mathrm{e}^{-2}$, we have:

$$\frac{\sum_{i=1}^{k} p_j \log^2 \frac{p_j}{q_j}}{\sum_{i=1}^{k} p_j \log \frac{p_j}{q_j}} \leq 10 \log \frac{1}{\varepsilon}.$$

The claim then holds by identifying k with k^2, p with πP and q with πQ.

We show that for each term we have the following inequality:

$$p_j \log \frac{p_j}{q_j} \geq (p_j - q_j) + \frac{1}{10} \frac{p_j \log^2 \frac{p_j}{q_j}}{\log \frac{1}{\varepsilon}}, \tag{4.22}$$

and therefore we can write $\sum_j p_j \log \frac{p_j}{q_j} \geq \sum_j (p_j - q_j) + \frac{1}{10} \frac{p_j \log^2 \frac{p_j}{q_j}}{\log \frac{1}{\varepsilon}} = \frac{1}{10 \log \frac{1}{\varepsilon}} \sum_j p_j \log^2 \frac{p_j}{q_j}$. To prove (4.22), we consider different cases and prove for each case separately.

If $p_j > q_j$ and $p_j - q_j \leq \frac{p_j}{2}$, then

$$\begin{aligned} p_j - q_j + \frac{p_j \log^2 \frac{p_j}{q_j}}{10 \log \frac{1}{\varepsilon}} &\overset{(a)}{\leq} (p_j - q_j) + \frac{2}{10 \log \frac{1}{\varepsilon}} \frac{(p_j - q_j)^2}{p_j} \\ &\overset{(b)}{\leq} (p_j - q_j) + \frac{(p_j - q_j)^2}{2 p_j} + \sum_{j=3} \frac{(p_j - q_j)^j}{j p_j^{j-1}} \\ &\overset{(c)}{=} p_j \log \frac{p_j}{q_j}, \end{aligned}$$

where (a) follows from $\log^2 \frac{p_j}{q_j} = \log^2 \frac{1}{1 - \frac{p_j - q_j}{p_j}}$ and $\log^2(1-x) \leq 2x^2$ for $x \leq 0.5$. Also (b) follows from $5 \log \frac{1}{\varepsilon} \geq 2$ and (c) from Taylor expansion of $\log \frac{q_j}{p_j}$ around 1.

Similarly if $q_j < p_j$ and $q_j - p_j \leq \frac{p_j}{2}$,

$$\begin{aligned} p_j - q_j + \frac{p_j \log^2 \frac{p_j}{q_j}}{10 \log \frac{1}{\varepsilon}} &\overset{(a)}{\leq} (p_j - q_j) + \frac{1}{10 \log \frac{1}{\varepsilon}} \frac{(q_j - p_j)^2}{p_j} \\ &\overset{(b)}{\leq} (p_j - q_j) + \frac{(q_j - p_j)^2}{2 p_j} + \sum_{j=3} \frac{(p_j - q_j)^j}{j p_j^{j-1}} \\ &\overset{(c)}{=} p_j \log \frac{p_j}{q_j}, \end{aligned}$$

where (a) follows from $\log^2 \frac{p_j}{q_j} = \log^2 \frac{1}{1+\frac{q_j-p_j}{p_j}}$ and $\log^2(1+x) \leq x^2$ for $x \leq 0.5$, (b) follows from the fact that $\sum_{j=3} \frac{(p_j-q_j)^j}{jp_j^{j-1}} \geq -\frac{1}{2}\frac{(p_j-q_j)^2}{2p_j}$ and $10\log\frac{1}{\varepsilon} \geq 4$ and (c) from Taylor expansion of $\log\frac{q_j}{p_j}$ around 1.

Now consider the case where $q_j > p_j$ and $q_j - p_j > \frac{p_j}{2}$. Then,

$$(p_j - q_j) + \frac{p_j \log^2 \frac{p_j}{q_j}}{10\log\frac{1}{\varepsilon}} \overset{(a)}{\leq} (1 - \frac{2}{10\log\frac{1}{\varepsilon}})(p_j - q_j)$$

Where (a) follows from $p_j \log^2 \frac{p_j}{q_j} = p_j \log^2 \frac{1}{1+\frac{q_j-p_j}{p_j}}$ and $\log^2(1+x) \leq x$. Using Taylor expansion and the inequality $\log(1+x) \leq 0.9x$ for $x > 0.5$, we have $|p_j \log\frac{p_j}{q_j}| \leq 0.9|p_j - q_j|$. Hence, $(p_j - q_j)(1 - \frac{2}{10\log\frac{1}{\varepsilon}}) \leq p_j \log\frac{p_j}{q_j}$ if $\varepsilon < \frac{1}{e^2}$.

Finally, for the case $p_j > q_j$ and $p_j - q_j > \frac{p_j}{2}$, $\log\frac{p_j}{q_j} \leq \log\frac{1}{\varepsilon}$ and

$$(p_j - q_j) + \frac{p_j \log^2 \frac{p_j}{q_j}}{10\log\frac{1}{\varepsilon}} \leq (p_j - q_j) + \frac{1}{10} p_j \log\frac{p_j}{q_j}.$$

By Taylor expansion of $\log\frac{q_j}{p_j}$ and the inequality $\frac{(p_j-q_j)}{p_j} \leq \frac{1}{2}$,

$$p_j \log\frac{p_j}{q_j} = (p_j - q_j)\left[1 + \sum_{j=1} \frac{(p_j - q_j)^j}{(j+1)p_j^j}\right] \geq \frac{10}{9}(p_j - q_j),$$

and therefore $(p_j - q_j) + \frac{1}{10}p_j \log\frac{p_j}{q_j} \leq p_j \log\frac{p_j}{q_j}$. For the final result over the entire matrix of P and Q, we average the inequality across all i. □ □

Proof. **(Lemma 63, Integration).**

Let us write $Y = Y_+ - Y_-$, where $Y_+ = Y \cdot \mathbf{1}\{Y \geq 0\}$ and $Y_- = Y \cdot \mathbf{1}\{Y \leq 0\}$. First let's consider the non-negative part of Y and note that $\frac{X-Y_+}{\sqrt{X}} \geq \tau$ is equivalent to $X \geq \left(\sqrt{Y + \frac{\tau^2}{4}} + \frac{\tau}{2}\right)^2$, by solving a quadratic equation and eliminating the one of two implied inequalities that gives

$\sqrt{X} < 0$. Because $Y_+ \geq Y$, $\frac{X-Y_+}{\sqrt{X}} > \tau$ implies $\frac{X-Y}{\sqrt{X}} > \tau$ and thus

$$\Pr\left\{\frac{X-Y_+}{\sqrt{X}} > \tau\right\} \leq \Pr\left\{\frac{X-Y}{\sqrt{X}} > \tau\right\} \leq f(\tau).$$

We then have that, for all $t > 0$:

$$\begin{aligned}
1 - f(\tau) &\leq \Pr\left\{X < \left(\sqrt{Y_+ + \tfrac{\tau^2}{4}} + \tfrac{\tau}{2}\right)^2\right\} && (4.23)\\
&\leq \Pr\left\{X \leq t \quad \text{or} \quad t < \left(\sqrt{Y_+ + \tfrac{\tau^2}{4}} + \tfrac{\tau}{2}\right)^2\right\} && (4.24)\\
&\leq \Pr\{X \leq t\} + \Pr\left\{t < \left(\sqrt{Y_+ + \tfrac{\tau^2}{4}} + \tfrac{\tau}{2}\right)^2\right\}
\end{aligned}$$

where the second line follows form the fact that the original event implies that one or the other of the two events hold (if both fail then the original event doesn't hold), and the second inequality is a union bound. It follows that for all $\tau > 0$ and $t > 0$

$$\Pr\{X > t\} \leq \Pr\left\{\left(\sqrt{Y_+ + \tfrac{\tau^2}{4}} + \tfrac{\tau}{2}\right)^2 > t\right\} + f(\tau).$$

In particular, we can choose τ according to t. Let $\tau^2 = at$ for some $a \in (0,1)$. Then $Y_+ > (1-\sqrt{a})t$ is equivalent to $\left(\sqrt{Y_+ + \frac{\tau^2}{4}} + \frac{\tau}{2}\right)^2 > t$, and we can write for all $t > 0$

$$\Pr\{X > t\} \leq \Pr\left\{\frac{Y_+}{1-\sqrt{a}} > t\right\} + f(\sqrt{at}). \tag{4.25}$$

Now write $\Lambda = \int_0^\infty f(\sqrt{t})\mathrm{d}t$, use the fact that for a non-negative random variable Z, $\int_0^\infty \Pr\{Z > t\}\mathrm{d}t = \mathbb{E}[Z]$, and integrate both sides of Equation (4.25), to obtain:

$$\mathbb{E}[X] \leq \frac{1}{1-\sqrt{a}}\,\mathbb{E}[Y_+] + \frac{\Lambda}{a}.$$

Choosing a good trade-off with $a = \frac{1}{2}$ and noting that $1 - \frac{1}{\sqrt{2}} > \frac{1}{4}$, we get:

$$\mathbb{E}[X] \leq 4\mathbb{E}[Y_+] + 2\Lambda. \tag{4.26}$$

Now consider the non-positive part of Y. Observe that $Y_- > t$ implies that $\frac{X-Y}{\sqrt{X}} > \frac{X+t}{\sqrt{X}} > \sqrt{4t}$. That last inequality follows from optimizing the middle expression over all $X > 0$ (or equivalently completing the square). Thus

$$\Pr\{Y_- > t\} \leq \Pr\left\{\frac{X-Y}{\sqrt{X}} > \sqrt{4t}\right\} \leq f(\sqrt{4t}).$$

By integrating both sides as before and rearranging, we get:

$$\mathbb{E}[Y_-] \leq \frac{\Lambda}{4} \quad \Rightarrow \quad 0 \leq -4\mathbb{E}[Y_-] + \Lambda. \tag{4.27}$$

By combining Equations (4.26) and (4.27), we obtain:

$$\mathbb{E}[X] \leq 4\mathbb{E}[Y] + 3\Lambda. \tag{4.28}$$

Lastly, we integrate f. Recall that f is of the form:

$$f(\tau) = \min\left\{\mathrm{e}^{-\frac{\tau^2}{2v+c\tau}+H}, 1\right\}$$

Perform the change of variable $\tau = \sqrt{t}$, and let t_0 be the point of switch of $f(\sqrt{t})$ between 1 and the exponential tail. By solving a quadratic equation, we obtain:

$$t_0 = \left[\sqrt{2vH + \frac{c^2H^2}{4}} + \frac{cH}{2}\right]^2 \leq 4vH + c^2H^2,$$

where for the last expression we have used Jensen's inequality, $\frac{1}{2}\sqrt{a} + \frac{1}{2}\sqrt{b} \leq \sqrt{\frac{a+b}{2}}$. It follows that:

$$\int_0^{t_0} f(\sqrt{t})\mathrm{d}t \leq t_0 \leq 4vH + c^2H^2.$$

Now note that $g(\tau) = \frac{\tau^2}{2v+c\tau}$ is convex and $g(\sqrt{t_0}) = 0$. We can thus bound it from below with the tangent line at $\sqrt{t_0}$. We have the derivative

$$g'(\tau) = \frac{2\tau(2v+c\tau) - c\tau^2}{(2v+c\tau)^2} = g(\tau)\left(\frac{2}{\tau} - \frac{c}{\tau^2}g(\tau)\right).$$

Recalling that $g(\sqrt{t_0}) = H$, we get that:

$$g(\sqrt{t}) - H \geq H\left(\frac{2}{\sqrt{t_0}} - \frac{cH}{t_0}\right)(\sqrt{t} - \sqrt{t_0}).$$

By using this tangent line bound with $f(\sqrt{\tau}) = \mathrm{e}^{-g(\sqrt{t})+H}$, we obtain:

$$\int_{t_0}^{\infty} f(\sqrt{t})\mathrm{d}t \leq \int_{t_0}^{\infty} \mathrm{e}^{-H\left(\frac{2}{\sqrt{t_0}} - \frac{cH}{t_0}\right)(\sqrt{t}-\sqrt{t_0})}\mathrm{d}t$$

Now apply the change of variable $s = \sqrt{t} - \sqrt{t_0}$ and note that $\mathrm{d}t = 2(s + \sqrt{t_0})\mathrm{d}s$ we get

$$\int_{t_0}^{\infty} f(\sqrt{t})\mathrm{d}t \leq 2\int_0^{\infty}(s+\sqrt{t_0})\mathrm{e}^{-H\frac{2\sqrt{t_0}-cH}{t_0}s}\mathrm{d}s \leq 2\frac{1+H(2\sqrt{t_0}-cH)/\sqrt{t_0}}{H^2(2\sqrt{t_0}-cH)^2/t_0}t_0,$$

where for the last bit we used $\int_0^{\infty}(x+a)\mathrm{e}^{-bx} = \frac{1+ab}{b^2}$.

Note that by using the definition of t_0, we have

$$\frac{\sqrt{t_0}}{(2\sqrt{t_0} - cH)} = \frac{1}{2} + \frac{1}{2}\frac{cH/2}{\sqrt{2vH + c^2H^2/4}} \in [\tfrac{1}{2}, 1]$$

using the lower end in the numerator and the upper in the denominator:

$$\int_{t_0}^{\infty} f(\sqrt{t})\mathrm{d}t \leq 2\frac{1+2H}{H^2}t_0.$$

This shows that for $H > 5$, we have $\int_{t_0}^{\infty} f(\sqrt{t})\mathrm{d}t \leq t_0$

Adding both contributions to the integral, we obtain:

$$\int_{t_0}^{\infty} f(\sqrt{t}) \leq 2t_0 \leq 8vH + 2c^2H^2.$$

□ □

Chapter 5

Towards Competitive N-gram Smoothing: Contextual Distribution Estimation

5.1 Introduction

Statistical N-gram language models have a long and rich history, and it is hard to give the literature justice in such a short space. The classical works are covered in the comprehensive survey [CG99],which empirically studied different smoothing techniques. Smoothing is critical to learning these models, since so much of the N-gram space remains unobserved. For a long time, the most successful smoothing technique was the one proposed by Kneser and Ney [KN95b]. This led to several efforts to explain its properties, mainly its use of backoff: reverting to a simpler model when data is scarce. Some of the best forays in this direction were the Bayesian perspective described by the hierarchical Pitman-Yor language models in [Teh06b] and, more pertinent to our work, the more recent developments exploring rank-reduction properties [HOF15, PSDX13, FOO16]. Despite these, no clear and complete understanding of the joint mechanisms of smoothing and back-off in Kneser-Ney have been suggested.

Perhaps this is due to the surge of neural networks and in particular recurrent neural language models, which led to a significant jump in performance [MKB$^+$10]. Neural language models have since continued to achieve increasingly better results [MKS17, YDSC17, GHT$^+$18, TSN18, DYY$^+$19]. Interestingly, *N*-gram techniques are still relevant as they usually run much faster, and, can be used in conjunction with neural models to improve performance even further [XWL$^+$17]. Moreover, for low-resource languages, non-neural methods or a combination of neural and non-neural methods are known to achieve the best performance [GML14].

For these reasons, we were motivated to lead the current effort to get a first theoretical handle into *N*-gram models, and in particular the principles behind the practice of back-off. As evidence of the promise of this exploration, we also report on a powerful generalization of Kneser-Ney backoff, which is empirically able to compete with neural models, albeit those limited to feed-forward architectures. It is worth mentioning that the smoothing aspect of *N*-gram models, understood primarily as high-dimensional categorical distribution estimation, has received attention. Part of the novelty of our perspective is to study back-off through the very same lens, namely that of competitive distribution estimation. This notion was expressed most clearly in the result of [OS15], where it was used to give a clear justification to the Good-Turing estimator [Goo53], which is intimately related to Kneser-Ney.

Our contributions are as follows:

- We study this problem as a *contextual distribution estimation* problem. Apart from the fact that this means we aim to learn conditional distributions, the objective function and notions of competitivity have to both be carefully set. We do this in Section 5.3 and show that competitivity is possible in the contextual setting, and give some evidence for the advantage that Kneser-Ney has.

- We generalize the Good-Turing estimator to the contextual setting, in Section 5.4. The idealized expression of this estimator cannot be used directly and needs to be smoothed, just like in the non-contextual setting. We show that with the proper smoothing, contextual Good-Turing recovers the Kneser-Ney estimator, when the tails of the distributions are power laws and are aligned.

- The idealized Good-Turing formula is much more powerful than the special case of Kneser-Ney. We conjecture that it could potentially offer competitivity versus oracles that are aware of intricate relationships between distributions in various contexts. We illustrate this potential by giving a strict generalization of Kneser-Ney back-off, which we call *Partial Low-Rank*, since it applies the rank structure only to the rare part of the data. Kneser-Ney corresponds to the rank-1 special case.

- In Section 5.6, we show that Partial Low-Rank uniformly improves on Kneser-Ney on various benchmarks. Furthermore, a nested trigram-level implementation of this approach meets and slightly exceeds the performance of the feed-forward neural models on the Penn Tree Bank data set. This advantage is only enhanced by considering that it can be trained with a fraction of the time and space resources required for the neural model.

We start with some preliminaries in Section 5.2.

5.2 Preliminaries

We describe the problem generally. Let the context space be $\mathcal{X}$ and the prediction space be $\mathcal{Y}$. When finite, identify these spaces with $\mathcal{X} = [K]$ and $\mathcal{Y} = [k]$ respectively. Data is modeled as n context/prediction pairs $(X_t, Y_t)_{t=1,\cdots,n}$. How is this data generated? Various scenarios may be considered. In modeling sequence data, as in the case of language modeling, the ideal context is usually the whole history. Namely, given an infinite history $X_t \overset{\text{def}}{=} (\cdots, Y_{t-2}, Y_{t-1})$, there is a conditional probability of observing the next word Y_t. When the history is truncated to $N-1$ words, this is called an N-gram model. Other history-to-context mappings $X_t \overset{\text{def}}{=} f(\cdots, Y_{t-2}, Y_{t-1})$ may also be considered, such as skip-grams or word embeddings [MCCD13, MSC$^+$13]. If the data consists of just a single long sequence, such as a text $Y_1, Y_2, Y_3, \ldots, Y_n$ of n words, the context/prediction pairs that result from partitioning the text are correlated.

We simplify this by assuming that X_t are independently and identically drawn from some distribution π over $\mathcal{X}$. In this case, independently for each context X_t, we take Y_t to be distributed according to a conditional distribution $P_{ij} \overset{\text{def}}{=} \mathbb{P}(Y = j|X = i)$. Note that for a given i, $p_i \overset{\text{def}}{=} (P_{ij})_{j\in[k]}$ is a distribution over $\mathcal{Y}$. The matrix $C_{i,j} = \sum_t \mathbb{1}\{X_t = i, Y_t = j\}$ then summarizes the data.

Our goal can now be concisely stated as: given $(X_t, Y_t)_{t=1,\cdots,n}$ or $(C_{i,j})_{i\in\mathcal{X},j\in\mathcal{Y}}$, estimate $(P_{ij})_{i\in\mathcal{X},j\in\mathcal{Y}}$. We judge the performance of an *estimator* Q that maps data to contextual probabilities, according to a suitably defined statistical risk. The primary goal of contextual probability estimation is to make accurate predictions, requiring the estimated conditional probability to be close to its true value on new data.

We consider the underlying risk of an estimator as being the *KL-risk*, defined as the averaged per-context Kullback-Leibler divergence

$$D_\pi(P\|Q) \overset{\text{def}}{=} \sum_i \pi_i \sum_j P_{ij} \ln \frac{P_{ij}}{Q_{ij}}, \tag{5.1}$$

which captures relative closeness of estimated and true probabilities, on average. It is the risk associated with the *log*-loss and, up to the entropy, is the population cross-entropy of Q. Since Q is random, guarantees are often given in terms of the expected KL-risk $r_n(\pi, P, Q) = \mathbb{E}[D_\pi(P\|Q)]$ or as high-probability or worst-case bounds.

Ideally, we would like to have the best Q possible, an optimal one. In the non-contextual setting, when $K = 1$, one measure of optimality is worst-case risk with respect to a class $\mathcal{P}$, $r_n(Q, \mathcal{P}) = \max_{P\in\mathcal{P}} r_n(\pi, P, Q)$. The best possible such risk is known as the minimax risk of the class, $r_n(\mathcal{P}) = \min_Q r_n(Q, \mathcal{P})$. A minimax optimal Q (either exactly or in rate) is desirable but pessimistic: minimax optimality does not capture the possibility of the truth being in a smaller class. The competitive loss with respect to a family $\mathcal{F}$, which contains many such (some small, some big) classes, is a more optimistic notion. It is defined as $\varepsilon(Q,\mathcal{F}) = \max_{\mathcal{P}\in\mathcal{F}}[r_n(Q,\mathcal{P}) - r_n(\mathcal{P})]$. This is related to the notion of adaptivity in statistics.

Think of a family and an oracle that can determine exactly which $\mathcal{P}$ in $\mathcal{F}$ we have, and does the best for it. When an estimator has small ε it means it manages to do as well as this oracle itself. To see the reason for optimism, say $\mathcal{P}$ is nice, with a small $r_n(\mathcal{P})$. The estimator will achieve this small risk, even if $\mathcal{F}$ contains such large classes for which $r_n(\mathcal{P})$ is enormous. If an estimator has ε that is of the same order as the minimax risk, we call it *competitive*. It effectively discovers the true $\mathcal{P}$. We currently have a nascent theory for non-contextual estimators that are competitive with respect to rich families. These include the works that show that the Good-Turing estimator combined with the empirical estimator has dimension-free competitivity, [OS15], and that the absolute discounting estimator competes with oracles aware of the effective alphabet size of the distribution, [FOOP17]. A similar notion can also be found in [VV15].

In the bigram setting, $K = k$, Kneser-Ney back-off can be described as follows. For every context / row of C, perform absolute discounting, defined, for a given $\alpha < 1$ as:

$$Q_{ij} = (C_{ij} - \alpha)/n_i, \qquad \text{when } C_{ij} > 0, \text{ and where } n_i = \textstyle\sum_j C_{ij},$$

the overall missing mass discounted through the α has to then be redistributed over unseen predictions with $C_{ij} = 0$. If we perform this redistribution uniformly, then this is row-wise absolute discounting, equivalent roughly to row-wise Good-Turing [FOOP17]. But instead, Kneser-Ney *backs-off* to an alternate distribution, by redistributing the missing mass proportionally to backoff counts $b_j = \sum_i \mathbf{1}\{C_{ij} > 0\}$. The variant proposed by Chen-Goodman [CG99] performs a further absolute discounting on b, and redistributes proportionally to $(b_j - \alpha)/\sum_{j'} b_{j'}$, and once again the missing mass within the b is uniformly distributed over the b_j that are zero. Despite being such a simple estimator, Kneser-Ney backoff, and especially the Chen-Goodman variant held state-of-the-art performance for a while.

5.3 Theoretical Insights

We now give a tentative theoretical exploration of the advantage of back-off through the lens of competitive distribution estimation. Let us first define the contextual competitive loss of an estimator Q, with respect to a family $\mathcal{F}$ of classes. In general $\mathcal{F}$ contains classes $\mathcal{C}$, which are sets containing pairs (π, P). To simplify, we take π to be arbitrary, or equivalently each $\mathcal{C}$ is of the form $\Delta_K \times \mathcal{P}$ where $\mathcal{P}$ is a class of Ps only. The competitive loss of an estimator is then:

$$\varepsilon_n(Q,\mathcal{F}) := \max_{\mathcal{C}\in\mathcal{F}} \max_{(\pi,P)\in\mathcal{C}} \left[r_n(\pi,P,Q) - r_n(\mathcal{P})\right], \tag{5.2}$$

$$= \max_{\Delta_K\times\mathcal{P}\in\mathcal{F}} \max_{\pi\in\Delta_K} \max_{P\in\mathcal{P}} \left[r_n(\pi,P,Q) - r_n(\mathcal{P})\right], \tag{5.3}$$

where the risk of the estimator Q is its expected KL-risk,

$$r_n(\pi,P,Q) = \mathbb{E}_{(x^n,y^n)\sim\pi P}\left[\textstyle\sum_i \pi_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}}\right], \tag{5.4}$$

and the minimax risk of the class $\mathcal{C} = \Delta_K \times \mathcal{P}$ achieved by an optimal estimator $Q^{\mathcal{C}}$,

$$r_n(\mathcal{P}) \overset{\text{def}}{=} \min_{Q} \max_{\pi\in\Delta_K} \max_{P\in\mathcal{P}} r_n(\pi,P,Q). \tag{5.5}$$

The choice of family $\mathcal{F}$ is again equivalent to competing with an oracle/genie that can determine the true (π,P) up to a class $\mathcal{C}$ which it belongs to. We are in particular considering oracles that are uninformed about π, but know that P belongs to some class $\mathcal{P}$ allowed by $\mathcal{F}$.

5.3.1 Basic competitivity

Consider an oracle $\mathcal{F}_1$ that knows each row of P up to permutation. Is it possible to compete with it? Intuitively, one ought to be able to, by reducing to the non-contextual competitive estimator in each context. There are some subtle points to consider, however. One is the fact that the number of samples that each context receives is random. More importantly, the number K of contexts plays a role in how competitive we can be.

Let Q^{GT} be the per-context Good-Turing estimator. This analysis also describes absolute-discounting applied to data from each context separately. Out of n total samples, let n_i denote those that fall in context i. In the non-contextual case, the Good-Turing estimator with n samples has a competitive loss of $O(\min\{\frac{1}{\sqrt{n}}, \frac{k}{n}\})$. Note that it is dimensionless in the high-dimensional regime. So we intuitively expect the same to hold per context. We formally extend this to the overall contextual case.

Theorem 64. *We have* $\varepsilon_n(Q^{\mathsf{GT}}, \mathcal{F}_1) \le O\left(\min\{1, \sqrt{\frac{K}{n}}, \frac{K \cdot k}{n}\}\right)$. *This implies three distinct regimes:*

$$\varepsilon_n(Q^{\mathsf{GT}}, \mathcal{F}_1) \le \begin{cases} O(\frac{K\cdot k}{n}) & n > K \cdot k^2 \\ O(\sqrt{\frac{K}{n}}) & K < n < K \cdot k^2 \\ O(1) & n < K \end{cases}$$

The proof is in the supplements (Appendix 5.A). Theorem 64 thus generalizes non-contextual results in a data-diluted form: effectively replacing n by n/K. The first case is the low-dimensional regime. Perhaps the most relevant is the middle high-dimensional regime. This often holds in the case of bigrams ($K = k$) and trigrams ($K = k^2$). In this case we recover the prediction-dimensionless (in k) bound. For large K and k, this loss is negligible compared to the minimax risk [FOO16], implying true competitivity (for more on minimax risks see the supplements, Appendix 5.B). Not that in the third (extreme high-dimensional) regime, the unobserved contexts give no advantage to this oracle, leading to a competitivity that does not decay but also does not depend on the number of contexts.

5.3.2 Stronger competitivity

In this chapter, we conjecture that the advantage of back-off is in providing a much stronger form of competitivity. We use the following intuition. The competitivity of the Good-Turing estimator shows that the difficulty of the problem is *not* in estimating the multiset of probabilities *as much as it is* in identifying in which permutation they map to the categories, the only task of the oracle given data. One can think of this as aligning the tail of the distribution. In the contextual setting, this intuition still persists. But another joins it: tails are often related across contexts, and since the identities of the categories are shared across contexts, the oracle then ought to be able to better align within each context too. To make this intuition concrete, consider the following idealized scenario.

Consider an oracle $\mathcal{F}_2$, that knows P has exactly $m \leq k$ non-zero columns, but not which ones they are. Thus $\mathcal{P} \in \mathcal{F}_2$ are indexed simply by m. This idealizes two aspects of the problem. First, there is a non-ambiguous tail (the zeros) in each context. And second, all these tails are clearly correlated across contexts by being aligned. It turns out that the Kneser-Ney back-off estimator Q^{KN} strongly competes with this oracle.

Theorem 65. *If $n \gg k$ then*

$$\varepsilon(Q^{\mathsf{KN}}, \mathcal{F}_2) \leq O\left(\frac{k}{n}\right).$$

Consider the regime where $n > k^2/K$. Usually $K \geq k$, think of N-grams, in which case it would suffice that $n > K$. It is easy to verify that in this regime the bound of Theorem 65 is strictly better than that of Theorem 64, and has the distinct benefit of not scaling with the dimension of contexts. It is also worth mentioning that the proof of this result (supplements, Appendix 5.A) gives the finer class-by-class competitive loss of $O(m/n)$. Q^{KN} achieves this without prior knowledge of m.

We believe this simple case reinforces the idea that tail alignment *across contexts* is a fruitful avenue for competitivity in the contextual case, just as tail alignment *within contexts* was a fruitful one in the non-contextual case. Classes that ease the latter alignment, such as power law decay or small effective support size, enjoy lower competitive loss by Q^{GT}. These factors are invariant under permutations. This suggests that unlike the oracle $\mathcal{F}_1$ that permutes within each context separately, the natural notion of invariance in the contextual case ought to be under simultaneous permutation across contexts, i.e. permutation of entire *columns* of P.

5.4 Contextual Good-Turing

Motivated by this theoretical foray, and with the goal of giving a principled underpinning to Kneser-Ney smoothing and the hope of deriving estimators with more favorable competitive properties, we revisit the derivation of the original Good-Turing estimator and extend it to the contextual case.

Good-Turing is based upon an empirical Bayes construction. To parallel it in the contextual setting, assume the multiset of the columns of P is known and that P is instanced via a uniformly random permutation of these columns. Let x be some context and y be some prediction, such that our ultimate goal is to estimate P_{xy}. The chance that y is any particular $j \in [k]$ is a priori $1/k$, and thus $\mathbb{E}[P_{xy}] = 1/k$. But having made some observations, we would like to determine the conditional expectation of P_{xy} given that there are $\mathbf{n} := (n_i)_{i\in[K]}$ samples in each context and given that y has been observed in each context $\mathbf{c} := (C_{i,y})_{i\in[K]}$ times[1]. Starting with the simple observation that

$$\Pr\{\mathbf{c} \mid \mathbf{n},\ y = j\} = \prod_i \binom{n_i}{C_{ij}} P_{ij}^{C_{ij}} (1-P_{ij})^{n_i - C_{ij}},$$

one can show that

$$\mathbb{E}[P_{xy} \mid \mathbf{n}, \mathbf{c}] = \frac{\sum_j \prod_i P_{ij}^{C_{ij}+\mathbf{1}\{i=x\}} (1-P_{ij})^{n_i - C_{ij}}}{\sum_j \prod_i P_{ij}^{C_{ij}} (1-P_{ij})^{n_i - C_{ij}}} \equiv \frac{c_x + 1}{n_x + 1} \frac{\mathbb{E}[K_{\mathbf{n}+\mathbf{1}_x, \mathbf{c}+\mathbf{1}_x} | \mathbf{n}]}{\mathbb{E}[K_{\mathbf{n},\mathbf{c}} | \mathbf{n}]}. \tag{5.6}$$

[1] Note that this is *not* the entire information useful to determine the permutation, just local information that makes the task tractable.

Here $K_{\mathbf{n},\mathbf{c}}$ is the number of columns that have exactly the $\mathbf{c}$ count pattern. The expectation of this quantity is column-permutation invariant, thus can be computed from the multiset. The Good-Turing approach is to use this expression as an estimator, substituting the expectations with their empirical counterparts (with the shift of $n_x + 1$ to n_x, since the additional sample is not available empirically):

$$\hat{P}_{xy} = \frac{c_x+1}{n_x}\frac{K_{\mathbf{n},\mathbf{c}+\mathbf{1}_x}}{K_{\mathbf{n},\mathbf{c}}}. \tag{5.7}$$

The challenge is that, even in the non-contextual case, these can be highly unreliable, and one needs to smooth them, such as by combining with the empirical distribution in the abundant range or by using absolute discounting. In the contextual case, even more smoothing is needed: $\mathbf{c}$ may be observed, but it's very unlikely that $\mathbf{c}+\mathbf{1}_x$ is, and the estimator degenerates. How can we remedy this?

5.4.1 From Contextual Good-Turing to Classical Back-off

Back-off is an intuitive notion, but was originally proposed in an ad hoc fashion. We now show that contextual Good-Turing naturally gives rise to back-off. We start by observing that if one sums the total probability assigned to all symbols that appear μ times in context x, the estimator (5.7) gives us back the non-contextual Good-Turing estimate of that probability:

$$\sum_{y:\mathbf{c}(y),c_x(y)=\mu} K_{\mathbf{n},\mathbf{c}(y)}\hat{P}_{xy} = \frac{\mu+1}{n_x+1}\sum_{y:\mathbf{c}(y),c_x(y)=\mu} K_{\mathbf{n}+\mathbf{1}_x,\mathbf{c}(y)+\mathbf{1}_x} = \frac{\mu+1}{n_x+1}K_{n_x,\mu+1}(x)$$

This shows that (5.7) simply redistributes this mass. This is the main premise of Kneser-Ney back-off. Does it redistribute it similarly to Q^{KN}? In general, no. But we can identify when exactly it does. We give the following general smoothing strategy, which we can think of as binning. For a given x, choose an equivalence $\sim$, compatible with the contextual Good-Turing estimator, namely that satisfies (1) $\mathbf{c}\sim\mathbf{c}'$ implies $\mathbf{c}_x = \mathbf{c}'_x$ (fixes x) and (2) if $c \sim c'$ then $\mathbf{c}_\sigma \sim \mathbf{c}'_\sigma$ for any permutation σ of $[K]$ (invariant under permutations of contexts). We smooth by spreading probability within each equivalence bin and counting all equivalent $\mathbf{c}$ as being identical:

$$\tilde{P}_{xy} = \frac{K_{\mathbf{n}+\mathbf{1}_x,\sim\mathbf{c}+\mathbf{1}_x}}{K_{\mathbf{n},\sim\mathbf{c}}}. \tag{5.8}$$

We can verify that this preserves the mass redistribution property. Let $\mathsf{nnz}(\mathbf{c}) = \sum_i \mathbf{1}\{c_i > 0\}$ count the number of non-zero entries of $\mathbf{c}$. Then the following defines a possible equivalence class:

$$\mathbf{c} \sim \mathbf{c}' \iff \begin{cases} c_x = c'_x \\ \mathsf{nnz}(c) = \mathsf{nnz}(c') \end{cases} \tag{5.9}$$

In this case we can characterize the redistribution accurately, at least in its idealized form.

Theorem 66. *Use the equivalence relation of Equation* (5.9) *in the smoothed contextual Good-Turing estimator* (5.8)*, where the counts K are substituted by their idealized expectations. Then:*

$$\check{P}_{xy} = \frac{\mu+1}{n_x+1} \frac{\sum_j \binom{n_x+1}{\mu+1} P_{xj}^{\mu+1}(1-P_{xj})^{n_x-\mu}\rho_{bj}}{\sum_j \binom{n_x}{\mu} P_{xj}^{\mu}(1-P_{xj})^{n_x-\mu}\rho_{bj}},$$

where $b = \mathsf{nnz}(\mathbf{c})$ *and* $\rho_{bj} = \sum_{S\subset[K]\setminus\{x\}\,:\,|S|=b} \prod_{i\in S}[1-(1-P_{xj})^{n_i}] \prod_{i\in S^c\setminus\{x\}} (1-P_{xj})^{n_i}.$

We omit the proof of this result, since it's straightforward manipulations. It is more important to observe that, apart from ρ, this is exactly the non-contextual Good-Turing expression. Thus ρ acts as a redistribution coefficient. In general, it does not quite redistribute like Kneser-Ney: unlike it, ρ depends on the context x. Observe however that only the small values (of the order of $1/n_i$) of P_{xj} contribute to ρ. Let us assume that these values are aligned across rows (do not depend on x), that they have a power law decay of index α, and that the n_i are roughly uniform. We can then show that the effect of ρ is asymptotically approximately given by (see supplements, Appendix 5.A):

$$\check{P}_{xy} \sim \frac{\mu + b_y - \alpha}{n + b_y}. \tag{5.10}$$

For the unseen symbols, when $\mu = 0$, this recovers the Chen-Goodman version of Kneser-Ney smoothing (see Section 5.2, and note that b_y is negligible with respect to n.)

5.5 Partial Low-Rank *N*-gram Backoff

It is enlightening that contextual Good-Turing, an empirical Bayes estimator derived from column-permutation invariance, when properly smoothed, recovers classical forms of back-off under the kind of tail alignment conditions that make these competitive in the first place. It is then natural to ask whether contextual Good-Turing's competitive properties extend further than such obvious alignment, especially that it is not explicitly aware of it. Indeed, column-permutation invariance has the potential to capture a much richer family of tail structures: the rank of P, its sparsity, the dimension of the manifold on which each row of P lies, such as the the embedding dimension in typical neural embeddings, and many other classical structures, are all invariant under such permutation.

What is needed to achieve this generality is a more flexible smoothing of the idealized contextual Good-Turing formula of Equation (5.6). Based on this idea, we now give a direct generalization of bigram Kneser-Ney smoothing. First, refine the equivalence relation given by (5.9), and use instead

$$\mathbf{c} \sim \mathbf{c}' \iff \begin{cases} c_x = c'_x \\ \forall i \; \mathbf{1}\{c_i > 0\} = \mathbf{1}\{c'_i > 0\} \end{cases} \tag{5.11}$$

Two columns are thus considered equivalent if their non-zero patterns align. This is clearly a coarser binning than maintaining the full identity of $\mathbf{c}$, but is much finer than the partition induced by (5.9). It is indeed too fine to effectively smooth $K_{\mathbf{n},\mathbf{c}}$ in general. However, it allows us to create a hierarchy of refinements of which itself is one extreme, and Kneser-Ney is another, a projection onto a subspace of one dimension. If we allow this subspace to be of a larger dimension m, then we are equivalently representing the non-zero indicator matrix

$B = (B_{i,j})_{i\in[X],j\in[k]}$, $B_{i,j} = \mathbf{1}\{C_{ij} > 0\}$ by a rank-m approximation. Since this is not a low-rank representation of the raw count matrix C, effectively only of its rare component, we dub it *partial low-rank* or PLR.

The rank-1 approximation recovers Kneser-Ney backoff, and is easily obtained by collapsing the counts $b_y = \sum_i B_{iy}$, which are then discounted to create the Chen-Goodman backoff distribution. A general simultaneous rank-reduction and smoothing was proposed recently by [FOOP17] based on multiplicative-update non-negative matrix factorization. That is what we propose to apply to B for the general rank case. For a given partial low-rank m, the full algorithm is given in Algorithm 1.

Algorithm 1 Partial Low-Rank (PLR)

```
inputs
    Count matrix $C$, rank $m$, discount $\alpha$, number of iterations $T$, (optional) initial $W_0$ and $H_0$
outputs
    Distribution matrix $Q^{\mathsf{PLR}}$, (optional) intermediate components $A$, $\nu$, $W$, and $H$
Perform $\alpha$ absolute discounting on each row of $C$ to obtain:
    The abundant component $A$, $A_{ij} = (C_{ij} - \alpha)/n_i$
    The missing mass vector $\nu$, $\nu_i = d_i\alpha/n_i$, where $d_i = \sum \mathbf{1}\{C_{ij} > 0\}$
Generate the indicator matrix $B$, $B_{i,j} = \mathbf{1}\{C_{ij} > 0\}$
If $W_0$ and $H_0$ are *not* provided:
    Initialize the $K \times m$ matrix $W_0$ to be uniform $1/m$ in each row
    Split $B$ into $m$ random row-blocks, collapse each to obtain the $m \times k$ soft-count matrix $\tilde{H}_0$
    Perform $\alpha$ absolute-discounting in each row of $\tilde{H}_0$ to obtain $H_0$
for $t = 1$ to $T$ do
    $\tilde{W}_t \longleftarrow \left[(C \oslash W_{t-1}H_{t-1})H_{t-1}^{\mathsf{T}}\right] \otimes W_{t-1}$
    Add $1/2$ to each row of the soft-count matrix $\tilde{W}$ and normalize each row to obtain $W_t$
    $\tilde{H}_t \longleftarrow \left[W_t^{\mathsf{T}}(C \oslash W_t H_{t-1})\right] \otimes H_{t-1}$
    Perform $\alpha$ absolute-discounting on each row of the soft-count matrix $\tilde{H}_t$ to obtain $H_t$
end for
return $Q^{\mathsf{PLR}} = A + \mathsf{diag}(\nu) \cdot W_T H_T$
```

Here $\otimes$ and $\oslash$ denote element-wise multiplication and division. When $K = k$, the PLR algorithm applies as-is to perform bigram smoothing. With $m = 1$, it is identical to Kneser-Ney. With larger m, it describes a much richer set of tail alignments. One could extend this to N-grams in two ways, directly by setting $K = k^{N-1}$, or by nesting like Chen-Goodman's recursive application of back-off, by fixing sub-contexts [CG99]. This general nested back-off N-gram smoothing algorithm, NPLR, along with other technical details of both algorithms, is presented in the supplements (Appendix 5.C.) Iterations can be very efficiently implemented using either sparse matrix manipulations or dictionaries. They run in linear time in n, as only observed contexts need to be tracked.

5.6 Experiments

In this section, we reinforce the theory and concepts so far with experiments. The goal is to show both that we can improve traditional smoothing techniques and that we can better aid neural language models. Toward the first goal, we show improved performance not only in language modeling, that is predicting the the next word given its history, but also the Global Terrorism Database, where we predict the target type for the next attack for a given city.

5.6.1 N-gram language models

We perform word-level language modeling on the Penn TreeBank (PTB) data set [Mik12] using standard splits (929k training tokens, 73k validation tokens, and 82k test tokens.) The vocabulary size is 10k. We compare different models in terms of their perplexity, the exponentiated $exp(\text{cross-entropy})$. These express the uncertainty in prediction, therefore, the lower, the better. We compare the original Kneser-Ney (KN) with the NPLR algorithm (nested version of PLR). We look at two variants, both of which are 5-gram models: partial low-rank applied either at only the bigram level PLR ($m_2 = 30$), or at both the trigram *and* bigram levels ($m2 = 18, m_3 = 4$). At higher levels, we maintain the KN structure in both cases, i.e. $m_4 = m_5 = 1$. We set $(\alpha_2, \cdots, \alpha_5) = (0.8, 0.9, 0.9, 0.9)$. We also include three neural models, two 5-grams, one feedforward and one LSTM, and one 13-gram feedforward, all reported in [CNB17]. Table 5.1 summarizes these results. We show significant improvement over KN and a modest improvement over the 5-gram feedforward model. The gap with the 5-gram LSTM is expected, considering its extensive weight-sharing, not done here. To the best of our knowledge, no other direct N-gram smoothing technique, especially none of the attempts to explain and generalize back-off, have reported such dramatic gains.

Table 5.1: Perplexity on PTB - NPLR smoothing surpasses KN and competes with feedforward NNs

	Method	Test Perplexity
5-gram	KN	143
5-gram	NPLR $m_2 = 30, m_3 = m_4 = m_5 = 1$	**131**
5-gram	NPLR $m_2 = 18, m_3 = 4, m_4 = m_5 = 1$	**126**
5-gram	Feedforward NN [CNB17]	127
13-gram	Feedforward NN [CNB17]	125
5-gram	LSTM [CNB17]	103

5.6.2 LSTM language models with data noising

Recently, [XWL$^+$17] utilized smoothing techniques as a data noising method for LSTM language models. Replacing some words in the input data changes the counts of N-grams in a way that applying an empirical estimator to the noised data is similar to applying N-gram smoothing techniques to the non-noisy data. In Table 5.2 we show that PLR, if used as a data noising technique, improves the perplexity of an LSTM language model more than all the other techniques, even the best noising based on Kneser-Ney. The LSTM setup is the same as the large-network used in [XWL$^+$17] (2-layer LSTM with 1500 hidden units) and we trained our models using the same setup as in [XWL$^+$17, ZSV14]. The noising parameter is tuned based on the validation data and the result is reported for the best noising parameter.

Table 5.2: Effect of smoothing as data noising on validation and test cross-entropy for PTB

Noise scheme	Validation	Test
none [XWL$^+$17]	81.6	77.5
unigram [XWL$^+$17]	79.4	76.1
bigram Kneser-Ney [XWL$^+$17]	76.2	73.4
PLR	**75.5**	**72.7**

5.6.3 Global terrorism target prediction

Language modeling is *the* flagship application of smoothing techniques and most of the new techniques are put to test there. However, the challenge of predicting rarely seen elements conditioned on some context, is present in a multitude of other natural applications. Even the power-law structure of language is also present in many natural phenomena. Here we explore one such alternative application: predicting terror incidents. We use the Global Terrorism Database (GTD) [ftSoTtTS18], which includes systematic data about more than $180,000$ cases of domestic and international terrorist events from 1970 through 2017 for more than $36,000$ cities around the globe. The task that we consider is to estimate the probability of the next attack in a given city has a specific target. This is a contextual probability estimation: the context is the city and the prediction is the target type. There are 114 different target sub-types, such as restaurants, banks, hotels, and etc.

We predict the type of the next attack in each city based on the prior incidents that happened in that year and compare four different estimators: row-wise add-half, row-wise absolute discounting, Kneser-Ney (KN), and PLR. Figures 5.1(b) and 5.1(a) show the benefit of using data from different cities (different rows of the count matrix) when predicting for the target type in a particular city. PLR and KN always have significantly better performances in predicting the next attack's type than row-wise estimators such as add-half and absolute discounting. Also, PLR shows an edge over KN in terms of generalization power. In all experiments, PLR is set to use rank $m = 5$ and the discount factor $\alpha = 0.9$, and is run for 100 iterations.

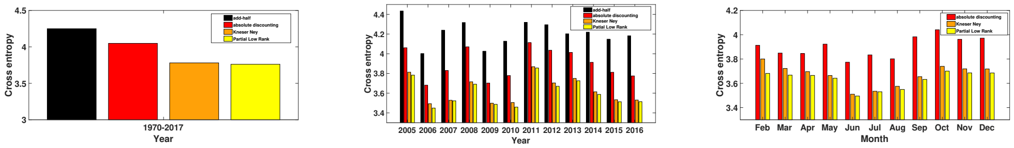
Lastly, we study how performance changes as the amount of data available for estimation varies. For this, we base our predictions for a particular year and predict the target type for two weeks in a month, using all the incidents before that time period in that year. As we move towards the end of the year, the amount of data available for estimation increases. Figure 5.1(c) shows how cross-entropy changes for different months in the year 2017. It is clear that the relative gain of PLR is more in the case when data is not abundant. More experiments are reported in the supplements (Appendix 5.D.)

5.7 Conclusion

We initiated a first theoretical exploration of N-gram smoothing, through the lens of competitivity, and we discovered powerful new generalizations. We hope this to provide momentum toward a mature theory and practice of competitive contextual distribution estimation.

5.8 Acknowledgment

Chapter 5, in full, has been submitted for the publication of the material as it may appear in *Neurips*. Falahatgar, Moein, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati, 2019. The dissertation author was one of the primary investigators and authors of this paper.



(a) All data

(b) Yearly data

(c) Effect of training size

Figure 5.1: Test cross-entropy with training / test periods: (a) dates before / after 2015, (b) the first 10 months / the last two months of the year, and (c) all dates prior to a month / that month of the year.

5.A Deferred Proofs

5.A.1 Proof of Theorem 64

We will need the following lemma.

Lemma 67. *Let $X \sim Poi(\lambda)$, then,* $\mathbb{E}[\frac{1}{\sqrt{x+1}}] \leq \frac{1}{\sqrt{\lambda}}$ *and* $\mathbb{E}[\frac{1}{X}] \leq \frac{1}{\lambda}$.

Proof.

$$\begin{aligned}
\mathbb{E}[\frac{1}{\sqrt{X+1}}] &= \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\frac{1}{\sqrt{x+1}} \\
&= \sum_{x=0}^{\infty} \frac{\sqrt{x+1}}{\lambda}e^{-\lambda}\frac{\lambda^{x+1}}{(x+1)!} \\
&= \sum_{x=0}^{\infty} \frac{\sqrt{x}}{\lambda}e^{-\lambda}\frac{\lambda^x}{x!} \\
&= \frac{\mathbb{E}[\sqrt{X}]}{\lambda} \\
&\leq \frac{\sqrt{\mathbb{E}[X]}}{\lambda} = \frac{1}{\sqrt{\lambda}}
\end{aligned}$$

where the last line is by Jensen's inequality. Proof of the other part is similar and thus omitted. □

Theorem 64. First note that if we do not observe any samples from a row, the competitive loss will be zero as we do same as the oracle.

Rewriting the competitive loss in (5.4):

$$
\begin{aligned}
\varepsilon(Q^{\mathsf{GT}}, \mathcal{F}_1) &\overset{(a)}{\leq} \max_{\pi} \max_{\mathcal{P}} \max_{P \in \mathcal{P}} \Bigg[\mathbb{E}_{(x^n, y^n) \sim \pi P} \sum_i \pi_i \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathsf{GT}}_{ij}((x^n, y^n))} \\
&\qquad\qquad - \max_{\pi} \max_{P \in \mathcal{P}} \mathbb{E}_{(x^n, y^n) \sim \pi P} \sum_i \pi_i \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathcal{C}}_{ij}((x^n, y^n))} \Bigg] \\
&\overset{(b)}{\leq} \max_{\pi} \max_{\mathcal{P}} \max_{P \in \mathcal{P}} \Bigg[\mathbb{E}_{(x^n, y^n) \sim \pi P} \sum_i \pi_i \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathsf{GT}}_{ij}((x^n, y^n))} \\
&\qquad\qquad - \max_{P \in \mathcal{P}} \mathbb{E}_{(x^n, y^n) \sim \pi P} \sum_i \pi_i \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathcal{C}}_{ij}((x^n, y^n))} \Bigg] \\
&\overset{(c)}{=} \max_{\pi} \mathbb{E}_{x^n \sim \pi} \sum_i \pi_i \Bigg[\max_{\mathcal{P}_i} \max_{p_i \in \mathcal{P}_i} \mathbb{E}_{(y^n | x^n) \sim p_i} \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathsf{GT}}_{ij}((x^n, y^n))} \\
&\qquad\qquad - \max_{p_i \in \mathcal{P}_i} \mathbb{E}_{(y^n | x^n) \sim p_i} \sum_j P_{ij} \log \frac{P_{ij}}{Q^{\mathcal{C}}_{ij}((x^n, y^n))} \Bigg] \\
&\overset{(d)}{\leq} \max_{\pi} \mathbb{E}_{n_i \sim \pi} \sum_{i=1}^{K} \pi_i \min\{\frac{1}{\sqrt{n_i}}, \frac{k}{n_i}\} \\
&\overset{(e)}{\leq} \max_{\pi} \sum_{i=1}^{K} \pi_i O(\min\{1, \frac{1}{\sqrt{n\pi_i}}, \frac{k}{n\pi_i}\}) \\
&\overset{(f)}{\leq} O(\min\{1, \sqrt{\frac{K}{n}}, \frac{K \cdot k}{n}\})
\end{aligned}
$$

where (a) is because any estimator gives us an upper bound on the competitive loss, (b) is by removing $\max_{\pi}$ from the second term in the bracket and make the term inside the bracket larger. Also, (c) follows from the independence of rows' distributions where $\mathcal{P}_i$ indicates the class of distributions for the ith row induced by $\mathcal{P}$, and p_i is ith row distribution. Results of [OS15] yields (d), (e) is by Lemma 67 and (f) follows from concavity of $\sqrt{x}$ function. Note that adding $O(1)$ is because the competitive loss for each row is always less than a constant and this trivial bound yields a better upper bound for the range $n < K$. □

5.A.2 Proof of Theorem 65

We show that Kneser-Ney estimator is competitive with respect to this oracle. Consider the case where $n \gg m$. Since the oracle knows there are m columns with non-zero entries, with enough samples, it gets to identify all those columns. Estimating each row of the transition probability then becomes an easier task for the oracle in the sense that the effective support size for each row is $m \ll k$. For example, if there is no observation from a particular row, the oracle will assign unifrom distribution over m non-zero columns, whereas without knowing the fact that the effective support size is m, one needs to assign uniform distribution over $k \gg m$ elements. We show that using Kneser-Ney estimator, the mass we assign to $k-m$ zero columns in each row is negligible and therefore we can compete with the oracle. To show competitivity we only need to bound the mass our estimator assigns to the zero columns, since the oracle does not have any extra knowledge about rest of the elements.

Lemma 68. *If an estimator q for a given row assigns an expected mass η to the zero columns, then*

$$\varepsilon(q) \leq O\left(\log \frac{1}{1-\eta}\right),$$

where ε is the contribution of that row to the competitive loss.

Proof. Since the oracle does not have any extra knowledge about the probabilities except that the support size is m instead of k, we can do as well as the oracle, except for the fact that we are assigning $1-\eta$ mass to the m non-zero columns. Therefore the loss for that particular row will be $O(\log \frac{1}{1-\eta})$. □

Lemma 69 ([FOOP17]). *Let q^{GT} be the Good-Turing estimator applied to estimate the back-off distribution over k elements using n samples and let $\mathbb{E}[\Phi_1]$ be the expected number of elements appearing exactly once in the sample set. Then the expected mass assigned to the unobserved elements in the samples is*

$$\eta \propto \frac{\mathbb{E}[\Phi_1]}{n}$$

Theorem 65. We first bound the amount of mass the Kneser-Ney estimator assigns to the zero-columns. The missing mass assigned to unobserved columns is the same as the mass assigned to the unobserved elements by Good-Turing estimator applied to the vector of the number of distinct elements seen in each column. By Lemma 69 we can bound the missing mass assigned by Kneser-Ney by $\frac{\mathbb{E}[\Phi_1]}{n}$, where $\mathbb{E}[\Phi_1]$ is the expected number of elements appeared once in the vector of column distinct elements. But $\mathbb{E}[\Phi_1] < m$ trivially. Therefore, using Lemma 68 we have $\varepsilon(Q^{\mathrm{KN}}, \mathcal{F}_2) \leq O\left(\log \frac{1}{1-\frac{m}{n}}\right) \leq O\left(\frac{m}{n}\right)$ for every $m \leq k$, and the claim follows by choosing the loosest bound. □

5.A.3 Derivation of Equation (5.10)

Recall the expression of the smoothed Good-Turing estimator from Theorem 66, with the common factors simplified:

$$\check{P}_{xy} = \frac{\sum_j P_{xj}^{\mu+1}(1-P_{xj})^{n_x-\mu}\rho_{bj}}{\sum_j P_{xj}^{\mu}(1-P_{xj})^{n_x-\mu}\rho_{bj}},$$

where $b = \mathsf{nnz}(\mathbf{c})$ and

$$\rho_{bj} = \sum_{S\subset[K]\setminus\{x\} \,:\, |S|=b} \prod_{i\in S}[1-(1-P_{xj})^{n_i}] \prod_{i\in S^{\mathrm{c}}\setminus\{x\}} (1-P_{xj})^{n_i}.$$

Without ρ, this is the non-contextual Good-Turing estimator. In order to characterize the effect of ρ, let us make the following simplifying assumption:

- Let $n_i = m := n/K =$ for all i.
- Let all P_{xj} (at least at contributing values near $1/m$) be approximated by the same p_j.
- We have a power-law p_j, that is $p_j \sim j^{-1/\alpha}$.

Then, we can write:

$$\begin{aligned}
\rho_{bj} &= \sum_{S\subset[K]\setminus\{x\}\,:\,|S|=b} \prod_{i\in S}[1-(1-P_{xj})^{n_i}] \prod_{i\in S^{\mathrm{c}}\setminus\{x\}} (1-P_{xj})^{n_i} && (5.12)\\
&\approx \sum_{S\subset[K]\setminus\{x\}\,:\,|S|=b} \prod_{i\in S}[1-(1-p_j)^{m}] \prod_{i\in S^{\mathrm{c}}\setminus\{x\}} (1-p_j)^{m} && (5.13)\\
&= \binom{K-1}{b}[1-(1-p_j)^m]^b(1-p_j)^{m(K-1-b)} && (5.14)\\
&= \binom{K-1}{b}[(1-p_j)^{-m}-1]^b(1-p_j)^{m(K-1)} && (5.15)\\
&\approx \binom{K-1}{b}(mp_j)^b(1-p_j)^{m(K-1)}
\end{aligned}$$

It follows that:

$$\begin{aligned}
\check{P}_{xy} &\approx \frac{\sum_j p_j^{\mu+b+1}(1-p_j)^{n-\mu}}{\sum_j p_j^{\mu+b}(1-p_j)^{n-\mu}}, && (5.16)\\
&= \frac{\mu+b+1}{n+b+1}\frac{\sum_j \binom{n+b+1}{\mu+b+1} p_j^{\mu+b+1}(1-p_j)^{n-\mu}}{\sum_j \binom{n+b}{\mu+b} p_j^{\mu+b}(1-p_j)^{n-\mu}}, && (5.17)\\
&\sim \frac{\mu+b-\alpha}{n+b} \quad \text{asymptotically, for } \mu+b>0.
\end{aligned}$$

The last asymptotic expression follows from the typical analysis using regular variation / power-laws (see, for example, [OD12a] and [BBO17]). Since the b term in the denominator will be generally negligible compared to n (for power-laws, the *overall* number of distinct symbols grows as n^{α} with $\alpha \in (0,1)$), it is clear that whenever $\mu = 0$ and $b \geq 1$, the mass is distributed proportionally to $b-\alpha$. This doesn't inform about the $b = 0$ case, but by redistributing it uniformly one recovers a single-depth version of Kneser-Ney back-off. More generally, one could iterate this as is done by Chen-Goodman and as generalized in Section 5.5.

5.B Minimax risks

To place the competitivity results in context, we also provide some minimax analysis for the estimation of the conditional probability matrix of size $K \times k$ using n samples. The minimax risk for the class of *all* distributions with K contexts and prediction space size k is:

$$r_n(\Delta_{K,k}) = \min_{Q} \max_{\pi,P} \mathbb{E}_{(x^n,y^n)\sim\pi P} \sum_{i=1}^{K} \pi_i \sum_{j=1}^{k} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$

For the special case of $K = 1$ this definition will be the same as one-dimensional minimax risk (see [KOPS15]). The non-contextual minimax risk has been widely studied and fully resolved in the most general case, namely all *i.i.d.* distributions. In particular, [BS04] showed that for the range $n \gg k$,

$$r_n(\Delta_{1,k}) = \frac{k-1}{2n}\left(1+o(1)\right),$$

and [Pan04] showed that for the range $n \ll k$,

$$r_n(\Delta_{1,k}) = \log \frac{k}{n}.$$

A more refined non-contextual minimax risk was defined in [FOOP17], capturing the dependence of minimax risk based on the number of distinct elements D observed in n samples. For a given n and k, let $\mathcal{P}_d$ be the set of all distributions for which $\mathbb{E}[D] \le d$. Then for some constant c,

$$r_n(\mathcal{P}_d) \le \frac{d}{n} \log \frac{k-\frac{d}{2}}{\frac{d}{2}} + c \cdot \frac{d}{n}$$

Using an estimator for each row independently based on the samples observed from that particular row upper bounds the minimax risk, namely,

$$\begin{aligned}
r_n(\Delta_{K,k}) &\leq \max_{\pi} \mathbb{E}_{x^n \sim \pi} \sum_{i=1}^{K} \pi_i \min_{Q_i} \max_{P_i} \mathbb{E}_{y^n|x^n \sim P_i} \sum_{j=1}^{k} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \\
&\leq \max_{\pi} \mathbb{E}_{x^n \sim \pi} \sum_{i=1}^{K} \pi_i r_{n_i}(1,k) \\
&\overset{(b)}{\leq} \max_{\pi} \mathbb{E}_{x^n \sim \pi} \sum_{i=1}^{K} \pi_i \left(\frac{d_i}{n_i} \log \frac{k - \frac{d_i}{2}}{\frac{d_i}{2}} + c \cdot \frac{d_i}{n_i} \right) \qquad (5.18)
\end{aligned}$$

where d_i is the expected number of distinct elements observed from n_i observations of row i, maximum over all distributions possible for that row. While equation (5.18) is convoluted, it is insightful in the extreme ranges. In the following examples, we elucidate some of these case.

Example 70. *Consider the case where $n \gg K \cdot k$, therefore, in each row, there will be enough samples and $d_i = k$. In this case, based on equation* (5.18)*,*

$$r_n(\Delta_{K,k}) \leq \max_{\pi} \mathbb{E}_{\pi} \sum_{i=1}^{K} \pi_i c \cdot \frac{k}{n_i} \leq O(\frac{K \cdot k}{n})$$

Theorem 71. *For the range $n \gg K * k$,*

$$r_n(\Delta_{K,k}) \geq \Omega \left(\frac{K \cdot k}{n} \right)$$

Proof. We fix π to be uniform, then with high probability each row will have $\frac{n}{K} \pm \sqrt{\frac{n}{K}}$ samples and then the proof follows from the non-contextual distribution estimation lower bound. □

Based on Theorem 71, in the range where $n \gg K \cdot k$, an $O((K \cdot k)/n)$ competitivity gives optimal rates, whereas $O(\sqrt{K/n})$ competitivity implies an optimal estimator, even to the constant.

Example 72. *Consider the case where $n \ll k$, namely, the sample size is much smaller than the alphabet size. Therefore, the worst case in* (5.18) *happens when all d_is are equal to n_is and* (5.18) *gives us the upper bound of*

$$r_n(\Delta_{K,k}) \leq \max_{\pi} \mathbb{E}_{\pi} \sum_{i=1}^{K} \pi_i \log \frac{k}{n_i} \leq O\left(\log \frac{K \cdot k}{n}\right)$$

Theorem 73. *For the range $n \ll k$,*

$$r_n(\Delta_{K,k}) \geq \Omega(\log k)$$

Proof. If $K \gg n$, we don't get to see most of the rows and since for the unobserved rows the minimax risk is $\log k$, we have the Theorem. In the other case, where we get to observe most of the rows, the number of observations from each row is $\leq n \ll k$ and the loss in each row will be $\Omega(\log k)$, hence the Theorem. □

Based on Theorem 73, in the regime where $n \ll k$, an $O(1)$ competitivity is acceptable.

5.C NPLR Algorithm

The nested-recursive version NPLR for applying PLR to smooth N-grams is given by Algorithm 2.

It is worth noting that when all m's are set to 1, NPLR reduces to the nested version of Kneser-Ney back-off suggested by Chen and Goodman (see Section 4.1.6 in [CG99]). It is therefore a strict generalization of that approach.

Further details

Similarly to most local-search algorithms, though its objective function is implicit, multiplicative updates in both PLR and NPLR can benefit from acceleration and noising. We did not include these in the pseudocodes in order not to clutter them, but they are easy to describe.

Algorithm 2 N-gram Partial Low-Rank (NPLR)

```
inputs
    N-gram count matrix $C$, number of iterations $T$
    Persistently maintain $W_z$ and $H_z$ for all subcontexts $z \in [k]^{N-2}$
    Partial ranks $m_2, \cdots, m_N$, and discounts $\alpha_2, \cdots, \alpha_N$ predefined
outputs
    Distribution matrix $Q^{\mathsf{NPLR}}$
for $t = 1$ to $T$ do
    for each subcontext $z \in [k]^{N-2}$ do
        Call PLR on the $z$-restricted bigram of $C$:
            Use $m_N$, $\alpha_N$, and just $T = 1$ iteration
            Unless $t = 1$, initialize with persistent $W_z$ and $H_z$,
        Recover the components $A_z$, $\nu_z$, $W_z$ (create/update persistent copy), and $\tilde{H}_z$
    end for
    if $N > 2$ then
        Call $N - 1$ NPLR with rank $m_{N-1}$ on the combined $(\tilde{H}_z)_{z\in[k]^{N-2}}$ for 1 iteration
        Recover the distribution matrix $(H_z)_{z\in[k]^{N-2}}$ (create/update persistent copy)
    end if
end for
Append and create $A$ from $A_z$, $\nu$ from $\nu_z$, $W$ from $W_z$, and $H$ from $H_z$.
return $Q^{\mathsf{NPLR}} = A + \mathsf{diag}(\nu) \cdot WH$
```

Acceleration can be achieved by moving further in the direction of the update. As for noising, we found that a very effective way to do noising is to periodically reset H by random blocking, just like in the initialization, and then perform several (of the order of the rank) updates of H. This has the effect of bringing H back towards the previous iteration (since W is fixed), while potentially pulling it away from local optima or slowing saddle points.

Another subtle comment is that when NPLR calls PLR, it passes through *soft* counts, which can take values between 0 and 1. A soft count version of absolute discounting was proposes in [FOO16], which can then be used as follows:

$$\hat{A}_{ij} = \begin{cases} (C_{ij} - \alpha)/n_i & C_{ij} \geq 1 \\ C_{ij}(1-\alpha)/n & C_{ij} < 1 \end{cases}$$

which means that the total subtracted missing mass is still $\nu_i = d_i \alpha / n_i$, but where the number of distinct elements now also incorporates the fractional counts:

$$d_i = \sum \mathbf{1}\{C_{ij} \geq 1\} + C_{ij}\mathbf{1}\{C_{ij} < 1\}$$

The only aspect to clarify is that this also requires to modify the indicator matrix B to account for the fractional counts, as follows (one can easily check that this is equivalent to $\mathbf{1}\{C_{ij>0}\}$ when the counts are whole):

$$B_{ij} = (C_{ij} - n_i A_{ij})/\alpha.$$

Finally, it is worth noting that we can ensemble the learned models by averaging. We find that blindly averaging a couple of runs with random restarts performs better than choosing the best by validation.

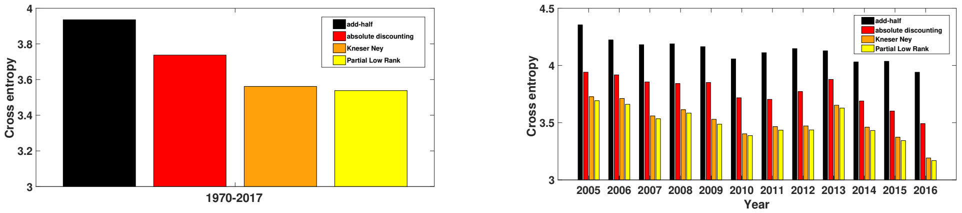
5.D Additional Experiments

5.D.1 Non-temporal data split

We perform the target type prediction, with the difference that the data split between train (used in forming the estimator) and test is by a random permutation and split, rather than temporal. Put differently, we randomly select 83% of the data pairs to be used in forming the estimator and the rest is used for testing. Figures 5.2(b) and 5.2(a) are equivalent to the ones in Figure 5.1 for the non-temporal data split case. The results are consistent with the ones in 5.6.

5.D.2 Joint prediction of target and weapon types on GDT

We repeat our experiments on a slightly modified prediction task. Instead of just predicting the target type for the next attack, we simultaneously predict both the target and weapon types in each city. Weapon type is a field recorded in the GDT and contains information about the weapon



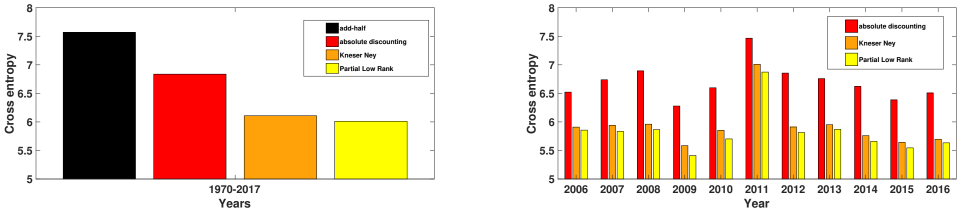
(a) All data

(b) Yearly data

Figure 5.2: Test cross entropy for a non-temporal split of (a) the whole data, and (b) a year's data.

used in an attack. This joint prediction task has a bigger output space (around 3000).

Figures 5.3(a) and 5.3(b) show the benefit of using data from different cities (different rows of the count matrix) when predicting the target and weapon types in a particular city. Similar to the experiments mentioned in Section 5.6, PLR and KN always have significantly better performances in predicting the next incident's attack and weapon types than row-wise estimators such as add-half and absolute discounting.



(a) All data

(b) Yearly data

Figure 5.3: Test cross entropy for the joint prediction of target and weapon types when (a) trained on data prior to 2005 and tested on dates after that, and (b) trained on the first 10 months of a year and tested on the last two months.

Bibliography

[AAJN13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[ADJ$^+$13] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Tight bounds for universal compression of large alphabets. In *ISIT*, pages 2875–2879, 2013.

[ADO12] Jayadev Acharya, Hirakendu Das, and Alon Orlitsky. Tight bounds on profile redundancy and distinguishability. In *NIPS*, 2012.

[ADW13] Armen E. Allahverdyan, Weibing Deng, and Q. A. Wang. Explaining Zipf's law via a mental lexicon. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6), 2013.

[AGMM15] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

[AH99] Lada Adamic and Bernardo A Huberman. The nature of markets in the world wide web. 1999.

[AJOS13a] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal Probability Estimation with Applications to Prediction and Classification. In *COLT*, pages 764–796, 2013.

[AJOS13b] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *COLT*, 2013.

[AJOS14a] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Efficient compression of monotone and m-modal distributions. In *Proceedings of IEEE Symposium on Information Theory*, 2014.

[AJOS14b] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Universal compression of envelope classes: Tight characterization via poisson sampling. *CoRR*, abs/1405.7460, 2014.

[ALY10] José A Adell, Alberto Lekuona, and Yaming Yu. Sharp bounds on the entropy of the poisson law and related quantities. *Information Theory, IEEE Transactions on*, 56(5):2299–2306, 2010.

[AWT91] Naoki Abe, Manfred K Warmuth, and Junichi Takeuchi. Polynomial learnability of probabilistic concepts with respect to the Kullback-Leibler divergence. In *COLT*, 1991.

[BBO17] Anna Ben Hamou, Stéphane Boucheron, and Mesrob I Ohannessian. Concentration Inequalities in the Infinite Urn Scheme for Occupancy Counts and the Missing Mass, with Applications. *Bernoulli*, 2017.

[BGG09] Stéphane Boucheron, Aurelien Garivier, and Elisabeth Gassiat. Coding on countably infinite alphabets. *IEEE Transactions on Information Theory*, 55(1):358–373, 2009.

[BGO14] Stéphane Boucheron, Elisabeth Gassiat, and Mesrob Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *CoRR, abs/1402.6305*, 2014.

[BGO15] Stéphane Boucheron, Elisabeth Gassiat, and Mesrob I Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *IEEE Transactions on Information Theory*, 61(9), 2015.

[BHBO14] Anna Ben-Hamou, Stephane Boucheron, and Mesrob I Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *CoRR*, abs/1412.8652, 2014.

[BJ04] Wray Buntine and Aleks Jakulin. Applying discrete PCA in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 59–66. AUAI Press, 2004.

[BKS15] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*, 2015.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[BPX$^+$07] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. In *EMNLP*, 2007.

[BS04] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[BZ08] Shashi Borade and Lizhong Zheng. Euclidean information theory. In *2008 IEEE International Zurich Seminar on Communications*, pages 14–17. IEEE, 2008.

[CB97] Mark E Crovella and Azer Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *Networking, IEEE/ACM Transactions on*, 5(6):835–846, 1997.

[CG99] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

[CNB17] Ciprian Chelba, Mohammad Norouzi, and Samy Bengio. N-gram language modeling using recurrent neural network estimation. *arXiv preprint arXiv:1703.10724*, 2017.

[Cov91] T.M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E J Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[Dav73] L.D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, Nov. 1973.

[DMPW81] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace. Efficient universal noiseless source codes. *IEEE Transactions on Information Theory*, 27(3):269–279, 1981.

[DYY$^+$19] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[FJO$^+$15] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Universal compression of power-law distributions. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2001–2005. IEEE, 2015.

[FNT16] Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Rediscovery of {Good–Turing} estimators via {B}ayesian nonparametrics. *Biometrics*, 72(1):136–145, 2016.

[FOO16] Moein Falahatgar, Mesrob I Ohannessian, and Alon Orlitsky. Near-Optimal Smoothing of Structured Conditional Probability Matrices. In *NIPS*, pages 4860–4868, 2016.

[FOOP17] Moein Falahatgar, Mesrob I Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati. The power of absolute discounting: all-dimensional distribution estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6660–6669. Curran Associates, Inc., 2017.

[ftSoTtTS18] National Consortium for the Study of Terrorism and Responses to Terrorism (START). Global terrorism database [data file]. retrieved from https://www.start.umd.edu/gtd, 2018. data retrieved from, `https://www.start.umd.edu/gtd`.

[Gar09] A. Garivier. A lower-bound for the maximin redundancy in pattern coding. *Entropy*, 11(4):634–642, 2009.

[GHP$^+$07] Alexander Gnedin, Ben Hansen, Jim Pitman, et al. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv*, 4(146-171):88, 2007.

[GHT$^+$18] Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, pages 1334–1345, 2018.

[GML14] Ankur Gandhe, Florian Metze, and Ian Lane. Neural network language models for low resource languages. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[Goo53] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, pages 237–264, 1953.

[GR44] Beno Gutenberg and Charles F Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944.

[GS95] William A Gale and Geoffrey Sampson. {Good–Turing} frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[HKKV16] Qingqing Huang, Sham M Kakade, Weihao Kong, and Gregory Valiant. Recovering structured probability matrices. *arXiv preprint arXiv:1602.06586*, 2016.

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR*, 1999.

[HOF11] Brian Hutchinson, Mari Ostendorf, and Maryam Fazel. Low rank language models for small training sets. *IEEE Signal Processing Letters*, 18(9):489–492, 2011.

[HOF15] Brian Hutchinson, Mari Ostendorf, and Maryam Fazel. A sparse plus low-rank exponential language model for limited resource scenarios. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):494–504, 2015.

[JM85] Frederick Jelinek and Robert Mercer. Probability distribution estimation from sparse data. *IBM technical disclosure bulletin*, 28:2591–2594, 1985.

[JVS$^+$16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[Kat87] Slava M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, 1987.

[Kie78] J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674–682, Nov. 1978.

[KN95a] Reinhard Kneser and Hermann Ney. Improved Backing-Off for {M}-Gram Language Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Detroit, MI, may 1995.

[KN95b] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE, 1995.

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On Learning Distributions from their Samples. In *COLT*, pages 1066–1100, 2015.

[KT81] R.E. Krichevsky and V.K. Trofimov. The preformance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199–207, Mar. 1981.

[LDMN16] Gary LaFree, Laura Dugan, Erin Miller, and National Consortium for the Study of Terrorism and Responses to Terrorism. Global Terrorism Database, 2016.

[LG14] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *NIPS*, 2014.

[LS01] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mik12] Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80, 2012.

[Mit04] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[MKB$^+$10] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[MKB$^+$11] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Honza Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *ICASSP*, 2011.

[MKS17] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

[MSC^{+}13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[MU05] M. Mitzenmacher and E. Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005.

[NEK94] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

[Neu13] Edward Neuman. Inequalities and bounds for the incomplete gamma function. *Results in Mathematics*, pages 1–6, 2013.

[New05] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[NP00] Pierpaolo Natalini and Biagio Palumbo. Inequalities for the incomplete gamma function. *Mathematical Inequalities & Applications*, 3(1):69–77, 2000.

[OD12a] Mesrob I Ohannessian and Munther A Dahleh. Rare Probability Estimation under Regularly Varying Heavy Tails. In *COLT*, page 21, 2012.

[OD12b] Mesrob I Ohannessian and Munther A Dahleh. Rare probability estimation under regularly varying heavy tails. In *COLT*, 2012.

[OS04] A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, To appear, 2004.

[OS15] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is {Good–Turing} good. In *NIPS*, pages 2143–2151, 2015.

[OSZ03] Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always {Good–Turing}: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

[OSZ04] 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.

[Pan04] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

[Par64] Vilfredo Pareto. *Cours d'économie politique*. Librairie Droz, 1964.

[PBS^{+}60] August Heinrich Petermann, Ernst Behm, Alexander Georg Supan, Paul Max Harry Langhans, Nikolaus Creutzburg, and Hermann Haack. *Petermanns Geographische Mitteilungen*, volume 6. H. Haack, 1860.

[PSDX13] Ankur P Parikh, Avneesh Saluja, Chris Dyer, and Eric P Xing. Language modeling with power low rank ensembles. *arXiv preprint arXiv:1312.7077*, 2013.

[PTRV98] Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *ACM SIGACT-SIGMOD-SIGART*, 1998.

[PY97] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[Ris96] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.

[SAJ04] Nathan Srebro, Noga Alon, and Tommi S Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *NIPS*, 2004.

[Sha04] Gil Shamir. A new upper bound on the redundancy of unknown alphabets. In *CISS, Princeton*, 2004.

[Sha06] G. Shamir. Universal lossless compression with unknown alphabets—the average case. *IEEE Transactions on Information Theory*, 52(11):4915–4944, Nov. 2006.

[Sha13] Gil I. Shamir. Universal source coding for monotonic and fast decaying monotonic distributions. *IEEE Transactions on Information Theory*, 59(11):7194–7211, 2013.

[Sht87] Yurii Mikhailovich Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

[SLE$^+$03] Felisa A Smith, S Kathleen Lyons, S K Ernest, Kate E Jones, Dawn M Kaufman, Tamar Dayan, Pablo A Marquet, James H Brown, and John P Haskell. Body mass of late Quaternary mammals. *Ecology*, 84(12):3403, 2003.

[Sor06] Didier Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Taylor & Francis US, 2006.

[SPC14] Noam Shazeer, Joris Pelemans, and Ciprian Chelba. Skip-gram Language Modeling Using Sparse Non-negative Matrix Probability Estimation. *arXiv preprint arXiv:1412.1454*, 2014.

[SW10] Wojciech Szpankowski and Marcelo J. Weinberger. Minimax redundancy for large alphabets. In *ISIT*, pages 1488–1492, 2010.

[SW12] Wojciech Szpankowski and Marcelo J Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets. *Information Theory, IEEE Transactions on*, 58(7):4094–4104, 2012.

[Szp98] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34(2):142–146, 1998.

[Teh06a] Yee-Whye Teh. A Hierarchical Bayesian Language Model Based on Pitman-Yor processe. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, (July):985–992, 2006.

[Teh06b] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

[TSN18] Sho Takase, Jun Suzuki, and Masaaki Nagata. Direct output connection for a high-rank language model. *arXiv preprint arXiv:1808.10143*, 2018.

[Tsy09] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* Springer series in statistics. Springer, 2009.

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[VV15] Gregory Valiant and Paul Valiant. Instance optimal learning. *arXiv preprint arXiv:1504.05321*, 2015.

[WPM$^+$15] Will Williams, Niranjani Prasad, David Mrva, Tom Ash, and Tony Robinson. Scaling Recurrent Neural Network Language Models. *arXiv preprint arXiv:1502.00512*, 2015.

[WST95] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

[WY22] John C Willis and G Udny Yule. Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, 109(2728):177–179, 1922.

[XB00] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

[XWL$^+$17] Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*, 2017.

[YB13] Xiao Yang and Andrew Barron. Large alphabet coding and prediction through poissonization and tilting. In *The Sixth Workshop on Information Theoretic Methods in Science and Engineering, Tokyo*, 2013.

[YDSC17] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*, 2017.

[YVdS65] P Yu and H Van de Sompel. Networks of scientific papers. *Science*, 169:510–515, 1965.

[Zip35] George Kingsley Zipf. *The psycho-biology of language.* Houghton, Mifflin, 1935.

[Zip13] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

[ZM01] Damian H Zanette and Susanna C Manrubia. Vertical transmission of culture and the distribution of family names. *Physica A: Statistical Mechanics and its Applications*, 295(1):1–8, 2001.

[ZSV14] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[ZYO13] Zhanxing Zhu, Zhirong Yang, and Erkki Oja. Multiplicative updates for learning with stochastic matrices. In *Scandinavian Conference on Image Analysis*, pages 143–152. Springer, 2013.