UNIVERSITY OF CALIFORNIA, SAN DIEGO

Probability estimation and compression involving large alphabets

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

in

Electrical Engineering
(Communication Theory and Systems)

by

Narayana Santhanam

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Mihir Bellare
Professor Russell Impagliazzo
Professor Jack Wolf
Professor Ken Zeger

2006

The dissertation of Narayana Santhanam is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2006

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

The first thanks are for my advisor, Alon Orlitsky. My first meeting with him was on a late Friday evening when I flew into the United States from India. I decided I would see how the campus looked like, and out of curiosity I went up to Alon's office—and found him working there. While I was taken aback to find him at work then, in retrospect, I realize that surprise was completely unwarranted given the dedication he puts into his work. However, from that day till now, time and again, I have found that the effort he puts into work is matched, or surpassed by the effort he has put into his students. It has really been a pleasure working with Alon.

I have worked with some remarkable students—Aldebaro, Junan, Krishna and Nikola from day one, and Anand, In Tae and Sajama later on. Not only was it a lot of fun working with them, it has definitely a learning experience for me.

During the course of my PhD, I have come across several researchers who have not only been involved in varying degrees with me work, but have provided lot of support too. Of them, I would like to mention, in the order I came to know them—Wojciech Szpankowski, who I met thanks to the first publication I had with Alon and Nikola; Gil Shamir, who has since published several excellent papers related to the subject matter of the dissertation; Ken Zeger, who is on my committee too; and Dharmendra Modha, with whom I spent a productive summer in IBM Almaden Research Center.

Thanks also to the several professors I took classes with, not only in UCSD, but my whole career so far. And to all the friends at UCSD without whom the stay would have seemed longer.

Last but not the least, thanks to my family—my parents, brother and grandparents for their support and encouragement.

Chapters 3 and 4 are adapted from A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004.

Chapter 5 is adapted from 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.

Chapter 6 is adapted from N. Jevtić, A. Orlitsky, and N.P. Santhanam. A lower bound on compression of unknown alphabets. Theoretical Computer Science, Feb 2005.

Chapter 8 is adapted from A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427—431, October 17 2003. (See also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003).

Chapter 9 is adapted from A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, July 2006.

Chapter 10 is adapted from A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy of large alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2006.

The author was a primary researcher and author of the sections reproduced here.

VITA

September 8, 1978	Born, Bangalore, India
2000	B. Tech., IIT Chennai
2003	M. S., Department of ECE, University of California San Diego
2006	Ph. D., University of California San Diego

PUBLICATIONS

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, July 2006.

N. Jevtić, A. Orlitsky, and N.P. Santhanam. A lower bound on compression of unknown alphabets. Theoretical Computer Science, Feb 2005.

A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004.

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

A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427—431, October 17 2003. See also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003.

A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy of large alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2006.

D. Modha and N.P. Santhanam. Making the correct mistakes. In *Proceedings of the Data Compression Conference*, 2006.

A. Orlitsky and N.P. Santhanam. On the redundancy of gaussian distributions. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2005.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Convergence of profile based estimators. In *Proceedings of the IEEE Symposium on Information Theory*, 2005.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Innovation and pattern entropy of stationary processes. In *Proceedings of the IEEE Symposium on Information Theory*, 2005.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. In *Information Theory Workshop*, 2004.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Information theoretic approach to modeling low probabilities. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2004.

A. Orlitsky, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang. Practical algorithms for modeling sparse data. Proceedings of the 2004 Proceedings of IEEE Symposium on Information Theory.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J.Zhang. On modeling profiles instead of values. In *Uncertainty in Artificial Intelligence*, 2004.

A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy: A more stringent performance guarantee for universal coding. In *Proceedings of IEEE Symposium on Information Theory*, 2004.

A. Orlitsky, N.P. Santhanam, and J. Zhang. Bounds on compression of unknown alphabets. In *Proceedings of IEEE Symposium on Information Theory*, July 2003.

A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. On compression and modeling of sparse data. In *Third Asian European Workshop on Coding and Information Theory*, June 2003.

A. Orlitsky and N.P. Santhanam. Performance of universal codes over infinite alphabets. In *Proceedings of the Data Compression Conference*, March 2003.

N. Jevtić, A. Orlitsky, and N.P. Santhanam. Universal compression of unknown alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

ABSTRACT OF THE DISSERTATION

Probability estimation and compression involving large alphabets

by

Narayana Santhanam
Doctor of Philosophy in Electrical Engineering
(Communication Theory and Systems)

University of California San Diego, 2006

Professor Alon Orlitsky, Chair

Many results in statistics and information theory are asymptotic in nature, with the implicit assumption that we operate in a regime where the data size is much larger than the alphabet size. In this dissertation, we will be concerned with *large alphabets*, namely alphabets for which the above assumption does not hold.

We consider *universal compression*, *i.e.,* compression when data statistics are unknown, and probability estimation involving data drawn from large alphabets. Both these problems are tackled using a notion of the structure of the string, the string's *pattern*. For example the pattern of "abracadabra" is 12314151231.

It has long been known that universal compression of even independent identically distributed (*i.i.d.*) strings incurs unbounded extra number of bits over the entropy of the source as the alphabet size grows. For such applications, we describe new approaches that isolate the pattern of the string from the dictionary. These approaches are analyzed using results in analysis as well as those on integer partitions studied by Hardy and Ramanujan.

We then consider a related problem of estimating a distribution over a large alphabet using samples drawn from it. The problem considered was posed by a prolific statistician, I.J. Good and Alan Turing. The large alphabet size renders practical sample sizes too small for conventional approaches, hence Good and Turing developed new estimators (without proof) for this problem. The Good-Turing estimator is empirically known to work well.

We use the framework developed for the universal compression problem above and provide an explanation of why the estimator developed by Good and Turing works well and propose other provably optimal variants.

We show that for a large class of processes, the entropy rate of patterns equals the process entropy rate. We also state an asymptotic equipartition property for patterns.

Chapter 1

Introduction

The availability of unprecedented communication, computation, and storage resources has made possible complex systems such as the Internet as well as helped scientific advances like the Human Genome Project. Parallely, to better utilize these advances and to facilitate them, several new problems have come to occupy researchers' efforts. Routing, speech recognition, and data mining are just few of many such applications that spring to mind.

Two aspects of these new problems stand out.

First, a fair number of these problems require solutions for very large *alphabets*. For instance, language models for speech recognition estimate distributions over English words, and thousands of genes are clustered by their expression levels for applications in diagnosis and drug response prediction.

On the other hand, a lot of work in both statistics and information theory is asymptotic in nature. It assumes that we operate in a regime where the data size is much larger than the alphabet size. For the large alphabet problems mentioned above, this is often not the case. We are therefore forced to rework some topics where conventional approaches no longer apply.

Second, problems posed in different contexts may be interconnected. For example, text compression and language modeling for speech recognition, which requires estimation of word probabilities given a text sample. However, while it is folklore that compression and estimation are closely linked, some very commonly used estimators had not even been considered from a compression perspective till recently [7].

We consider two such interconnected large-alphabet problems in this dissertation: universal data compression and probability estimation.

1.1 Universal data compression

While Huffman or Shannon codes achieve compression of data by using the underlying distribution to assign variable length codewords, in most applications, we do not know the underlying distribution. In such situations, the source is usually modeled as an unknown distribution in a collection $\mathcal{P}$ of distributions, *e.g.* the collection of *i.i.d.*, markov, context tree sources, depending on the application [8, 9].

The objective then is to compress the data almost as well as when the distribution is known in advance, namely to find a *universal* compression scheme that performs almost optimally by approaching the entropy no matter which distribution in $\mathcal{P}$ generates the data. The following is a brief introduction to universal compression. Extensive overviews can be found in [10, 11, 12, 13].

Let a source X be distributed over a support set $\mathcal{X}$ according to a probability distribution p. An *encoding* of X is a prefix free 1-1 mapping $\phi : \mathcal{X} \to \{0,1\}^*$. It can be shown that every encoding of X corresponds to a probability assignment q over $\mathcal{X}$ where the number of bits allocated to $x \in \mathcal{X}$ is approximately $\log(1/q(x))$. Roughly speaking, the optimal encoding that is selected based on the distribution $p \in \mathcal{P}$ and achieves its entropy, allocates $\log(1/p(x))$ bits to every $x \in \mathcal{X}$.

The extra number of bits required to encode x when q is used instead of p is therefore

$$\log\frac{1}{q(x)} - \log\frac{1}{p(x)} = \log\frac{p(x)}{q(x)}.$$

The *worst-case redundancy* of q with respect to the distribution $p \in \mathcal{P}$ is

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

the largest number of extra bits allocated for any possible x. The *worst-case redundancy* of q with respect to the collection $\mathcal{P}$ is

$$\hat{R}(\mathcal{P},q) \stackrel{\text{def}}{=} \max_{p\in\mathcal{P}} \hat{R}(p,q),$$

the number of extra bits used for the worst distribution in $\mathcal{P}$ and worst $x \in \mathcal{X}$. The *worst-case redundancy* of $\mathcal{P}$ is

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_q \hat{R}(\mathcal{P}, q) = \min_q \max_{p \in \mathcal{P}} \max_{x \in \mathcal{X}} \log \frac{p(x)}{q(x)}, \tag{1.1}$$

the lowest number of extra bits required in the worst case by any possible encoder q.

For any pair of distributions p and q, $\hat{R}(p, q)$ is non-negative, and therefore $\hat{R}(\mathcal{P})$ is always non-negative. Note that when the redundancy $\hat{R}(\mathcal{P})$ is small, there is an encoding that assigns to every x a probability not much smaller than that assigned to x by the most favorable distribution in $\mathcal{P}$.

Remark For average-case definitions, see [9]. The worst case redundancy is an upper bound on the average case redundancy. □

In most of the problems we consider in the dissertation, the support is a collection of strings of a particular length, say n, which we refer to as the *blocklength.* The encoder is hence a probability distribution on strings of length n. The redundancy will therefore depend on n, and we will be interested in how the redundancy increases with n.

If the redundancy grows $o(n)$, the excess number of bits we use per symbol is asymptotically zero. In such cases, the encoder can effectively compress as well as every source in the collection $\mathcal{P}$.

An important variation of the above problem that we consider in some detail is *sequential* universal compression. Now, the universal encoders for different blocklength cannot be arbitrary. For all $n \geq 1$, the encoders q_n and q_{n+1} on length-n and length-$(n+1)$ strings respectively must satisfy for all strings $x_1 \ldots x_n$,

$$q_n(x_1, \ldots, x_n) = \sum_{x_{n+1}} q_{n+1}(x_1, \ldots, x_n x_{n+1}),$$

namely the marginals must be “consistent”. The question again is how the redundancy of sequential encodings grows with the blocklength.

1.1.1 Large Alphabet compression

The approach of universal compression tackles the problem of not knowing the distribution. But in several applications such as text, speech, and image compression, involve alphabets that are very large, comparable to or even larger than the size of the data sample. Yet the most common universal compression algorithms, such as Lempel-Ziv (LZ) or context-tree weighting (CTW), typically operate on small—usually even binary—alphabets.

To compress a source, these algorithms convert the original signal to a binary string, which they then compress. For example, in text compression, common implementations of both the LZ and CTW algorithms convert words into letters and letters into bits, and then compress the resulting sequence of bits.

Such algorithms risk losing the natural correlation between the source symbols. For example, in the above application, the probability of a word may depend on several previous words, hence on tens of letters, namely hundreds of bits. Most programs truncate their memory at significantly fewer bits.

One reason for this alphabet-size reduction is that the performance guarantees for the algorithms mentioned above implicitly require that the data sample to be compressed is much larger than the alphabet size. Specifically, the redundancy of universal encodings typically increases with the alphabet size.

This phenomenon, first observed by Davisson [9], was studied by Kieffer [11] who showed that even *i.i.d.* distributions over infinite alphabets entail an infinite per-symbol redundancy. Kieffer also provided a necessary and sufficient condition for a collection of sources to have a diminishing per-symbol redundancy.

Recently, there has been renewed interest in universal compression of sources over large alphabets. One line of work [14, 15, 16] follows Elias [17] and considers compression of collections that satisfy Kieffer's condition. Results in this genre typically describe universal algorithms for such collections or find bounds on their redundancy. The most recent results [18] show that all collections satisfying Kieffer's condition can be universally compressed using grammar-based codes.

In this dissertation, we pursue a second direction [19, 20] that separates the description of strings over large alphabets into two parts: description of the symbols appearing in the string, and of the order in which the symbols appear. For example, in text compression, this approach separates the description of the order of the words from the specification of each word's binary representation. The rationale is that the two problems are inherently different, hence best addressed separately.

Results along this line of work typically show [21, 2] that the order of symbols generated by *i.i.d.* distributions over any alphabet, even infinite or unknown, can be compressed with diminishing per-symbol redundancy. We will see how these results can be used [4] to derive asymptotically-optimal probability solutions for the *Good-Turing probability estimation* problem. Related average case results were subsequently been proven in [22, 23].

We start with some results on the standard compression of strings over large alphabets, and then present some results in the latter approach to compression of strings over large alphabets. Specifically, we consider three approaches: the standard compression of the string itself, and two other description methods: *shapes*, where the relative magnitude of the symbols, is conveyed, and *patterns*, where the relative precedence of the symbols in the string is conveyed.

For example, consider the string "abracadabra" over the Roman alphabet. Shapes use an ordering of the alphabet, and in this case, we use $a < b < c \ldots < z$. This approach conveys "abracadabra" using the relative magnitude of the symbols in the string, its *shape*

$$12513141251,$$

followed by the set of letters appearing in the string,

$$\{a, b, c, d, r\}.$$

On the other hand, patterns do not require the alphabet to be ordered. This approach would convey the relative precedence of the symbols in the string, its *pattern*

$$12314151231,$$

followed by the *dictionary*,

$$\{a \to 1, b \to 2, r \to 3, c \to 4, d \to 5\}.$$

Comparing the redundancy of the original strings, their patterns, and their shapes, the three methods display a gradation of redundancy rates.

For standard compression we determine the rate at which the per-symbol redundancy increases to infinity when the alphabet size k grows with the blocklength n.

It is known [10, 24, 25, 26, 27, 28, 29, 30, 31] that when k is fixed and n grows, the per-symbol redundancy diminishes to zero at the rate of $\frac{k-1}{2n}\log n\,(1+o(1))$. We first use techniques in [26] to extend this result and show that when the alphabet size grows with, but slower than, the blocklength, namely $k=o(n)$, the per-symbol redundancy still diminishes to zero, albeit at the lower rate of $\frac{k-1}{2n}\log\frac{n}{k}\,(1+o(1))$. This coincides with a lower bound on average-case redundancy independently proven in [23]. We then show that when k is a constant fraction of the blocklength, namely, $k=\Theta(n)$, the per-symbol redundancy is strictly positive and bounded, and that when k is much larger than the blocklength, namely, $n=o(k)$, the per-symbol redundancy increases to infinity as $\log\frac{k}{n}\,(1+o(1))$.

An observation we make in this dissertation is that most of the redundancy for large alphabets arises from describing just the set of symbols occuring in the string, not their locations within it. At the same time, if the distribution is known, describing the symbols occuring in the string requires negligible number of bits compared to the description of the location of the symbols in the string.

Shapes and patterns abstract the actual symbols occuring, describing their relative locations alone. Therefore, their descriptions may entail potentially lower redundancy when the alphabet size grows, while capturing a substantial amount of information in the string. We show that this is indeed the case.

Unlike standard per-symbol redundancy that increases to infinity with the alphabet size, we show that the maximum per-symbol redundancy of shapes is always between .027 and 1. We also parametrize the upper bound by the number of distinct symbols occuring in the string, tightening it for sources with small alphabets.

While per-symbol standard- and shape-redundancies are strictly positive, the per-symbol redundancy of patterns is at most $\left(\pi\sqrt{2/3}\log e\right)/\sqrt{n}$ regardless of the alphabet size, hence diminishes to zero as the blocklength increases. As with shapes, we also tighten the redundancy bound for sources with small alphabets.

1.2 Distribution Estimation

A related problem is one of distribution estimation. In fact, the framework we will use is one of sequential universal compression. The set of possible distributions that could be in effect is $\mathcal{P}$, and our estimator will be a sequential (online) probability distribution over the same support. This approach is applied in a variety of fields other than universal compression: finance [25], online algorithms, and learning, *e.g.* [32, 33, 34].

To evaluate the performance of an estimator, we apply it not just once but repeatedly to a sequence of elements, all drawn according to the same underlying distribution. Before each element is revealed, we use the estimator to evaluate its conditional probability given the previous elements. Multiplying the conditional probability estimates together, we obtain the probability that the estimator assigns to the whole sequence.

We derive sequential estimators that assign to every sequence a probability that is not much lower than the highest probability assigned to it by any distribution. We therefore define the *sequence attenuation* of an estimator q for a sequence x_1^n to be

$$\hat{A}(q, x_1^n) \overset{\text{def}}{=} \frac{\hat{p}(x_1^n)}{q(x_1^n)},$$

the ratio between the highest probability assigned to x_1^n by any distribution and the probability assigned to it by q. The *worst-case sequence attenuation* of q for patterns of length n is

$$\hat{A}_n(q) \overset{\text{def}}{=} \max_{\psi_1^n \in \Psi^n} \hat{A}(q, \psi_1^n),$$

the largest sequence attenuation of q for any length-n pattern. Note that $(\hat{A}_n(q))^{1/n}$ is the *worst-case symbol attenuation* of q for patterns of length n, namely, the largest possible ratio between the *per-symbol* probability assigned by any distribution to symbols of length-n patterns and the corresponding probability assigned by q. Finally, the *(asymptotic, worst-case, symbol) attenuation* of q is

$$\hat{A}^*(q) \overset{\text{def}}{=} \limsup_{n\to\infty} \left(\hat{A}_n(q)\right)^{1/n},$$

the largest possible ratio between the per-symbol probability assigned to any asymptotically long pattern by any distribution and the corresponding distribution assigned by q.

Recall that the above definitions are exactly analogous to redundancy definitions in Section 1.1. Correspondingly, the attenuation of any estimator is always at least one. Attenuation of a constant $c > 1$ implies that the estimator assigns to each n-symbol sequence a probability which is at most a factor of c^n lower than its best probability. Attenuation of one, which we call *diminishing attenuation* in analogy with diminishing per symbol redundancy, implies that the estimator assigns to each sequence a probability that is at most sub-exponentially smaller than the best possible. Hence the per-symbol probability assigned by the estimator would be asymptotically the best possible.

In using the above framework, a natural objection that may arise is that while estimating distributions overestimation is also undesirable, not just underestimation. It is however, easy to see and explained in Chapter 2, that overestimation by a large factor will not happen with high probability, no matter what the underlying distribution or the estimator is.

1.2.1 Large alphabet distribution estimation

In your next safari, say you observe a random sample of African animals. You find 3 giraffes, 1 zebra, and 2 elephants. How would you estimate the probability of the various species you may encounter on your trip?

A naive, *empirical-frequency*, estimator may assign probability 1/2 to giraffes, 1/6 to zebras, and 1/3 to elephants. But the poor estimator will be completely unprepared for an encounter with an offended lion.

To address this unseen-elements problem, Laplace [35] proposed adding one to the count of each species, including to the collection of unseen ones, thereby assigning probability $(3 + 1)/10 = 0.4$ to giraffes, $(1 + 1)/10 = 0.2$ to zebras, $(2 + 1)/10 = 0.3$ to elephants, and $(0 + 1)/10 = 0.1$ to unseen species.

The *Laplace* and other *add-constant* estimators have since been applied and studied extensively. In particular, the *add half*, or *Krichevski-Trofimov* [24], estimator was shown to possess certain optimality properties when the number of possible elements is fixed and the sample size increases to infinity [36, 37].

However, when the number of possible elements is large compared to the sample size, add-constant estimators are lacking too [38]. Suppose that during your safari trip you evaluate the distribution of animals' DNA sequences. You observe a large number n of animals and, predictably, find that each has a unique DNA sequence. You therefore have a sample of n sequences, each observed once, from which you would like to estimate the distribution of all sequences. An add-c estimator would assign probability $(1+c)/(n+nc+c)$ to each observed sequence and probability $c/(n+nc+c)$ to all unseen ones. It follows that the probability assigned to all observed sequences, $(n+nc)/(n+nc+c) \approx 1$, while that assigned to all unseen sequences is close to zero. Clearly, the opposite would be a better model.

Good and Turing encountered this problem while trying to break the Enigma cipher during World War II [39]. The British intelligence was in possession of the *Kenngruppenbuch*, the German cipher book that contained all possible secret keys, and used previously decrypted messages to document the page numbers of keys used by various U-boat commanders. They wanted to use this information to estimate the distributions of pages that each U-boat commander picked secret keys from.

Good and Turing came up with a surprising estimator that bears little resemblance to either the empirical-frequency or the add-constant estimators above. After the war, Good published the estimator [40] mentioning that Turing had an "intuitive demonstration" for it, but not describing what this intuition was.

The *Good-Turing estimator* has since been incorporated into a variety of applications such as information retrieval [41], spelling correction [42], word-sense disambiguation [43], and speech recognition, *e.g.* [44] where it is applied to estimate the probability distribution of words.

While the Good-Turing estimator performs well in general, it is suboptimal for elements that appear frequently, hence was modified in subsequent estimators, *e.g.* the Jelinek-Mercer, Katz, Witten-Bell, and Kneser-Ney estimators [44].

On the theoretical side, interpretations of the Good-Turing estimator have been proposed [45, 46, 47], and its convergence rate was analyzed [48, 49]. Yet, lacking a measure for assessing the performance of an estimator, no objective evaluation or optimality results for the Good-Turing estimator have been established.

Note that all the estimators above assign probabilities that the next element is one of the elements that has appeared before, and assign a probability that the next element is hitherto unseen. This represents a simplification from estimating the distribution over the whole support at each step, and we will see in Chapter 8 that this approach is analogous to doing sequential pattern compression.

Our objective is then to evaluate the performance of existing estimators and to construct diminishing-attenuation variants of Good-Turing like estimators when the alphabets could be infinite. We show that add-constant estimators have infinite attenuation. On the other hand, the Good-Turing estimator performs well in the sense that its attenuation is low, however for some sequences it assigns a probability that is exponentially smaller than the best possible, namely its attenuation is strictly above one.

We construct estimators over unknown, potentially large, even infinite, alphabets, by abstracting the actual symbols that appear in the sequence and considering only their pattern. We derive two diminishing-attenuation estimators. The first is computationally more efficient and requires only a constant number of operations per symbol. The second is more complex but its attenuation approaches one faster. To determine the estimators' attenuations we use potential functions and results of Hardy and Ramanujan [50] on the number of partitions of an integer.

1.3 Entropy rate and AEP for patterns

The universal compression results mentioned above [21, 2] show that patterns of strings generated by *i.i.d.* distributions over any alphabet, even infinite or unknown, can be compressed with diminishing per-symbol redundancy. These results have were used [4] to derive asymptotically-optimal solutions for the Good-Turing probability estimation problem. Related average case results have subsequently been proven [23].

It is therefore natural to consider the entropy of patterns themselves. Observe that the mutual information between sequences and patterns is

$$I(X, \Psi) = H(\Psi) - H(\Psi|X) = H(\Psi),$$

the entropy of patterns. Therefore the pattern entropy is the information about the sequence contained in the patterns. Shamir and Song [22, 51], bounded the entropy of patterns of *i.i.d.* distributions in terms of the source entropy and alphabet size. See also discussion after (9.10).

In the dissertation we determine the entropy rate of patterns of a large class of processes, and for *i.i.d.* processes, we bound the speed at which the per-symbol pattern entropy converges to this rate, and show that patterns satisfy an asymptotic equipartition property. To derive some of these results, we upper bound the probability that the n'th variable in a random process differs from all preceding ones. We note that related entropy-rate results were independently derived by Gemelos and Weissman [52, 53], and that subsequent results appeared in [54, 55, 56].

For finite alphabets, namely $|\mathcal{A}| = k < \infty$, it is easy to see that the pattern entropy rate should equal that of the distribution, since the pattern $\overline{\psi}$ is determined by the sequence X and can derive from at most $k!$ sequences.

Clearly, the above argument is too simplistic to generalize to large, or infinite distributions, however the above result does extend to a all finite entropy discrete stationary processes.

Equivalently, this result also asymptotically bounds the average number of bits needed to describe a sequence given the pattern because

$$H(X_1,\ldots,X_n) = H(\Psi_1,\ldots,\Psi_n) + H(X_1,\ldots,X_n|\Psi_1,\ldots,\Psi_n).$$

Thus, for discrete processes, the above result implies that the entropy rate of sequences given the pattern is zero.

In effect, an interpretation of the above result is that patterns contain effectively all the information in the sequences.

It is not possible to obtain a uniform rate at which $\frac{1}{n}H(\Psi_1,\ldots,\Psi_n)$ converges to H for all finite entropy distributions. Nor is it possible to obtain a rate that just depends on the entropy and blocklength. Instead we obtain the convergence rate in terms of the variance of the codelengths of the symbols in the alphabet [57].

Furthermore, we also derive an asymptotic equipartition property for *i.i.d.* induced distributions on patterns in [58].

Chapter 2

Universal compression: preliminaries

We outline several universal compression results that form the intuitive background of the chapters to follow. We derive an expression (*Shtarkov's sum* for worst case redundancy, outline a simple but general technique to upper bound worst case redundancy and show that sequential universal compression incurs at most a small additional redundancy over block compression. We then show that universal compression schemes with diminishing redundancy correspond to good probability estimators.

2.1 Shtarkov's sum

To evaluate the redundancies, we will frequently use a result by Shtarkov, showing that [59] the distribution achieving $\hat{R}(\mathcal{P})$ in Equation (1.1) is

$$q^*(x) = \frac{\sup_{p\in\mathcal{P}} p(x)}{\sum_{x\in\mathcal{X}} \sup_{p\in\mathcal{P}} p(x)}.$$

It follows that the redundancy of a collection $\mathcal{P}$ of distributions over $\mathcal{X}$ is determined by *Shtarkov's sum*,

$$\hat{R}(\mathcal{P}) = \log\left(\sum_{x\in\mathcal{X}} \sup_{p\in\mathcal{P}} p(x)\right). \tag{2.1}$$

2.1.1 A general upper bound

Let $\mathcal{P}$ be a collection of distributions. The following bound on the redundancy of $\mathcal{P}$ is easily obtained.

Lemma 1. For all $\mathcal{P}$,

$$\hat{R}(\mathcal{P}) \le \log |\mathcal{P}| .$$

Proof The claim is obvious when $\mathcal{P}$ is infinite, and for finite $\mathcal{P}$, Shtarkov's sum implies that

$$\begin{aligned}
\hat{R}(\mathcal{P}) &= \log \sum_{x \in \mathcal{X}} \sup_{p \in \mathcal{P}} p(x) \\
&\le \log \sum_{p \in \mathcal{P}} \sum_{\substack{x \in \mathcal{X}: \\ \sup_{p' \in \mathcal{P}} p'(x) = p(x)}} p(x) \\
&\le \log \sum_{p \in \mathcal{P}} 1 \\
&= \log |\mathcal{P}| . \qquad \square
\end{aligned}$$

Intuitively, for finite $\mathcal{P}$, the lemma corresponds to first identifying the maximum-likelihood distribution of x from all distributions in $\mathcal{P}$ and then describing x using this distribution. If not all distributions are candidates for maximum-likelihood distributions, the above bound can be improved as follows.

A collection $\hat{\mathcal{P}}$ of distributions dominates $\mathcal{P}$ if for all $x \in \mathcal{X}$,

$$\sup_{p \in \hat{\mathcal{P}}} p(x) \ge \sup_{p \in \mathcal{P}} p(x),$$

namely, the highest probability of any $x \in \mathcal{X}$ in $\hat{\mathcal{P}}$ is at least as high as that in $\mathcal{P}$. The next lemma then follows immediately from Shtarkov's sum.

Lemma 2. If $\hat{\mathcal{P}}$ dominates $\mathcal{P}$, then

$$\hat{R}(\mathcal{P}) \le \hat{R}(\hat{\mathcal{P}}). \qquad \square$$

The above lemmas imply that the redundancy is upper bounded by the logarithm of the size of $\hat{\mathcal{P}}$.

Corollary 3. If $\hat{\mathcal{P}}$ dominates $\mathcal{P}$, then

$$\hat{R}(\mathcal{P}) \leq \log \left|\hat{\mathcal{P}}\right|. \qquad \square$$

To illustrate this bound, we bound the standard redundancy of *i.i.d.* strings over finite alphabets.

Example 1. Consider the collection $\mathcal{I}_2^n$ of all *i.i.d.* distributions over length-n strings drawn from an alphabet of size 2, which, without loss of generality, we assume to be $\{0,1\}$. Clearly,

$$\hat{\mathcal{I}_2^n} = \left\{p \in \mathcal{I}_2^n : p(0) = \frac{k}{n} \text{ where } 0 \leq k \leq n \text{ and } k \in \mathbb{Z}\right\},$$

dominates $\mathcal{I}_2^n$, hence from Corollary 3,

$$\hat{R}(\mathcal{I}_2^n) \leq \log \left|\hat{\mathcal{I}_2^n}\right| = \log(n+1).$$

Similarly, since

$$\left|\hat{\mathcal{I}_m^n}\right| = \binom{n+k-1}{k-1},$$

it follows that for every k and n,

$$\hat{R}(I_k^n) \leq (k-1)\log\left(e \cdot \frac{n+k-1}{k-1}\right). \qquad \square$$

2.2 Overestimation

We adopted the universal compression framework for probability estimation in Chapter 1. Universal compression forces the codelengths to be small, namely it attempts to prevent underestimation of probabilities, but probability estimation should not allow for overestimation either.

The following result justifies that overestimation is not a serious issue.

Lemma 4. For all distributions p and q over $\mathcal{X}$.

$$p\{x : q(x) \geq Ap(x)\} \leq \frac{1}{A}.$$

Proof The result follows since

$$1 \geq q\{x : q(x) \geq Ap(x)\} \geq Ap\{x : q(x) \geq Ap(x)\}. \qquad \square$$

Suppose a universal scheme compresses a collection of distributions $\mathcal{P}$ over $\mathcal{X}$ with redundancy R, and therefore attenuation $A = 2^R$. It follows that with probability $\geq 1 - \frac{1}{2^R}$,

$$\left|\log \frac{p(X)}{q(X)}\right| \leq R.$$

where X is a random variable distributed according to p.

Typically we consider the support to be strings of length n. Therefore, for all sources in the collection, with probability $\geq 1 - \frac{1}{2^{R_n}}$,

$$\frac{1}{n}\left|\log \frac{p(X^n)}{q(X^n)}\right| \leq \frac{R_n}{n}. \tag{2.2}$$

Therefore, if the per symbol redundancy, $R_n/n \to 0$, then

$$\frac{1}{n}\left|\log \frac{p(X^n)}{q(X^n)}\right| \to 0$$

in probability. The above statement follows directly if $R_n \to \infty$, and it is easy to see that the above statement is true even if R_n is bounded.

Therefore, if a collection of sources incurs diminishing per-symbol redundancy, any source in the collection can be estimated well using a scheme with diminishing per-symbol redundancy.

2.3 Sequential universal compression

As mentioned in Chapter 1, in this setting, we force the encodings over different blocklengths to be consistent with each other, namely for all $n \geq 1$, if q_n and q_{n+1} are encodings on length-n and length-$(n+1)$ strings respectively, they must satisfy for all strings $x_1 \ldots x_n$,

$$q_n(x_1, \ldots, x_n) = \sum_{x_{n+1}} q_{n+1}(x_1, \ldots, x_n x_{n+1}).$$

The question now is how the redundancy of sequential encodings grows with the blocklength. More precisely, let $\overline{q} = q_1, q_2, \ldots$ be a sequence of distributions over $\mathcal{X}, \mathcal{X}^2, \ldots$ respectively. Let

$$\hat{R}(\overline{q}, n) \stackrel{\text{def}}{=} \sup_{p \in \mathcal{P}^n} \max_{\overline{x} \in \mathcal{X}^n} \log \frac{p(\overline{x})}{q_n(\overline{x})},$$

be the redundancy of q_n, where $\mathcal{P}^n$ is the collection of distributions on length n strings.

Note that

$$\hat{R}(\overline{q}, n) \geq \hat{R}(\mathcal{P}_n).$$

This lower bound can be achieved for every n.

Theorem 5. For all n, $\exists \overline{q}$ such that

$$\hat{R}(\overline{q}, n) = \hat{R}(\mathcal{P}_n).$$

Proof For all $x_1, \ldots, x_n = x^n \in \mathcal{X}^n$, let

$$q^*(x^n) = \frac{\sup_{p \in \mathcal{P}} p(x^n)}{\sum_{x^n \in \mathcal{X}^n} \sup_{p \in \mathcal{P}} p(x^n)}$$

be the worst case optimal encoder for length-n strings. For $j \leq n$, let

$$q_j(x^j) = \sum_{x_{j+1}, \ldots, x_n} q^*(x^n).$$

Assume without loss of generality that $1 \in \mathcal{X}$. For $j > n$, let

$$q_j(x^j) = \begin{cases} q^*(x^n) & x_{n+1} = \ldots = x_j = 1, \\ 0 & \text{else.} \end{cases}$$

Clearly $\overline{q} = q_1, q_2, \ldots, q_n, \ldots$ corresponds to a sequential encoding, and by construction $\hat{R}(\overline{q}, n) = \hat{R}(\mathcal{P}_n)$. □

We construct $\overline{q}$ that almost matches the lower bound in Theorem 5

Theorem 6. $\exists \overline{q}$ such that for all n,

$$\hat{R}(\overline{q}, n) \leq \hat{R}(\mathcal{P}_n) + 2 \log n + \log \frac{\pi^2}{6}.$$

Proof Let $\overline{q}_n$ correspond to the sequential encoding described in Theorem 5 satisfying

$$\hat{R}(\overline{q}_n, n) = \hat{R}(\mathcal{P}_n).$$

Consider the linear weighting

$$\overline{q} \stackrel{\text{def}}{=} \frac{6}{\pi^2} \sum_{i \geq 1} \frac{\overline{q}_i}{i^2}.$$

Clearly, $\overline{q}$ also corresponds to a sequential encoding since each of $\overline{q}_i$ does so. Further, for all $n \geq 1$ and all $x^n \in \mathcal{X}^n$,

$$\overline{q}(x^n) \geq \frac{6}{\pi^2} \frac{\overline{q}_n}{n^2},$$

which implies the theorem. □

2.4 Mathematical preliminaries

We approximate binomial and multinomial coefficients that will be encountered several times in the dissertation,

2.4.1 Approximation of binomial coefficients

While finite-alphabet results typically involve binomial coefficients of form $\binom{n}{\alpha n}$ for some constant α, large alphabets often require the calculation of $\binom{n}{o(n)}$. The following lemma provides a convenient approximation.

Lemma 7. When $m \to \infty$, and $m = \mathcal{O}(\sqrt{n})$,

$$\binom{n}{m} = \Theta\Big(\frac{e\,n}{m}\Big)^m,$$

and when, in addition, $m = o(\sqrt{n})$,

$$\binom{n}{m} = \frac{1}{\sqrt{2\pi m}}\Big(\frac{e\,n}{m}\Big)^m (1 + o(1)).$$

Proof Feller's bounds on Stirling's approximation [60] state that for every $n \ge 1$,

$$\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \le n! \le \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n}}. \tag{2.3}$$

Hence for all $m \le n$,

$$\begin{aligned}&\frac{e^{-\frac{1}{12}\left(\frac{1}{m}+\frac{1}{(n-m)}\right)}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\Big(\frac{n}{m}\Big)^m\Big(\frac{n}{n-m}\Big)^{n-m}\\ &\le \binom{n}{m} \le \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\Big(\frac{n}{m}\Big)^m\Big(\frac{n}{n-m}\Big)^{n-m}.\end{aligned}$$

Taking derivatives, it is easy to see that for all $x \ge 0$,

$$e^{x-x^2/2} \le 1 + x \le e^x,$$

hence for all $m \le n$,

$$\Big(\frac{n}{n-m}\Big)^{n-m} = \Big(1 + \frac{m}{n-m}\Big)^{n-m} \le \Big(e^{\frac{m}{n-m}}\Big)^{n-m} = e^m,$$

and

$$\begin{aligned}\left(\frac{n}{n-m}\right)^{n-m} &= \left(1+\frac{m}{n-m}\right)^{n-m}\\ &\geq \exp\left[\left(\frac{m}{n-m}-\frac{1}{2}\left(\frac{m}{n-m}\right)^2\right)(n-m)\right]\\ &= \frac{e^m}{e^{\frac{1}{2}\frac{m^2}{n-m}}}.\end{aligned}$$

Therefore for all $m \leq n$,

$$\begin{aligned}&\frac{1}{C\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{e\,n}{m}\right)^m\\ &\quad\leq \binom{n}{m} \leq \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{m(n-m)}}\left(\frac{e\,n}{m}\right)^m,\end{aligned}$$

where

$$C = \exp\left(\frac{1}{12m}+\frac{1}{12(n-m)}+\frac{1}{2}\frac{m^2}{n-m}\right),$$

proving the first part of the lemma. When $m \to \infty$ and $m = o(\sqrt{n})$, $C = 1 + o(1)$, and the second part follows. □

2.4.2 Approximation of multinomial coefficients

Recall Feller's bounds [60] for all $n \geq 1$

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

They imply that for all $0 < m < n$, the binomial coefficients are bounded by

$$\begin{aligned}e^{-\frac{1}{12}\left(\frac{1}{m}+\frac{1}{n-m}\right)} \cdot \sqrt{\frac{n}{2\pi m(n-m)}}\cdot 2^{nh\left(\frac{m}{n}\right)}&\\ \leq \binom{n}{m} \leq&\\ e^{\frac{1}{12n}} \cdot \sqrt{\frac{n}{2\pi m(n-m)}}\cdot 2^{nh\left(\frac{m}{n}\right)}&\end{aligned} \tag{2.4}$$

where h is the binary entropy function

$$h(x) = x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x}.$$

Furthermore, for all l and $1 \le u_1, \ldots ,u_l \le n$ with $\sum_i u_i = n$

$$\begin{aligned} & e^{-\frac{l}{12}} \sqrt{\left(\frac{l}{n}\right)^l \cdot \frac{n}{(2\pi)^{l-1}}} \\ \le & \binom{n}{u_1, \ldots ,u_l} \prod_{i=1}^{l} \left(\frac{u_i}{n}\right)^{u_i} \le \\ & e^{\frac{1}{12n}} \sqrt{\frac{1}{n-l+1} \cdot \frac{n}{(2\pi)^{l-1}}} \end{aligned} \tag{2.5}$$

where, in addition to Feller's bounds, the lower bound uses the arithmetic/geometric mean inequality and the upper bound uses the fact that

$$\prod_{i=1}^{l} u_i \ge n - l + 1.$$

2.4.3 The Gamma function

Shtarkov's sum for the collection $\mathcal{I}_m^n$ of *i.i.d.* sources involves terms of the form

$$\sum_{\overline{u} \in \mathcal{U}_l^n} \binom{n}{u_1, \ldots ,u_l} \prod_{j=1}^{l} \left(\frac{u_j}{n}\right)^{u_j},$$

which Feller's bounds (2.3) upper bound by

$$e^{\frac{1}{12n}} \left(\frac{1}{2\pi}\right)^{\frac{l-1}{2}} \sum_{\overline{u} \in \mathcal{U}_l^n} \sqrt{\frac{n}{u_1 \ldots u_l}}.$$

For fixed l, this upper bound is tight as $n \to \infty$

$$\sum_{\overline{u} \in \mathcal{U}_l^n} \sqrt{\frac{n}{u_1 \ldots u_l}} \sim n^{\frac{l}{2}} \int_{\substack{x_i \ge 0; \\ \sum_{x_i} = 1}} \frac{dx_1 \ldots dx_l}{\sqrt{x_1 \ldots x_l}} = n^{\frac{l}{2}} \frac{\Gamma^l(\frac{1}{2})}{\Gamma(\frac{l}{2})},$$

where the *Gamma function* is defined for all $z \in \mathbb{C}$ by

$$\Gamma(z) \overset{\text{def}}{=} \int_0^\infty x^{z-1} e^{-x} dx.$$

Some values of the Gamma function are well known. For example, $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(1) = 1$, and simple calculation shows that for all z

$$\Gamma(z+1) = z\Gamma(z) \tag{2.6}$$

hence, $\Gamma(z+1)$ generalizes the factorial function to complex numbers. In particular, for all positive real arguments, it satisfies Feller's bounds (2.3)

$$\sqrt{2\pi z}\left(\frac{z}{e}\right)^{z} \leq \Gamma(z+1) \leq \sqrt{2\pi z}\left(\frac{z}{e}\right)^{z} e^{\frac{1}{12z}}. \tag{2.7}$$

Chapter 3

Large alphabets

The aysmptotic redundancy of *i.i.d.* sequences when the alphabet size k is finite and the blocklength n tends to infinity has been studied by several researchers. It was shown, *e.g.* in [] that

$$\hat{R}(I_k^n) = \frac{k-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} + o_k(1),$$

and therefore determined the term that grows with n, and the constant term that depends on the alphabet size.

However, many applications do not reside in asymptopia. Often, n is not arbitrarily larger than k, and the asymptotic results do not apply. We now consider the redundancy for more general relations between n and k and show that in many cases, the constant term actually dominates the redundancy.

We show that when $k = o(n)$, namely the blocklength is much larger than the alphabet, the per symbol redundancy diminishes to zero as $\frac{k}{n}\log\frac{n}{k}$, that when $n = \Theta(k)$, namely the blocklength is proportional to the alphabet size, the per-symbol redundancy is a positive constant that we determine to within a factor of eight, and that when $n = o(k)$, namely the blocklength is much smaller than the alphabet, the per-symbol redundancy increases to infinity as $\log\frac{k}{n}$.

To prove these results we relate types of strings to various partitions of integers.

3.1 Ordered partitions of integers

A *positive ordered partition* of $n \in \mathbb{Z}^+$ is a tupple of *parts*, positive integers summing to n. The set of positive ordered partitions of n into m parts is denoted by $\mathcal{U}_m^n$. For example

$$\mathcal{U}_3^5 = \{(1,1,3),(1,3,1),(3,1,1),(1,2,2),(2,1,2),(2,2,1)\}.$$

Any ordered partition $(u_1, u_2, \ldots, u_m) \in \mathcal{U}_m^n$ can be represented as a linear diagram of n dots and $m-1$ vertical bars where the first bar follows the left u_1 dots, the second bar follows the next u_2 dots, and so on till the $(m-1)$-st bar follows u_{m-1} dots and precedes the final u_m dots. For example, the diagrams corresponding to the partitions of $\mathcal{U}_3^5$ are

$$\cdot|\cdot|\cdot\cdot\cdot \qquad \cdot|\cdot\cdot\cdot|\cdot \qquad \cdot\cdot\cdot|\cdot|\cdot \qquad \cdot|\cdot\cdot|\cdot\cdot \qquad \cdot\cdot|\cdot|\cdot\cdot \qquad \cdot\cdot|\cdot\cdot|\cdot$$

Any positive ordered partition corresponds to a linear diagram of n dots with $m-1$ of the spaces between them marked with single bars. Conversely, every such diagram corresponds to to a positive ordered partition. Hence, for all $m, n \geq 1$,

$$|\mathcal{U}_m^n| = \binom{n-1}{m-1}. \tag{3.1}$$

For example, as we saw above,

$$|\mathcal{U}_3^5| = \binom{5-1}{3-1} = \binom{4}{2} = 6.$$

While not needed now, $\mathcal{U}^n$ denotes the set of all ordered partitions of n, into any number of parts. It follows from (3.1) (or can be shown directly) that for every n,

$$|\mathcal{U}^n| = 2^{n-1}. \tag{3.2}$$

Similar to positive ordered partitions, a *nonnegative ordered partition* of $n \in \mathbb{Z}^+$ is a tupple of *nonnegative* integers, again called parts, summing to n. The set of nonnegative ordered compositions of n into k parts is denoted by $\mathcal{T}_k^n$. For example

$$\mathcal{T}_3^2 = \{(2,0,0),(0,2,0),(0,0,2),(1,1,0),(1,0,1),(0,1,1)\}.$$

As with their positive counterparts, any nonnegative ordered partition $(n_1, n_2, \ldots, n_k) \in \mathcal{T}_k^n$ can be represented as a linear diagram of n dots and $k-1$ vertical bars where the first bar follows the left n_1 dots, the second bar follows the next n_2 dots, and so on till the $(k-1)$-st bar follows n_{k-1} dots and precedes the final n_k dots. For example, the diagrams corresponding to the partitions of $\mathcal{T}_3^2$ are

$$\cdot\cdot|| \qquad |\cdot\cdot| \qquad ||\cdot\cdot \qquad \cdot|\cdot| \qquad \cdot||\cdot \qquad |\cdot|\cdot$$

As with their positive counterparts, the diagram of a nonnegative ordered partition of n into k parts consists of n dots and $k-1$ bars, but here the diagrams may have multiple bars between adjacent dots and at the diagram's left or right ends, hence any ordering of the n dots and $k-1$ bars is now possible. The diagram corresponding to a nonnegative ordered partition is just a sequence of $n+(k-1)$ dots and bars where the $k-1$ bars can appear in any location, and therefore for all $n \geq 0$ and $k \geq 1$,

$$|\mathcal{T}_k^n| = \binom{n+k-1}{k-1}. \tag{3.3}$$

For example, as we found above,

$$|\mathcal{T}_3^2| = \binom{2+3-1}{3-1} = \binom{4}{2} = 6.$$

Positive and nonnegative ordered partitions are clearly closely related. Let $\mathcal{T}_{k,m}^n$ denote the set of nonnegative ordered partitions of n into k parts, m of which are non-zero. For example,

$$\mathcal{T}_{3,1}^2 = \{(2,0,0),(0,2,0),(0,0,2)\} \quad \text{and} \quad \mathcal{T}_{3,2}^2 = \{(1,1,0),(1,0,1),(0,1,1)\}.$$

Since the number of non-zero parts is positive and cannot exceed n or k,

$$\mathcal{T}_k^n = \bigcup_{m=1}^{\min(n,k)} \mathcal{T}_{k,m}^n,$$

and as the sets $\mathcal{T}_{k,m}^n$ are disjoint,

$$|\mathcal{T}_k^n| = \sum_{m=1}^{\min(n,k)} |\mathcal{T}_{k,m}^n|.$$

On the other hand, every partition in $\mathcal{T}_{k,m}^n$ can be specified by first describing which m of the k parts are nonnegative, followed by a description of these m positive values, namely of a partition of n into m parts. Hence,

$$|\mathcal{T}_{k,m}^n| = \binom{k}{m}|\mathcal{U}_m^n|.$$

Therefore,

$$|\mathcal{T}_k^n| = \sum_{m=1}^{\min\{n,k\}} \binom{k}{m} |\mathcal{U}_m^n|, \tag{3.4}$$

incidentally, explaining the combinatorial identity

$$\binom{n+k-1}{k-1} = \sum_{m=1}^{\min\{n,k\}} \binom{k}{m}\binom{n-1}{m-1}. \tag{3.5}$$

3.2 Upper bound

Considering the types of sequences and applying Shtarkov's sum, we derive a simple upper bound on $\hat{R}(I_k^n)$ that is always tight up to a factor of roughly eight, asymptotically tight when $n = o(k)$, and tight up to a factor of two when $k = o(n)$. We then derive another upper bound that is asymptotically tight in the latter regime.

Recall that the multiplicity μ_j of a symbol j is its number of occurances in a sequence. The *type* of a sequence $\overline{x}$ over $[k]$ is the k-tuple of multiplicities

$$\tau(\overline{x}) \stackrel{\text{def}}{=} (\mu_1, \ldots, \mu_k).$$

For example, over $[6]$,

$$\tau(51535) = (1, 0, 1, 0, 3, 0)$$

as 1 and 3 appear once, 5 appears thrice, and 2, 4, and 6 do not appear. Any sequence in $[k]^n$ has k nonnegative multiplicities summing to n, hence we identify its type with a nonnegative ordered partition of n into k parts, and let $\mathcal{T}_k^n$ denote also the set of types of $[k]^n$ sequences. From (3.3),

$$|\mathcal{T}_k^n| = \binom{n+k-1}{k-1}.$$

A basic property of types is that every *i.i.d.* distribution assigns the same probability to all sequences of the same type. Therefore the same distribution maximizes the probability of all sequences of the same type, and hence the probability of the type itself. It follows that the highest probability of a type is the sum of the highest probabilities of its sequences, namely, for every type τ,

$$\hat{p}(\tau) = \sum_{\overline{x} \in \tau} \hat{p}(\overline{x}) \tag{3.6}$$

where $\overline{x} \in \tau$ denotes $\tau(\overline{x}) = \tau$.

For example the type $(1,2)$ over $[2]$ is the set $\{122, 212, 221\}$. For any *i.i.d.* distribution, $p(122) = p(212) = p(221)$, hence the same distribution, in this case $(1/3, 2/3)$, maximizes the probability of all three sequences in the type, and of the type $\{011, 101, 110\}$ itself, hence

$$\hat{p}(\{011, 101, 110\}) = \hat{p}(011) + \hat{p}(011) + \hat{p}(011) = 3 \cdot (4/27) = 4/9.$$

Read backwards, (3.6) says that the sum of the maximum-likelihood probabilities of all sequences of a given type is the type's maximum-likelihood probability, and hence at most one. This implies that the attenuation is at most the number of types,

$$\hat{A}(I_k^n) = \sum_{\overline{x} \in [k]^n} \hat{p}(\overline{x}) = \sum_{\tau \in \mathcal{T}_k^n} \sum_{\overline{x} \in \tau} \hat{p}(\overline{x}) = \sum_{\tau \in \mathcal{T}_k^n} \hat{p}(\tau) \leq \sum_{\tau \in \mathcal{T}_k^n} 1 = |\mathcal{T}_k^n| = \binom{n+k-1}{k-1}, \quad (3.7)$$

and therefore we obtain,

Lemma 8. For all n and k,

$$\hat{R}(I_k^n) \leq \log \binom{n+k-1}{k-1} \leq (k+n) h\left(\frac{k}{k+n}\right). \qquad \square$$

As for binary alphabets, this bound has a simple information-theoretic interpretation. Every sequence can be specified by first describing its type and then the precise sequence within the type. Since *i.i.d.* distributions assign all sequences of a given type the same probability, using the same number of bits to identify all sequences within the type is optimal, and the redundancy $\hat{R}(I_k^n)$ is upper bounded by the number of bits needed to describe the type.

Surprisingly, in spite of its simplicity, as we will see that in Theorem 12, the upper bound is always tight up to a factor of 8. We now consider two extreme relations between the blocklength and the laphbet size. In one the bound is tight, and in the other it is twice the redundancy and we prove an asymptotically tight upper bound.

When $n = o(k)$, namely the alphabet is much larger than the blocklength, the upper bound simplifies to

$$\hat{R}(I_k^n) \leq \log \binom{n+k-1}{k-1} = \log \binom{n+k-1}{n} \sim n \log \frac{k}{n}. \quad (3.8)$$

This is logical since when $n = o(k)$ most symbols appear once, hence most of the bits are used to describe which n of the k symbols appear. It is easy to do that using $\sim n \log k$ bits. We will see in this range the bound is tight.

When $k = o(n)$, for example, when the alphabet is fixed and the sequence length grows, the upper bound simplifies to

$$\hat{R}(I_k^n) \leq \log \binom{n+k-1}{k-1} \sim (k-1) \log \frac{n}{k}. \tag{3.9}$$

This corresponds to describing the first $k-1$ multiplicities, which can be done using $\lesssim (k-1) \log \frac{n}{k}$ bits. In this range this bound is off by a factor of two. We now refine the bound, and in the next section we will show that it is tight.

The attenuation can be written as

$$\hat{A}(I_k^n) = \sum_{m=1}^{\min\{k,n\}} \binom{k}{m} \sum_{\overline{u} \in \mathcal{U}_m^n} \binom{n}{u_1, \ldots, u_m} \prod_{j=1}^{m} \left(\frac{u_j}{n}\right)^{u_j} \stackrel{\text{def}}{=} \sum_{m=1}^{\min\{k,n\}} T_m^n.$$

Lemma 9. For $k = o(n)$

$$\hat{R}(I_k^n) \lesssim \frac{k-1}{2} \log \frac{n}{k},$$

Proof We first show by induction on m that for all $m \leq n$,

$$T_m^n \leq e^{\frac{m}{12n}} \frac{\sqrt{\pi} \left(\frac{n}{2}\right)^{\frac{m-1}{2}}}{\Gamma\left(\frac{m}{2}\right)}. \tag{3.10}$$

From (2.6),

$$\Gamma\Big(\frac{m+1}{2}\Big) \leq \Gamma\Big(\frac{m}{2} + 1\Big) = \sqrt{\frac{m}{2}}\, \Gamma\Big(\frac{m}{2}\Big),$$

implying that the upper bound (3.10) on T_m increases with m for all $m \leq n$.

For $m = 2$, (3.10) holds since

$$T_2^n \leq \hat{A}(I_2^n) \leq \sqrt{\frac{\pi n}{2}}.$$

For all $m > 2$,

$$
\begin{aligned}
T_m^n &= \sum_{l=1}^{n-1} \binom{n}{l} \left(\frac{l}{n}\right)^l \left(1 - \frac{l}{n}\right)^{n-l} T_{m-1}^{n-l} \\
&\le \sum_{l=1}^{n-1} \binom{n}{l} \left(\frac{l}{n}\right)^l \left(1 - \frac{l}{n}\right)^{n-l} \cdot e^{\frac{m-1}{12n}} \frac{\sqrt{\pi}\left(\frac{n-l}{2}\right)^{\frac{m-2}{2}}}{\Gamma(\frac{m-1}{2})} \\
&\le \frac{e^{\frac{m}{12n}}\,\left(\frac{n}{2}\right)^{\frac{m}{2}-1}}{\sqrt{2}\,\Gamma(\frac{m-1}{2})} \cdot \frac{1}{\sqrt{n}} \sum_{l=1}^{n-1} \left(\frac{l}{n}\right)^{-\frac{1}{2}} \left(1 - \frac{l}{n}\right)^{\frac{m-2}{2}-\frac{1}{2}} \\
&\le \frac{e^{\frac{m}{12n}}\,\left(\frac{n}{2}\right)^{\frac{m-1}{2}}}{\Gamma(\frac{m-1}{2})} \cdot \frac{1}{n} \sum_{l=1}^{n-1} \left(\frac{l}{n}\right)^{-\frac{1}{2}} \left(1 - \frac{l}{n}\right)^{\frac{m-3}{2}} \\
&\le \frac{e^{\frac{m}{12n}}\,\left(\frac{n}{2}\right)^{\frac{m-1}{2}}}{\Gamma(\frac{m-1}{2})} \cdot \int_{x=0}^{1} x^{-\frac{1}{2}} (1-x)^{\frac{m-3}{2}} dx \\
&\le \frac{e^{\frac{m}{12n}}\,\left(\frac{n}{2}\right)^{\frac{m-1}{2}}}{\Gamma(\frac{m-1}{2})} \cdot \frac{\Gamma(\frac{1}{2})\Gamma(\frac{m-1}{2})}{\Gamma(\frac{m}{2})} \\
&= e^{\frac{m}{12n}} \frac{\sqrt{\pi}\left(\frac{n}{2}\right)^{\frac{m-1}{2}}}{\Gamma(\frac{m}{2})}.
\end{aligned}
$$

The lemma follows by applying Stirling's approximation for $\Gamma(m/2)$. □

3.3 Lower bound

Lemma 8 shows that

$$\hat{R}(I_k^n) \le \log \binom{n+k-1}{k-1} \le (k+n) h\left(\frac{k}{k+n}\right).$$

We now show that this bound is essentially tight up to a factor of 8,

$$\hat{R}(I_k^n) \gtrsim \frac{1}{8}(k+n) h\left(\frac{k}{k+n}\right).$$

To prove the lower bound, we group the terms in Shtarkov's Sum by the number of symbols they contain, find the number $\hat{m}$ of symbols that contributes the most to the *upper* bound, and lower bound its contribution to the sum. Rewrite the upper-bound sum as

$$\sum_{\overline{x} \in [k]^n} \hat{p}(\overline{x}) = \sum_{m=1}^{\min\{k,n\}} \sum_{\tau \in \mathcal{T}_{k,m}^n} \sum_{\overline{x} \in \tau} \hat{p}(\overline{x}) = \sum_{m=1}^{\min\{k,n\}} \sum_{\tau \in \mathcal{T}_{k,m}^n} \hat{p}(\tau) \le \sum_{m=1}^{\min\{k,n\}} \sum_{\tau \in \mathcal{T}_{k,m}^n} 1 = \sum_{m=1}^{\min\{k,n\}} |\mathcal{T}_{k,m}^n|.$$

As in Lemma 3.4, the number of types of m-symbol sequences is

$$|\mathcal{T}^n_{k,m}| = \binom{k}{m}\binom{n-1}{m-1},$$

where first binomial coeficient corresponds to the number of ways to select the m symbols appearing in the sequence, and the second, to the number of times each of these symbols appears.

It is easy to see that the largest value of $\mathcal{T}^n_{k,m}$ corresponds to

$$\hat{m} \overset{\text{def}}{=} \arg\max_m \mathcal{T}^n_{k,m} = \left\lceil \frac{nk}{n+k+1} \right\rceil \sim \frac{nk}{n+k}. \tag{3.11}$$

Observe that when the $n = o(k)$, namely the number of elements is much larger than the blocklength, $\hat{m} \sim n$ as almost all elements will be distinct, and that that when the $k = o(n)$, namely the number of elements is much smaller than the blocklength, $\hat{m} \sim k$ as almost all elements will appear.

Again, we begin with the extreme cases.

Theorem 10. For $n = o(k)$,

$$\hat{R}(I^n_k) \sim n \log \frac{k}{n}.$$

Proof The upper bound was derived in (3.8), and we prove the lower bound. As above, when $n = o(k)$, $\hat{m} \sim n$, and the contribution of just the nth term in Shtarkov's sum implies that

$$\hat{R}(I^n_k) \geq \log T_n = \log\left(\binom{k}{n}\frac{n!}{n^n}\right) \geq \log\left(\frac{k-n}{n}\right)^n \gtrsim n \log \frac{k}{n}. \qquad \square$$

Next consider alphabets much smaller than the blocklength. We show that the upper bound in Lemma 9 is tight. It will follow that in this regime, as the blocklength increases, the per-symbol redundancy diminishes to zero.

Theorem 11. For $k = o(n)$,

$$\hat{R}(I^n_k) \sim \frac{k-1}{2} \log \frac{n}{k}.$$

Proof The upper bound was derived in Lemma 9, and we prove the lower bound. As we saw, $\hat{m} \sim k$, and we consider the contribution of types with k symbols alone to Shtarkov's sum. There are

$$\binom{n-1}{k-1}$$

such types and as with the binary case, of them, the type with the lowest $\hat{p}$ is where each of the k symbols appears n/k times. The attenuation is

$$\begin{aligned}
\hat{A}(I_k^n) &\geq T_k \\
&\geq \binom{n-1}{k-1} \cdot \left(\frac{1}{k}\right)^n \cdot \binom{n}{\frac{n}{k}, \frac{n}{k}, \ldots, \frac{n}{k}} \\
&\geq \binom{n-1}{k-1} \cdot \left(\frac{1}{k}\right)^n \cdot \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(e^{\frac{k}{12n}}\sqrt{2\pi\frac{n}{k}}\left(\frac{n}{ke}\right)^{n/k}\right)^k} \\
&> \frac{\sqrt{n}}{e^{\frac{k^2}{12n}}} \binom{n-1}{k-1} \left(\frac{k}{2\pi n}\right)^{\frac{k}{2}}.
\end{aligned}$$

Therefore the redundancy,

$$\begin{aligned}
\hat{R}(I_k^n) &= \log \hat{A}(I_k^n) \\
&\geq (k-1)\log\frac{n-1}{k-1} - \frac{k}{2}\log\left(\frac{n}{k}\right) - \frac{k}{2}\log 2\pi + \frac{1}{2}\log\frac{n}{k} + \frac{1}{2}\log k + \frac{k^2\log e}{12n} \\
&> \frac{k-1}{2}\log\frac{n}{k} - \left(\frac{k}{2}\log 2\pi - \frac{1}{2}\log k\right) + \frac{k^2\log e}{12n} \\
&\geq \frac{k-1}{2}\log\frac{n}{k} + \frac{k^2\log e}{12n},
\end{aligned}$$

and the theorem follows since $\frac{k^2}{n} = o(k)$ if $k = o(n)$. □

For general k, we prove that the simple upper bound in Lemma 8 is always within a factor of essentially 8 from $\hat{R}(I_k^n)$.

Recall that if x is not an integer, if n is an integer $> x$

$$\binom{n}{x} = \frac{n!}{\Gamma(x+1)\Gamma(n-x+1)}.$$

Lemma 12. For all k and n such that $k+n \geq 8$,

$$\hat{R}(I_k^n) \geq \frac{k+n}{8} h\left(\frac{k}{k+n}\right)\left(1 - 8\frac{\log\log(k+n)}{\log(k+n)}\right) \gtrsim \frac{k+n}{8} h\left(\frac{k}{k+n}\right).$$

Proof From (3.11), $\hat{m} = \lceil \frac{nk}{n+k+1} \rceil$. Considering the contribution of T_k alone to Shtarkov's

Sum, and replacing the integer by a fraction we obtain,

$$\begin{aligned}\hat{R}(I_k^n) \geq \log T_{\hat{m}} &\geq \log\left(\frac{1}{kn}\binom{k}{\frac{nk}{n+k}} \cdot \sqrt{n}e^{-\frac{nk}{12(n+k)}} \cdot \binom{n-1}{\frac{nk}{n+k}-1}\left(\frac{k}{2\pi(n+k)}\right)^{\frac{nk}{2(n+k)}}\right)\\ &\geq (k+n)h\left(\frac{k}{k+n}\right) - \frac{kn}{2(k+n)}\log\frac{2\pi e^{\frac{1}{6}}(k+n)}{k} + \log\frac{n+k}{n^2k^2} + \frac{1}{2}\log\frac{1}{2\pi e^{\frac{1}{3}}}\\ &\geq (k+n)\left(h(x) - \frac{x(1-x)}{2}\log\frac{2\pi e^{\frac{1}{3}}}{x}\right) + \log\frac{1}{(k+n)x(1-x)} + \frac{1}{2}\log\frac{1}{2\pi e^{\frac{1}{3}}},\end{aligned}$$

where $x \stackrel{\text{def}}{=} \frac{k}{k+n}$.

Observe that $\frac{1}{k+n} \leq x \leq 1 - \frac{1}{k+n}$, hence

$$(k+n)h(x) \geq \log(k+n). \tag{3.12}$$

We bound the $\log\frac{1}{(k+n)x(1-x)}$ term as a fraction of the leading term,

$$\begin{aligned}\frac{\log\frac{1}{(k+n)x(1-x)}}{(k+n)h(x)} &\geq \frac{\log\frac{4}{(k+n)h(x)}}{(k+n)h(x)}\\ &= \frac{\log 4}{(k+n)h(x)} - \frac{\log\left((k+n)h(x)\right)}{(k+n)h(x)}\\ &\geq \frac{2}{(k+n)h(x)} - \frac{\log\log(k+n)}{\log(k+n)},\end{aligned}$$

where the first inequality follows since

$$h(x) \geq 4x(1-x).$$

To see the last inequality, observe that $\frac{\log y}{y}$ decreases for $y \geq e$, therefore last inequality follows from (3.12) because $\log(k+n) \geq 3$.

The constant term in the lower bound of $\hat{R}(I_k^n)$ is less than 2, therefore

$$\log\frac{1}{(k+n)x(1-x)} + \frac{1}{2}\log\frac{1}{2\pi e^{\frac{1}{3}}} \geq -(k+n)h(x)\cdot\frac{\log\log(k+n)}{\log(k+n)}$$

The lemma follows because for all $0 \leq x \leq 1$, the following inequalities

$$\begin{aligned}h(x) &\geq x(1-x)\log\frac{1}{x},\\ \frac{3}{4}h(x) &> x(1-x)\log 2\pi e^{\frac{1}{3}},\end{aligned}$$

imply that

$$\frac{x(1-x)}{2}\log\frac{2\pi e^{\frac{1}{3}}}{x} < \frac{7}{8}h(x). \qquad \square$$

Combining the lemma with the upper bound in Lemma 8 we obtain the following theorem.

Theorem 13. For all k and n such that $k+n \geq 8$,

$$\frac{1}{8}(k+n)h\left(\frac{k}{k+n}\right) \lesssim \hat{R}(I_k^n) \leq \log \binom{n+k-1}{k-1} \leq (k+n)h\left(\frac{k}{k+n}\right). \qquad \square$$

It follows that when k grows proportionally to n, say $k = \alpha n$,

$$\frac{1}{8}(1+\alpha)nh\left(\frac{\alpha}{1+\alpha}\right) \lesssim \hat{R}(I_k^n) \leq (1+\alpha)nh\left(\frac{\alpha}{1+\alpha}\right),$$

hence the per-symbol redundancy is a constant.

3.4 Remarks

For fixed alphabets as the blocklength grows, there are many results, for example [10, 24, 59, 25, 26, 27, 28, 29, 30, 31].

Acknowledgments

We thank K. Viswanathan, and J. Zhang for helpful discussions and comments. This chapter is adapted from A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004. The author was a primary researcher of the paper.

Chapter 4

Shapes

We study the redundancy incurred with the compression of the *shape* of *i.i.d.* strings, which describes its symbols' relative magnitude. We determine the rate at which per-symbol standard redundancy increases to infinity as the alphabet size increases, showing that unlike the redundancy of the original strings themselves, the shape redundancy, normalized by the number of symbols is bounded, and lies between 0.027 and 1.

4.1 Definitions

Let $\mathcal{A}$ be a possibly infinite, even uncountable alphabet with an order '$<$'. For $\overline{x} = x_1 \ldots x_n \in \mathcal{A}^n$, let

$$\mathcal{X}(\overline{x}) \stackrel{\text{def}}{=} \{x_1, \ldots, x_n\}$$

be the set of symbols appearing in $\overline{x}$. The *rank of* $x \in \mathcal{X}(\overline{x})$ is the number

$$\rho_{\overline{x}}(x) \stackrel{\text{def}}{=} |\{y \in \mathcal{X}(\overline{x}) : y \le x\}|$$

of distinct symbols in $\mathcal{X}(\overline{x})$ not larger than x. The *shape* of $\overline{x}$ is the concatenation

$$\mathsf{S}(\overline{x}) \stackrel{\text{def}}{=} \rho_{\overline{x}}(x_1)\rho_{\overline{x}}(x_2)\ldots\rho_{\overline{x}}(x_n)$$

of all ranks. Consider for example the Roman alphabet with the standard order $a < b < c < \ldots < z$ and the string $\overline{x} =$ "*abracadabra*". Then $\rho_{\overline{x}}(a) = 1$, $\rho_{\overline{x}}(b) = 2$, $\rho_{\overline{x}}(c) = 3$,

$\rho_{\overline{x}}(d) = 4$, and $\rho_{\overline{x}}(r) = 5$, hence

$$\mathsf{S}(abracadabra) = 12513141251.$$

Let

$$\mathsf{S}(\mathcal{A}^n) = \{\mathsf{S}(\overline{x}) : \overline{x} \in \mathcal{A}^n\}$$

denote the set of shapes of all strings in $\mathcal{A}^n$. For example, if $\mathcal{A}$ consists of two elements, then $\mathsf{S}(\mathcal{A}) = \{1\}$, $\mathsf{S}(\mathcal{A}^2) = \{11, 12, 21\}$, $\mathsf{S}(\mathcal{A}^3) = \{111, 112, 221, 121, 212, 211, 122\}$, etc. Let

$$\mathsf{S}^n \overset{\text{def}}{=} \cup_{\mathcal{A}} \mathsf{S}(\mathcal{A}^n)$$

be the set of all length-n *shapes*. For example

$$\begin{aligned} \mathsf{S}^0 &= \{\lambda\} \\ \mathsf{S}^1 &= \{1\} \\ \mathsf{S}^2 &= \{11, 12, 21\} \\ \mathsf{S}^3 &= \{111, 112, 221, 121, 212, 211, 122, \\ &\qquad 123, 321, 231, 132, 312, 213\} \end{aligned}$$

and so on, where λ is the empty string. Finally, let

$$\mathsf{S}^* \overset{\text{def}}{=} \cup_{n=0}^{\infty} \mathsf{S}^n$$

be the set of all shapes.

Observe that a string $x_1 \ldots x_n$ is a shape if and only if it consists of positive integers such that if any number $i \geq 2$ appears, so does $i - 1$, namely,

$$\mathcal{X}(\overline{x}) = \Big[\big|\{x_1, \ldots, x_n\}\big|\Big]$$

where as before

$$[n] = \{1, \ldots, n\}$$

and $[0] \overset{\text{def}}{=} \emptyset$. For example, 1, 11, 21, 212, and 321 are shapes, while 2, 13, and 131 are not.

Every probability distribution p over $\mathcal{A}^*$ induces the distribution p_{S} over S^* where

$$p_{\mathsf{S}}(\overline{\mathsf{s}}) \overset{\text{def}}{=} p(\{\overline{x} \in \mathcal{A}^* : \mathsf{S}(\overline{x}) = \overline{\mathsf{s}}\})$$

is the probability that a string in $\mathcal{A}^*$ generated according to p has shape $\bar{\mathsf{s}}$. When p_{S} is used to evaluate a specific shape probability $p_{\mathsf{S}}(\bar{\mathsf{s}})$, the subscript can be inferred, and hence omitted. For example, let p be the uniform distribution over $\{a, b\}^2$ with the usual ordering $a < b$. Then p induces over S^2 the distribution

$$\begin{aligned} p(11) &= p(\{aa, bb\}) = \frac{1}{2} \\ p(12) &= p(\{ab\}) = \frac{1}{4} \\ p(21) &= p(\{ba\}) = \frac{1}{4}. \end{aligned}$$

For a collection $\mathcal{P}$ of distributions over $\mathcal{A}^*$ or any of its subsets let

$$\mathcal{P}_{\mathsf{S}} \stackrel{\text{def}}{=} \{p_{\mathsf{S}} : p \in \mathcal{P}\}$$

be the collection of all distributions over S^* induced by distributions in $\mathcal{P}$. By (1.1), the *shape redundancy* of $\mathcal{P}$, namely the worst case redundancy of compressing shapes generated by an unknown distribution in $\mathcal{P}_{\mathsf{S}}$, is

$$\hat{R}(\mathcal{P}_{\mathsf{S}}) = \inf_q \sup_{p \in \mathcal{P}_{\mathsf{S}}} \sup_{\mathsf{s} \in \mathsf{S}^*} \log \frac{p(\mathsf{s})}{q(\mathsf{s})}.$$

Note that the shape redundancy of any $\mathcal{P}$ is non-negative.

We will be mainly concerned with $\hat{R}(\mathcal{I}^n_{\mathsf{S}})$, the shape redundancy of the collection $\mathcal{I}^n$ of all *i.i.d.* distributions over length-n strings drawn from any alphabet, finite or infinite. Without loss of generality, the distribution can be assumed to be over the real numbers.

4.2 Combinatorics of shapes

4.2.1 Shapes and ordered set partitions

We now relate shapes to the well-known combinatorial structure of ordered set partitions.

An *ordered partition of a set S* into m parts is an m-tuple of disjoint nonempty subsets of S whose union is S. The *type* of an ordered set partition $R = (R_1, \ldots, R_m)$ is

$$\overline{u}(R) \stackrel{\text{def}}{=} \Big(|R_1|, \ldots, |R_m|\Big),$$

the vector of cardinalities of the sets in R.

We are mainly interested in ordered partitions of $[n] = \{1, \ldots ,n\}$. For example, $(\{2\}, \{1, 4\}, \{3\})$ is an ordered set partition of $[4]$ into 3 parts, and its type is

$$\overline{u}((\{2\}, \{1, 4\}, \{3\})) = (|\{2\}|, |\{1, 4\}|, |\{3\}|) = (1, 2, 1).$$

Let $\mathcal{R}^n$ be the set of all ordered partitions of $[n]$, let $\mathcal{R}^n_m$ be the set of all ordered partitions of $[n]$ into m parts, and let $\mathcal{R}_{\overline{u}}$ be the set of ordered partitions of type $\overline{u}$.

To formalize the connection between shapes and ordered set partitions, define the mapping $\mathfrak{f}_{\mathsf{S}}$ from the set of all shapes to the set of all ordered set partitions by

$$\begin{aligned}\mathfrak{f}_{\mathsf{S}}(\mathsf{s}_1 \ldots \mathsf{s}_n) \overset{\text{def}}{=} \big(&\{i : \mathsf{s}_i = 1\}, \{i : \mathsf{s}_i = 2\}, \ldots , \\ &\{i : \mathsf{s}_i = \max\{\mathsf{s}_1, \ldots ,\mathsf{s}_n\}\}\big).\end{aligned}$$

For example

$$\begin{aligned}\mathfrak{f}_{\mathsf{S}}(12131) &= \big(\{i : \mathsf{s}_i = 1\}, \{i : \mathsf{s}_i = 2\}, \{i : \mathsf{s}_i = 3\}\big) \\ &= \big(\{1, 3, 5\}, \{2\}, \{4\}\big).\end{aligned}$$

The following result follows easily.

Lemma 14. The function $\mathfrak{f}_{\mathsf{S}}$ is a bijection. Furthermore, for all $n \geq 0$

$$\mathfrak{f}_{\mathsf{S}}(\mathsf{S}^n) = \mathcal{R}^n,$$

for all $0 \leq m \leq n$

$$\mathfrak{f}_{\mathsf{S}}(\mathsf{S}^n_m) = \mathcal{R}^n_m,$$

and for all types $\overline{u}$

$$\mathfrak{f}_{\mathsf{S}}(\mathsf{S}_{\overline{u}}) = \mathcal{R}_{\overline{u}}.$$ □

For $0 \leq m \leq n$, let

$$F(n, m) \overset{\text{def}}{=} \big|\mathcal{R}^n_m\big| = \big|\mathsf{S}^n_m\big|$$

and let

$$F(n) \overset{\text{def}}{=} \big|\mathcal{R}^n\big| = \big|\mathsf{S}^n\big| = \sum_{m=0}^{n} F(n, m).$$

For example, $F(0,0) = 1$, $F(1,0) = 0$, $F(1,1) = 1$, $F(2,0) = 0$, $F(2,1) = 1$, and $F(2,2) = 2$, hence, $F(0) = 1$, $F(1) = 1$, and $F(2) = 3$. Similarly, $F(3) = 13$ and $F(4) = 75$. Note that $F(n,0)$ is 1 for $n = 0$ and 0 otherwise. The numbers $F(n,m)$ and $F(n)$ are known as the *ordered Stirling numbers of the second kind*, and the *Fubini numbers*, respectively.

Several interpretations and results are known for both numbers, *e.g.* [61, 62]. For example, one interpretation of $F(n,m)$ is as the number of ways to distribute n distinguishable balls into m distinguishable urns so that no urn is left empty, and a common recursion for $F(n,m)$ for $1 \le m \le n$ is

$$F(n,m) = mF(n-1,m) + mF(n-1,m-1). \tag{4.1}$$

4.2.2 Types of shapes

We classify S^n into sets of equiprobable shapes, and in Section 4.3.2 we use this classification to upper bound shape redundancy using Corollary 3.

Recall that S^n is the collection of length-n shapes. Let

$$\mathsf{S}^n_m \overset{\text{def}}{=} \{\mathsf{s}_1 \ldots \mathsf{s}_n \in \mathsf{S}^n : \left|\{\mathsf{s}_1, \ldots, \mathsf{s}_n\}\right| = m\}$$

be the set of length-n shapes with m symbols. For example,

$$\begin{aligned}\mathsf{S}^3_1 &= \{111\}\\ \mathsf{S}^3_2 &= \{112, 211, 121, 212, 122, 211\}\\ \mathsf{S}^3_3 &= \{123, 231, 312, 321, 132, 213\}\end{aligned}$$

For all $1 \le i \le m \le n$, the *multiplicity* of i in $\overline{\mathsf{s}} \in \mathsf{S}^n_m$ is

$$\mu_i \overset{\text{def}}{=} \mu_i(\overline{\mathsf{s}}) \overset{\text{def}}{=} |\{1 \le j \le n : \mathsf{s}_j = i\}|,$$

the number of times i occurs in $\overline{\mathsf{s}}$. The *type* of $\overline{\mathsf{s}} \in \mathsf{S}^n_m$ is

$$\overline{u}(\overline{\mathsf{s}}) \overset{\text{def}}{=} (\mu_1, \ldots, \mu_m),$$

the m-tuple of multiplicities of the symbols in $\overline{\mathsf{s}}$. For example, the multiplicities of the symbols in 12131 are $\mu_1 = 3$, $\mu_2 = 1$, and $\mu_3 = 1$, hence $\overline{u}(12131) = (3,1,1)$.

Clearly, the type of any length-n shape corresponds to an ordered partition of n and vice versa, where the number of partition parts equals the number of shape symbols. Hence, for all $m \le n$

$$\overline{u}(\mathsf{S}^n_m) = \mathcal{U}^n_k$$

and

$$\overline{u}(\mathsf{S}^n) = \mathcal{U}^n.$$

Equations (3.1) and (3.2) therefore imply that the number of types of length-n shapes with m symbols is $\binom{n-1}{m-1}$ and that the number of types of length-n shapes is 2^{n-1}.

For $\overline{u} = (u_1, \ldots, u_m) \in \mathcal{U}^n_k$, let

$$\mathsf{S}_{\overline{u}} \overset{\text{def}}{=} \{\overline{\mathsf{s}} \in \mathsf{S}^* : \overline{u}(\overline{\mathsf{s}}) = \overline{u}\}$$

be the collection of shapes of type $\overline{u}$. A standard counting argument shows that

$$\left|\mathsf{S}_{\overline{u}}\right| = \binom{n}{u_1, \ldots, u_m} = \frac{n!}{u_1! \cdots u_m!}. \tag{4.2}$$

4.3 Redundancy of shapes

In the last chapter we saw that the standard per-symbol redundancy of *i.i.d.* distributions increases to infinity as the alphabet size grows. We now show that the per-symbol shape redundancy of *i.i.d.* distributions is bounded. Specifically, for all $n \in \mathbb{Z}^+$

$$0.027 \le \frac{1}{4} \log\left(\frac{4}{\pi e^{1/6}}\right) \le \hat{R}(\mathcal{I}^n_{\mathsf{S}})/n \le 1.$$

We also consider the shape redundancy of short strings, discuss, though do not analyze, sequential shape compression, and related these results to oredered set partitions.

4.3.1 The redundancy of shapes of length 1 and 2

There is only one distribution on $\mathsf{S}^1 = \{1\}$, hence

$$\hat{R}(\mathcal{I}^1_{\mathsf{S}}) = 0.$$

To determine $\hat{R}(\mathcal{I}_{\mathsf{S}}^2)$, the shape redundancy of the collection $\mathcal{I}_{\mathsf{S}}^2$ of *i.i.d.*-induced distributions over $\mathsf{S}^2 = \{11, 12, 21\}$, consider Shtarkov's sum (2.1)

$$\begin{aligned}\hat{R}(\mathcal{I}_{\mathsf{S}}^2) &= \log\left(\sum_{\bar{\mathsf{s}}\in\mathsf{S}^2} \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(\bar{\mathsf{s}})\right)\\ &= \log\left(\sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(11) + \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(12) + \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(21)\right).\end{aligned}$$

We show that

$$\sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(11) = 1$$

and

$$\sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(12) = \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(21) = \frac{1}{2}$$

hence

$$R(\mathcal{I}_{\mathsf{S}}^2) = \log\left(1 + \frac{1}{2} + \frac{1}{2}\right) = 1.$$

Since any constant *i.i.d.* distribution induces

$$p(11) = 1,$$

the maximum probability of the shape 11 is 1. To show that the maximum shape probability of the shapes 12 and 21 is 1/2, note that any *i.i.d.* distribution $p \in \mathcal{I}^2$ induces the same probability on 12 and 21, namely

$$p(12) = p(21)$$

hence

$$\sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(12) = \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(21) \le \frac{1}{2}.$$

On the other hand, a continuous distribution over an interval, *e.g.* the uniform distribution over $[0, 1]$, induces

$$p(12) = p(21) = \frac{1}{2}$$

hence

$$\sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(12) = \sup_{p\in\mathcal{I}_{\mathsf{S}}^2} p(21) = \frac{1}{2}.$$

4.3.2 Upper bound

In the previous section, we determined $\hat{R}(\mathcal{I}^1_{\mathsf{S}})$ and $\hat{R}(\mathcal{I}^2_{\mathsf{S}})$ by evaluating the maximum shape probabilities exactly. For large n, exact calculation of these probabilities seems difficult. Instead, we establish a relation between $\hat{R}(\mathcal{I}^n_{\mathsf{S}})$ and $|\mathcal{U}^n|$, and use it to prove that the per-symbol shape redundancy $\hat{R}(\mathcal{I}^n_{\mathsf{S}})/n$ is at most one.

Theorem 15. For all n

$$\hat{R}(\mathcal{I}^n_{\mathsf{S}}) \le n-1.$$

Proof Let $\overline{u} \in \mathcal{U}^n$ be a type. Every *i.i.d.* distribution induces the same probability on all shapes in $\mathsf{S}_{\overline{u}}$. Hence the probability assigned to every shape in $\mathsf{S}_{\overline{u}}$ is at most $1/|\mathsf{S}_{\overline{u}}|$. Let $\overline{p}_{\overline{u}}$ be the uniform distribution assigning probability $1/|\mathsf{S}_{\overline{u}}|$, to every shape in $\mathsf{S}_{\overline{u}}$, and let

$$\hat{\mathcal{I}^n_{\mathsf{S}}} = \{\overline{p}_{\overline{u}} : \overline{u} \in \mathcal{U}^n\}$$

be the collection of all such uniform distributions. $\hat{\mathcal{I}^n_{\mathsf{S}}}$ clearly dominates $\mathcal{I}^n_{\mathsf{S}}$ and the theorem follows from Corollary 3 and (3.2). □

Many applications give rise to sequences containing relatively few symbols. In the remainder of this section, we refine Theorem 15 and derive compression algorithms whose redundancy for these sequences is low.

Let $\mathcal{P}^n_{\mathsf{S}}$ be a collection of distributions over length-n shapes, and let q be a distribution over length-n shapes, not necessarily in $\mathcal{P}^n_{\mathsf{S}}$. The *redundancy* of q, *i.e.,* of the compression algorithm associated with it, for the collection S^n_m of length-n shapes with m symbols is

$$\hat{R}_m(\mathcal{P}^n_{\mathsf{S}}, q) \overset{\text{def}}{=} \sup_{p\in\mathcal{P}^n_{\mathsf{S}}} \max_{\overline{\mathsf{s}}\in\mathsf{S}^n_m} \log \frac{p(\overline{\mathsf{s}})}{q(\overline{\mathsf{s}})},$$

the highest excess number of bits used to encode any $m-$symbol shape.

For $m = 1, \ldots, n$, Theorem 15 can be used to construct a distribution q_m over length-n shapes, whose redundancy for $m-$symbol shapes is at most $\log \mathcal{U}^n_k$. Weighting each of q_m by $\frac{1}{n}$, we obtain the following.

Theorem 16. There is a distribution q over length-n shapes such that for all $m \le n$

$$\hat{R}_m(\mathcal{I}^n_{\mathsf{S}}, q) \le \log \binom{n-1}{m-1} + \log n. \qquad \square$$

4.3.3 Lower bound

We show that the per-symbol redundancy of shapes of *i.i.d.* distributions is at least 0.027.

Theorem 17. For all $n > 1$

$$\hat{R}(\mathcal{I}_{\mathsf{S}}^n) \geq \frac{n}{4} \cdot \log \frac{4}{\pi e^{\frac{1}{6}}} > 0.027n.$$

Proof We lower bound the maximum shape probabilities, and incorporate this bound into Shtarkov's sum.

For all $\bar{\mathsf{s}} \in \mathsf{S}^*$ let

$$\mathsf{S}_p^{-1}(\bar{\mathsf{s}}) = \{\overline{x} \in \mathcal{A}^* : \mathsf{S}(\overline{x}) = \bar{\mathsf{s}} \text{ and } p(\overline{x}) > 0\}$$

be the *support of a shape* $\bar{\mathsf{s}}$ with respect to the distribution p. For every $\bar{\mathsf{s}} \in \mathsf{S}^n$

$$\sup_{p \in \mathcal{I}_{\mathsf{S}}^n} p(\bar{\mathsf{s}}) = \sup_{p \in \mathcal{I}^n} p\big(\mathsf{S}_p^{-1}(\bar{\mathsf{s}})\big) \geq \sup_{p \in \mathcal{I}^n} \max_{\overline{x} \in \mathsf{S}_p^{-1}(\bar{\mathsf{s}})} p(\overline{x}).$$

For all $\bar{\mathsf{s}}$ with type $\overline{u} = (u_1, \ldots, u_m)$, standard maximum likelihood arguments imply that

$$\max_{p \in \mathcal{I}^n} \max_{\overline{x} \in \mathsf{S}_p^{-1}(\bar{\mathsf{s}})} p(\overline{x}) = \prod_{j=1}^{m} \left(\frac{u_j}{n}\right)^{u_j}$$

hence

$$\sup_{p \in \mathcal{I}_{\mathsf{S}}^n} p(\bar{\mathsf{s}}) \geq \prod_{j=1}^{m} \left(\frac{u_j}{n}\right)^{u_j}.$$

Incorporating this lower bound into Shtarkov's sum (2.1), we obtain

$$\begin{aligned}
\hat{R}(\mathcal{I}^n_{\mathsf{S}}) &= \log\left(\sum_{m=1}^{n} \sum_{\overline{u}\in\mathcal{U}^n_k} \sum_{\overline{\mathsf{s}}\in\mathsf{S}_{\overline{u}}} \sup_{p\in\mathcal{I}^n_{\mathsf{S}}} p(\overline{\mathsf{s}})\right)\\
&\ge \log\left(\sum_{m=1}^{n} \sum_{\overline{u}\in\mathcal{U}^n_k} \sum_{\overline{\mathsf{s}}\in\mathsf{S}_{\overline{u}}} \prod_{j=1}^{m} \left(\frac{u_j}{n}\right)^{u_j}\right)\\
&\overset{(a)}{=} \log\left(\sum_{m=1}^{n} \sum_{\overline{u}\in\mathcal{U}^n_k} \binom{n}{u_1,\ldots,u_m} \prod_{j=1}^{m} \left(\frac{u_j}{n}\right)^{u_j}\right)\\
&\overset{(b)}{\ge} \log\left(\sum_{m=1}^{n} \frac{1}{e^{\frac{m}{12}}} \sum_{\overline{u}\in\mathcal{U}^n_k} \sqrt{\frac{2\pi n}{\left(\frac{n}{m}\right)^m}} \left(\frac{1}{\sqrt{2\pi}}\right)^m\right)\\
&\overset{(c)}{\ge} \log\left(\sum_{m=1}^{n} \frac{1}{e^{\frac{m}{12}}} \binom{n-1}{m-1} \frac{\sqrt{2\pi n}}{\left(\frac{n}{m}\right)^{\frac{m}{2}}} \left(\frac{1}{\sqrt{2\pi}}\right)^m\right)\\
&\overset{(d)}{\ge} \log\left(\frac{1}{e^{\frac{n}{24}}} \binom{n-1}{\frac{n}{2}-1} \frac{\sqrt{2\pi n}}{2^{n/4}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n/2}\right)\\
&\ge \frac{n}{4}\log\left(\frac{4}{\pi}\right) - \frac{n}{24}\log e
\end{aligned}$$

where (a) follows from (4.2), (b) from (2.5), (c) from (3.1), and (d) by considering only the $m = \frac{n}{2}$ term. □

4.4 Sequential description

So far, we described the shape of a whole block of symbols. However in many applications, the symbols must be encoded as they arrive. Representations of shapes for such applications must be *sequential*, namely the shape of a string's prefix must be the prefix of the string's shape. In this brief section we describe a sequential representation of shapes.

Recall that

$$\mathcal{A}(x_1^n) \overset{\text{def}}{=} \{x_1,\ldots,x_n\}$$

is the set of symbols appearing in $x_1^n = x_1 \ldots x_n \in \mathcal{A}^n$, and if the alphabet $\mathcal{A}$ is ordered, the rank of x with respect to $\mathcal{A}(x_1^n)$ is

$$\rho_{x_1^n}(x) = |\{y \in \mathcal{A}(x_1^n) : y \le x\}|,$$

the number of distinct symbols in $\mathcal{A}(x_1^n)$ not larger than x.

Let the *sequential rank* of x_i be

$$\mathfrak{r}_{x_1^{i-1}}(x_i) \stackrel{\text{def}}{=} \begin{cases} 2\rho_{x_1^{i-1}}(x_i) & x_i \in \mathcal{A}(x_1^{i-1}), \\ 2\rho_{x_1^{i-1}}(x_i) + 1 & x_i \notin \mathcal{A}(x_1^{i-1}). \end{cases}$$

For example, in the string *abracadabra*, the sequential rank of the first a is $2 \cdot 0 + 1 = 1$, the sequential rank of all the other $a's$ is $2 \cdot 1 = 2$, the sequential rank of the first b is $2 \cdot 1 + 1 = 3$, while that of the second is $2 \cdot 2 = 4$ the sequential rank of the first r is $2 \cdot 2 + 1 = 5$, while that of the second is $2 \cdot 5 = 10$, and so on.

The *sequential shape* of x_1^n is

$$\mathfrak{r}_\lambda(x_1)\, \mathfrak{r}_{x_1^1}(x_2) \ldots \mathfrak{r}_{x_1^{n-1}}(x_n),$$

the concatenation of the sequential ranks. For example, the sequential shape of "abracadabra" is 135252724(10)2. It can be easily verified that the sequential shape of every prefix of a string is the corresponding prefix of its sequential shape.

Acknowledgments

We thank N. Jevtić, N. Kashyap, G. Shamir, W. Szpankowski, K. Viswanathan, and J. Zhang for helpful discussions and comments. This chapter is adapted from A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004. The author was a primary researcher of the paper.

Chapter 5

Patterns

We saw that strings can be described by separately conveying its symbols, and its *pattern*—the order in which the symbols appear. Concentrating on the latter, we show that the patterns of iid strings over all, including infinite and even unknown, alphabets, can be compressed with diminishing redundancy, both in block, and sequentially.

To establish these results, we show that the number of patterns is the Bell number, that the number of patterns with a given number of symbols is the Stirling number of the second kind, and that the redundancy of patterns can be bounded using a celebrated result of Hardy and Ramanujan on the number of integer partitions.

5.1 Results

In Section 5.2 we formally define patterns and the redundancy of compressing them. In Section 5.3 we derive some useful properties of patterns, including a correspondence between patterns and set partitions. We use this analogy to show that the number of patterns is the Bell number and that the number of patterns with a given number of symbols is the Stirling number of the second kind.

We are primarily interested in universal codes for the class $\mathcal{I}^n$ of all *i.i.d.* distributions, over all possible alphabets, even continuous. As mentioned earlier, for standard compression the per-symbol redundancy increases to infinity as the alphabet size grows. Yet in Section 5.4, we show that $\hat{R}(\mathcal{I}^n_\Psi)$, the block redundancy of compressing patterns of *i.i.d.* distributions over potentially infinite alphabets is bounded by

$$\left(\frac{3}{2}\log e\right)n^{\frac{1}{3}}(1+o(1)) \le \hat{R}(\mathcal{I}_\Psi^n) \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}.$$

Therefore the per-symbol redundancy of coding patterns diminishes to zero as the blocklength increases, irrespective of the alphabet size. The proofs use an analogy between patterns and set partitions which allows us to incorporate a celebrated result of Hardy and Ramanujan on the number of partitions of an integer.

In Section 5.5, we consider sequential pattern encoders. We first construct an encoder with redundancy of at most

$$\frac{2}{\sqrt{2}-1}\left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n} = \frac{4\pi\log e}{(2-\sqrt{2})\sqrt{3}}\sqrt{n}.$$

However, this encoder has high computational complexity. We therefore describe a linear-complexity encoder, which also has diminishing, albeit slightly higher, redundancy of at most

$$\mathcal{O}(n^{2/3}),$$

where the implied constant is less than 10. For both block and sequential compression, the redundancy grows sublinearly with the blocklength, hence the per symbol redundancy $\hat{R}(\mathcal{I}_\Psi^n)/n$ diminishes to zero.

5.2 Definitions

We formally describe patterns and their redundancy.

Let $\mathcal{A}$ be any alphabet. For $\overline{x} = x_1^n = x_1, \ldots, x_n \in \mathcal{A}^n$,

$$\mathcal{X}(\overline{x}) \stackrel{\text{def}}{=} \{x_1, \ldots, x_n\}$$

denotes the set of symbols appearing in $\overline{x}$. The *index of* $x \in \mathcal{X}(\overline{x})$ is

$$\imath_{\overline{x}}(x) \stackrel{\text{def}}{=} \min\{|\mathcal{X}(x_1^i)| : 1 \le i \le n \text{ and } x_i = x\},$$

one more than the number of distinct symbols preceding x's first appearance in $\overline{x}$. The *pattern of* $\overline{x}$ is the concatenation

$$\Psi(\overline{x}) \stackrel{\text{def}}{=} \imath_{\overline{x}}(x_1)\imath_{\overline{x}}(x_2)\ldots\imath_{\overline{x}}(x_n),$$

of all indices. For example, if $\overline{x}$ = "*abracadabra*", $\imath_{\overline{x}}(a) = 1$, $\imath_{\overline{x}}(b) = 2$, $\imath_{\overline{x}}(r) = 3$, $\imath_{\overline{x}}(c) = 4$, and $\imath_{\overline{x}}(d) = 5$, hence

$$\Psi(abracadabra) = 12314151231.$$

Let

$$\Psi(\mathcal{A}^n) = \{\Psi(\overline{x}) : \overline{x} \in \mathcal{A}^n\}$$

denote the set of patterns of all strings in $\mathcal{A}^n$. For example, if $\mathcal{A}$ contains two elements, then $\Psi(\mathcal{A}) = \{1\}$, $\Psi(\mathcal{A}^2) = \{11, 12\}$, $\Psi(\mathcal{A}^3) = \{111, 112, 121, 122\}$, etc. Let

$$\Psi^n = \cup_{\mathcal{A}} \Psi(\mathcal{A}^n)$$

denote the set of all length-n patterns, and let

$$\Psi^* = \cup_{n=0}^{\infty} \Psi^n$$

be the set of all patterns. For example,

$$\begin{aligned}
\Psi^0 &= \{\lambda\},\\
\Psi^1 &= \{1\},\\
\Psi^2 &= \{11, 12\},\\
\Psi^3 &= \{111, 112, 121, 122, 123\},\\
\Psi^* &= \{\lambda, 1, 11, 12, 111, 112, \ldots\},
\end{aligned}$$

where λ is the empty string. Figure 5.2 depicts a tree representation of all patterns of length at most 4.

It is easy to see that a string $\overline{\psi}$ is a pattern iff it consists of positive integers such that no integer $i > 1$ appears before the first occurrence of $i - 1$. For example, 1, 12, and 1213 are patterns, while 2, 21, and 131 are not.

Every probability distribution p over $\mathcal{A}^*$ induces a distribution p_Ψ over patterns on Ψ^*, where

$$p_\Psi(\overline{\psi}) \stackrel{\text{def}}{=} p(\{\overline{x} \in \mathcal{A}^* : \Psi(\overline{x}) = \overline{\psi}\}),$$

is the probability that a string generated according to p has pattern $\overline{\psi}$. When pattern probabilities $p_\Psi(\overline{\psi})$ are evaluated, the subscript Ψ can be inferred, and is hence omitted.

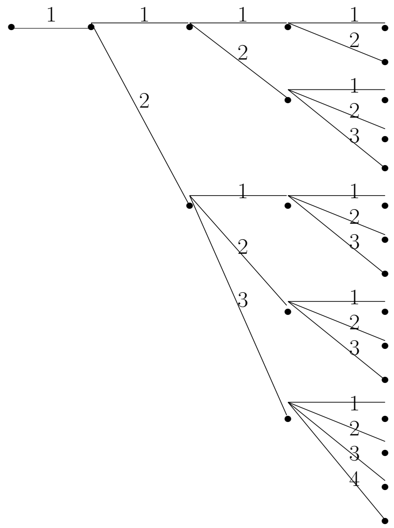

Figure 5.1: A tree representation of patterns of length ≤ 4.

For example, let p be a uniform distribution over $\{a, b\}^2$. Then p induces on Ψ^2 the distribution

$$p(11) = p(\{aa, bb\}) = \frac{1}{2},$$
$$p(12) = p(\{ab, ba\}) = \frac{1}{2}.$$

For a collection $\mathcal{P}$ of distributions over $\mathcal{A}^*$ let

$$\mathcal{P}_\Psi \stackrel{\text{def}}{=} \{p_\Psi : p \in \mathcal{P}\}$$

denote the collection of distributions over Ψ^* induced by probability distributions in $\mathcal{P}$. From the derivations leading to Equation (1.1), the worst case *pattern redundancy* of $\mathcal{P}$, *i.e.,* the worst case redundancy of patterns generated according to an unknown distribution in $\mathcal{P}_\Psi$ is

$$\hat{R}(\mathcal{P}_\Psi) = \min_q \max_{p \in \mathcal{P}_\Psi} \max_{\overline{\psi} \in \Psi^*} \log \frac{p(\overline{\psi})}{q(\overline{\psi})}, \tag{5.1}$$

where q is any distribution over Ψ^*. In particular, for all $\mathcal{P}$,

$$\hat{R}(\mathcal{P}_\Psi) \geq 0.$$

As mentioned earlier, we are mostly interested in $\hat{R}(\mathcal{I}^n_\Psi)$, the pattern redundancy of $\mathcal{I}^n$, the collection of arbitrary *i.i.d.* distributions over length-n strings. We show that the per-symbol redundancy $\hat{R}(\mathcal{I}^n_\Psi)/n$ diminishes to zero, and that diminishing per-symbol redundancy can be achieved both by block and sequential coding with a constant number of operations per symbol.

5.3 Combinatorics of patterns

5.3.1 Set partitions and patterns

A *partition* of a set S is a collection of disjoint nonempty subsets of S whose union is S. For $n \ge 0$, let $[n] \overset{\text{def}}{=} \{1,\ldots,n\}$ with $[0] \overset{\text{def}}{=} \emptyset$. Let $\mathcal{S}^n$ be the set of all partitions of $[n]$, and let

$$\mathcal{S}^* = \cup_{n=0}^{\infty} \mathcal{S}^n$$

be the collection of all partitions of $[n]$ for all $n \in \mathbb{Z}^+$. For example,

$$\begin{aligned}
\mathcal{S}^0 &= \{\emptyset\}\\
\mathcal{S}^1 &= \Big\{\{\{1\}\}\Big\}\\
\mathcal{S}^2 &= \Big\{\{\{1,2\}\},\{\{1\},\{2\}\}\Big\}.
\end{aligned}$$

For $0 \le m \le n$, let $B(n,m)$ be the number of partitions of $[n]$ into m sets, and let

$$B(n) = \sum_{k=0}^{n} B(n,m)$$

be the number of partitions of $[n]$. For example, $B(0,0) = 1$, $B(1,0) = 0$, $B(1,1) = 1$, $B(2,0) = 0$, $B(2,1) = 1$, $B(2,2) = 1$, and so on, hence,

$$B(0) = 1,\ B(1) = 1,\ B(2) = 2,\ B(3) = 5,\ B(4) = 15,\ldots$$

Note that $B(n,0)$ is 1 for $n = 0$ and 0 otherwise.

The numbers $B(n,m)$, are called *Stirling numbers of the second kind* while the numbers $B(n)$, are called *Bell numbers*. Many results are known for both [61]. In particular, it is easy to see that for all $n > 0$, Bell numbers satisfy the recursion

$$B(n+1) = \sum_{i=0}^{n} \binom{n}{i} \cdot B(i).$$

Set partitions are equivalent to patterns. To see that, let the mapping $\mathfrak{f}_\Psi : \Psi^* \to \mathcal{S}^*$ assign to $\overline{\psi} \in \Psi^*$ the set partition

$$\mathfrak{f}_\Psi(\overline{\psi}) \overset{\text{def}}{=} \left\{ \{i : \psi_i = j\} : 1 \le j \le \max_{1 \le i \le |\overline{\psi}|} \psi_i \right\},$$

where $|\overline{\psi}|$ denotes the length of $\overline{\psi}$. For example, for the pattern $\overline{\psi} = \psi_1 \ldots \psi_5 = 12131$,

$$\begin{aligned} \mathfrak{f}_\Psi(\overline{\psi}) &= \big\{\{i : \psi_i = 1\}, \{i : \psi_i = 2\}, \{i : \psi_i = 3\}\big\} \\ &= \big\{\{1,3,5\}, \{2\}, \{4\}\big\}. \end{aligned}$$

The following follows easily.

Lemma 18. The function $\mathfrak{f}_\Psi : \Psi^* \to \mathcal{S}^*$ is a bijection. Furthermore, for every n,

$$\mathfrak{f}_\Psi(\Psi^n) = \mathcal{S}^n. \qquad \square$$

5.3.2 Profiles

We classify patterns and set partitions by their *profile*, which will be useful in evaluating the redundancy of *i.i.d.*-induced distributions.

The *multiplicity* of $\psi \in \mathbb{Z}^+$ in $\overline{\psi}$ is

$$\mu_\psi \overset{\text{def}}{=} \mu_\psi(\overline{\psi}) \overset{\text{def}}{=} |\{1 \le i \le |\overline{\psi}| : \psi_i = \psi\}|,$$

the number of times ψ appears in $\overline{\psi}$. The *prevalence* of a multiplicity $\mu \in \mathbb{N}$ in $\overline{\psi}$ is

$$\varphi_\mu \overset{\text{def}}{=} \varphi_\mu(\overline{\psi}) \overset{\text{def}}{=} |\{\psi : \mu_\psi = \mu\}|,$$

the number of symbols appearing μ times in $\overline{\psi}$. The *profile* of $\overline{\psi}$ is

$$\overline{\varphi} \overset{\text{def}}{=} \varphi(\overline{\psi}) \overset{\text{def}}{=} \left(\varphi_{|\overline{\psi}|}, \ldots, \varphi_1\right)$$

the vector of prevalences of μ in $\overline{\psi}$ for $1 \le \mu \le |\overline{\psi}|$.

Similarly, the *profile* of S is the vector

$$\overline{\varphi}(S) = (\varphi_{|\overline{\psi}|}, \ldots, \varphi_1),$$

where

$$\varphi_\mu = \big|\{s \in S : |s| = \mu\}\big|$$

is the number of sets of cardinality μ in S. The following is easily observed.

Lemma 19. For all $\overline{\psi} \in \Psi^*$,

$$\varphi(\overline{\psi}) = \varphi\big(\mathfrak{f}_\Psi(\overline{\psi})\big). \qquad \square$$

For example, the pattern $\psi = 12131$ has multiplicities $\mu_1 = 3$, $\mu_2 = \mu_3 = 1$, and $\mu_\psi = 0$ for all other $\psi \in \mathbb{Z}^+$. Hence its prevalences are $\varphi_1 = 2$, $\varphi_2 = 0$, $\varphi_3 = 1$, $\varphi_4 = \varphi_5 = 0$, and its profile is $\varphi(\psi) = (0, 0, 1, 0, 2)$. On the other hand, we see that $\mathfrak{f}_\Psi(\overline{\psi}) = \big\{\{1, 3, 5\}, \{2\}, \{4\}\big\}$ with its profile $\varphi(S) = (0, 0, 1, 0, 2)$.

Let

$$\Phi^n = \{\overline{\varphi} : \exists\, \overline{\psi} \in \Psi^n : \varphi(\overline{\psi}) = \overline{\varphi}\}$$

be the set of profiles of all length-n patterns, and let

$$\Phi^* = \cup_{n=0}^{\infty} \Phi^n$$

be the set of profiles of all patterns. Clearly, Φ^n and Φ^* are also the set of profiles of all set partitions in $\mathcal{S}^n$ and $\mathcal{S}^*$ respectively, and for all $\overline{\varphi} \in \Phi^n$

$$\sum_{\mu=1}^{n} \mu \varphi_\mu = n.$$

For $\overline{\varphi} \in \Phi^*$, let

$$\Psi_{\overline{\varphi}} \overset{\text{def}}{=} \{\overline{\psi} \in \Psi^* : \varphi(\overline{\psi}) = \overline{\varphi}\}$$

be the collection of patterns of profile $\overline{\varphi}$, and equivalently let

$$\mathcal{S}_{\overline{\varphi}} \overset{\text{def}}{=} \{S \in \mathcal{S}^* : \varphi(S) = \overline{\varphi}\}$$

denote the collection of partitions whose profile is $\overline{\varphi}$. It follows that for all $\overline{\varphi} \in \Phi^*$,

$$\mathfrak{f}_\Psi(\Psi_{\overline{\varphi}}) = \mathcal{S}_{\overline{\varphi}}.$$

5.3.3 Useful results

In this Section, we evaluate the size of $\Psi_{\overline{\varphi}}$ and recall Shtarkov's result for computing the worst case redundancy.

Number of patterns of a given profile

Let

$$N(\overline{\varphi}) \overset{\text{def}}{=} \left|\Psi_{\overline{\varphi}}\right| = \left|\mathcal{S}_{\overline{\varphi}}\right|$$

be the number of patterns of profile $\overline{\varphi}$. It follows that

Lemma 20. For all $n \geq 0$ and $\overline{\varphi} = (\varphi_1, \ldots, \varphi_n) \in \Phi^n$,

$$N(\overline{\varphi}) = \frac{n!}{\prod_{\mu=0}^{n} (\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}.$$

Proof There is only one pattern of length 0, the empty string λ, hence the lemma holds.

There are many ways to derive this result. To see one, let $S \in \mathcal{S}_{\overline{\varphi}}$ be a profile-$\overline{\varphi}$ partition of $[n] = \{1, \ldots, n\}$. For $i = 1, \ldots, n$, let S_μ be the collection of elements in sets of size μ. Clearly,

$$|S_\mu| = \mu\varphi_\mu,$$

hence $[n]$ can be decomposed into the sets $S_1, \ldots, S_n$ in

$$\binom{n}{1\varphi_1, 2\varphi_2, \ldots, n\varphi_n} = \frac{n!}{\prod_{\mu=1}^{n} (\mu\varphi_\mu)!}$$

ways. Each set S_μ can be further decomposed into φ_μ interchangeable sets of size μ in

$$\binom{\mu\varphi_\mu}{\underbrace{\mu, \ldots, \mu}_{\varphi_\mu}} \frac{1}{\varphi_\mu!} = \frac{(\mu\varphi_\mu)!}{(\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}$$

ways. These two decompositions uniquely define the partition, hence the number of profile-$\overline{\varphi}$ partitions of $[n]$ is

$$\frac{n!}{\prod_{\mu=1}^{n} (\mu\varphi_\mu)!} \cdot \prod_{\mu=1}^{n} \frac{(\mu\varphi_\mu)!}{\mu!^{\varphi_\mu} \cdot \varphi_\mu!} = \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \cdot \varphi_\mu!}. \qquad \square$$

5.4 Redundancy of patterns

We show that for all $n \in \mathbb{Z}^+$,

$$\left(\frac{3}{2}\log e\right) \cdot n^{\frac{1}{3}}(1 + o(1)) \leq \hat{R}(\mathcal{I}_\Psi^n) \leq \left(\pi\sqrt{\frac{3}{2}\log e}\right)\sqrt{n},$$

namely, the redundancy of patterns of *i.i.d.* distributions is sublinear in the blocklength, implying that the per-symbol redundancy diminishes to zero. To obtain these bounds we rewrite Shtarkov's sum (2.1) as

$$\hat{R}(\mathcal{I}_\Psi^n) = \log\left(\sum_{\overline{\varphi}\in\Phi^n}\sum_{\overline{\psi}\in\Psi_{\overline{\varphi}}}\max_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi})\right). \tag{5.2}$$

In the next subsection we use this sum to compute $\hat{R}(\mathcal{I}_\Psi^1)$ and $\hat{R}(\mathcal{I}_\Psi^2)$. However, for larger n, exact calculation of the maximum-likelihood probabilities of patterns, namely $\max_{p\in\mathcal{I}_\Psi^n} p(\overline{\psi})$, seems difficult [5]. Hence we use the approach in Chapter 2 and saddle point analysis techniques in Chapter 6 for upper and lower bounds on the redundancy.

5.4.1 The redundancy of patterns of length 1 and 2

We determine the redundancies $\hat{R}(\mathcal{I}_\Psi^1)$ and $\hat{R}(\mathcal{I}_\Psi^2)$ of *i.i.d.*-induced distributions over patterns of length 1 and 2 respectively.

There is only one distribution on $\Psi^1 = \{1\}$, hence

$$\hat{R}(\mathcal{I}_\Psi^1) = 0.$$

For length 2, consider the collection of distributions over

$$\Psi^2 = \{11, 12\}$$

induced by the set $\mathcal{I}^2$ of *i.i.d.* distributions over strings of length 2. By Shtarkov's sum,

$$\begin{aligned}\hat{R}(\mathcal{I}_\Psi^2) &= \log\left(\sum_{\overline{\psi}\in\Psi^2}\max_{p\in\mathcal{I}_\Psi^2} p(\overline{\psi})\right)\\ &= \log\left(\max_{p\in\mathcal{I}_\Psi^2} p(11) + \max_{p\in\mathcal{I}_\Psi^2} p(12)\right).\end{aligned} \tag{5.3}$$

Since any constant *i.i.d.* distribution assigns $p(11) = 1$, the maximum-likelihood probability of 11 is 1, hence

$$\max_{p\in\mathcal{I}_\Psi^2} p(11) = 1.$$

Similarly, any continuous distribution over $[0,1]^2$ assigns $p(12) = 1$, hence

$$\max_{p\in\mathcal{I}_\Psi^2} p(12) = 1.$$

Incorporating into Equation (5.3), we obtain

$$R(\mathcal{I}_\Psi^2) = \log(1+1) = 1.$$

Unfortunately, calculation of maximum-likelihood probabilities for longer patterns, seems difficult. Therefore, instead of evaluating the sum in (5.2) exactly, we bound the maximum-likelihood probabilities of $\overline{\psi} \in \Psi^n$ to obtain bounds on $\hat{R}(\mathcal{I}_\Psi^n)$.

5.4.2 Upper bound

We show that $\hat{R}(\mathcal{I}_\Psi^n)$ is at most the logarithm of the number of profiles. Similar to the correspondence between patterns and set partitions, we obtain a correspondence between profiles of patterns and unordered partitions of positive integers, and use Hardy and Ramanujan's results on the number of unordered partitions of positive integers to bound the number of profiles, and hence the redundancy.

Lemma 21. For all n,

$$\hat{R}(\mathcal{I}_\Psi^n) \le \log|\Phi^n|.$$

Proof Every induced *i.i.d.* distribution assigns the same probability to all patterns of the same profile. Hence the probability assigned to every pattern of a given profile is at most the inverse of the number of patterns of that profile. Let $\hat{\mathcal{I}_\Psi^n}$ consist of $|\Phi^n|$ distributions, one for each profile. The distribution associated with any profile assigns to each pattern of that profile a probability equal to the inverse of the number of patterns of the profile. $\hat{\mathcal{I}_\Psi^n}$ clearly dominates $\mathcal{I}_\Psi^n$ and the lemma follows from Corollary 3. □

To count the number of profiles in Φ^n, we observe the following correspondence with unordered partitions of positive integers.

An *unordered partition* of a positive integer n is a multiset of positive integers whose sum is n. An unordered partition can be represented by the vector $\overline{\varphi} = (\varphi_n, \ldots, \varphi_1)$ where φ_μ denotes the number of times μ occurs in the partition. For example, the partition $\{1, 1, 3\}$ of 5 corresponds to the vector $(0, 0, 1, 0, 2)$. Unordered partitions of a positive integer n and profiles of patterns in Ψ^n are equivalent as follows.

Lemma 22. A vector $\overline{\varphi}$ is an unordered partition of n iff $\overline{\varphi} \in \Phi^n$. □

Henceforth we use the notation developed for profiles of patterns in Section 5.3.2 for unordered partitions also.

Lemma 23. [Hardy and Ramanujan [50], see also [63]] The number of unordered partitions of n is

$$\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}(1-o(1))\right) \le |\Phi^n| \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right). \qquad \square$$

Lemmas 21 and 23 imply the following upper bound on the pattern redundancy of $\mathcal{I}^n$.

Theorem 24. For all n,

$$\hat{R}(\mathcal{I}^n_\Psi) \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}. \qquad \square$$

In particular, the pattern redundancy of *i.i.d.* strings is sublinear in the block-length, and hence the per-symbol redundancy diminishes as the number of compressed symbols increases. We note that the number of integer partitions has also been used by Csiszár and Shields [64] to bound the redundancy of renewal processes.

5.4.3 Lower bound

In the last section we showed that the redundancy of patterns of *i.i.d.* strings is $\mathcal{O}(n^{1/2})$. We now show that it is $\Omega(n^{1/3})$. We provide a simple proof of this lower bound and mention a more complex approach that yields the same growth rate, but with a higher multiplicative constant.

Theorem 25. As n increases,

$$\hat{R}(\mathcal{I}^n_\Psi) \ge \log\left(\frac{e^{23/12}}{\sqrt{2\pi}}\right)\cdot n^{\frac{1}{3}}(1+o(1)).$$

Proof Let

$$\Psi_p^{-1}(\overline{\psi}) = \{\overline{x}\in\mathcal{A}^* : \Psi(\overline{x}) = \overline{\psi} \text{ and } p(\overline{x}) > 0\}$$

be the *support* of a pattern $\overline{\psi}$ with respect to a distribution p over an alphabet $\mathcal{A}$. As noted in [19], $\Psi_p^{-1}(\overline{\psi})$ can be partitioned into sets, each with $\prod_\mu \varphi_\mu!$ equi-probable sequences, where $\varphi = (\varphi_n, \ldots, \varphi_1)$ is the profile of $\overline{\psi}$. By standard maximum-likelihood arguments, the probability of any sequence with profile φ is at most $\prod_{\mu=1}^{n} \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}$, hence

$$\begin{aligned}
\sup_{p \in \mathcal{I}_\Psi^n} p(\overline{\psi}) &= \sup_{p \in \mathcal{I}^n} p\big(\Psi_p^{-1}(\overline{\psi})\big) \\
&\geq \prod_{\mu} \varphi_\mu! \cdot \max_{p \in \mathcal{I}^n} \max_{\overline{x} \in \Psi_p^{-1}(\overline{\psi})} p(\overline{x}) \\
&\geq \prod_{\mu} \varphi_\mu! \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}. \qquad (5.4)
\end{aligned}$$

From Shtarkov's sum (2.1),

$$\begin{aligned}
\hat{R}(\mathcal{I}_\Psi^n) &= \log\left(\sum_{\overline{\varphi} \in \Phi^n} \sum_{\overline{\psi} \in \Psi_{\overline{\varphi}}} \sup_{p \in \mathcal{I}_\Psi^n} p(\overline{\psi})\right) \\
&\overset{(a)}{\geq} \log\left(\sum_{\overline{\varphi} \in \Phi^n} \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!} \cdot \prod_{\mu=1}^{n} \varphi_\mu! \Big(\frac{\mu}{n}\Big)^{\mu\varphi_\mu}\right) \\
&\overset{(b)}{\geq} \log\left(\frac{e^n n!}{n^n} \sum_{m=1}^{n} \sum_{\overline{\varphi} \in \Phi_m^n} \frac{1}{\sqrt{\prod_{\mu=1}^{n} \mu^{\varphi_\mu}}} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{e^{m/12}}\right) \\
&\overset{(c)}{\geq} \log\left(\sum_{m=1}^{n} \sum_{\overline{\varphi} \in \Phi_m^n} \left(\frac{m}{n}\right)^{m/2} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{e^{m/12}}\right) \\
&\overset{(d)}{\geq} \log\left(\sum_{m=1}^{n} \binom{n-1}{m-1} \frac{1}{m!} \cdot \left(\frac{m}{n}\right)^{m/2} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{e^{m/12}}\right) \\
&\geq \log\left(\binom{n-1}{m-1} \cdot \frac{1}{m!} \cdot \left(\frac{m}{n}\right)^{m/2} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{e^{m/12}}\right)\Bigg|_{m=n^{\frac{1}{3}}} \\
&\overset{(e)}{\geq} \log\Bigg(\frac{m}{n} \frac{1+o(1)}{\sqrt{2\pi m}} \left(\frac{n\,e}{m}\right)^m \cdot \frac{1}{\sqrt{2\pi m} e^{\frac{1}{12m}}} \frac{e^m}{m^m} \cdot \\
&\qquad\qquad \left(\frac{m}{n}\right)^{m/2} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{e^{m/12}}\Bigg)\Bigg|_{m=n^{\frac{1}{3}}} \\
&\geq \log\left(\frac{e^{2n^{1/3}}}{\left(\sqrt{2\pi}\right)^{n^{1/3}} e^{n^{1/3}/12}}\right)(1+o(1)) \\
&= \log\left(\frac{e^{23/12}}{\sqrt{2\pi}}\right) \cdot n^{1/3}(1+o(1)),
\end{aligned}$$

where (a) follows from Lemma 20 and Equation (5.4), (b) from Feller's bounds (2.3), (c) from the arithmetic-geometric mean inequality, (d) because each unordered partition into m parts can be ordered in at most $m!$ ways, and (e) from Lemma 7. The theorem follows. □

Note that the constant in the bound can be increased by taking

$$m = \left(2\pi e^{-5/6}\right)^{-1/3} \cdot n^{1/3}$$

in the proof, yielding

$$\hat{R}(\mathcal{I}_{\Psi}^{n}) \geq \left(2\pi e^{-5/6}\right)^{-1/3} \cdot \frac{3}{2} \log e \cdot n^{1/3}(1+o(1)).$$

Generating functions and Hayman's Theorem can be used to evaluate the exact asymptotic growth of

$$\log\left(\sum_{\overline{\varphi}\in\Phi^n} \frac{n!}{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!} \prod_{\mu=1}^{n}\left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}\right), \tag{5.5}$$

thereby improving the lower bound to the following.

Theorem 26. As n increases,

$$\hat{R}(\mathcal{I}_{\Psi}^{n}) \geq \left(\frac{3}{2}\log e\right) n^{\frac{1}{3}}(1+o(1)). \qquad \square$$

Proof The proof is fairly involved, and is defered to the next chapter.

These lower bounds should be compared with those in Åberg, Shtarkov, and Smeets [19] who lower bounded pattern redundancy when the number m of symbols is fixed and finite and the block length n increases to infinity. While it is not clear whether their proof extends to arbitrary m, which may grow with n, the bound they derive may still hold in general. If so, it would yield a lower bound similar to those described here. For a more complete discussion, see [65]. Note also that subsequent to the derivation of Theorem 13, Shamir *et. al.* [22, 23] showed that the average-case pattern redundancy is lower bounded by $(\pi/2)^{1/3}\, 1.5 \log e\, n^{(1-\epsilon)/3}$ for arbitrarily small ϵ.

5.4.4 Distributions for block coding of patterns

We summarize some results on probability distributions relevant to the results we have obtained so far. For any n, $\overline{\varphi} \in \Phi^n$ and pattern $\psi_1^n \in \Psi_{\overline{\varphi}}$, let

$$\overline{p}_{\overline{\varphi}}(\psi_1^n) \stackrel{\text{def}}{=} \frac{1}{N(\varphi)} = \frac{\prod_{\mu=1}^{n} \mu!^{\varphi_\mu}\varphi_\mu!}{n!}$$

be the uniform distribution on patterns of profile $\overline{\varphi}$. As mentioned earlier, this uniform distribution upper bounds the maximum likelihood probability from $\mathcal{I}_\Psi^n$, namely for all ψ_1^n,

$$\overline{p}_{\varphi(\psi_1^n)}(\psi_1^n) \geq \hat{p}_{\psi_1^n}(\psi_1^n),$$

where we let

$$\hat{p}^{\psi_1^n} \stackrel{\text{def}}{=} \arg\max_{p \in \mathcal{I}_\Psi^n} p(\psi_1^n)$$

denote the distribution that maximizes the induced *i.i.d.* probability of ψ_1^n. For all $\psi_1^n \in \Psi^n$, the distribution

$$\tilde{p}(\psi_1^n) = \frac{\overline{p}_{\varphi(\psi_1^n)}(\psi_1^n)}{\sum_{\overline{\psi} \in \Psi^n} \overline{p}_{\varphi(\overline{\psi})}(\overline{\psi})}$$

over Ψ^n assigns a probability that is smaller than $\overline{p}_{\varphi(\psi_1^n)}(\psi_1^n)$ by a factor, which by Theorem 24, is at most $\Theta(\sqrt{n})$, namely

$$\sum_{\overline{\psi} \in \Psi^n} \overline{p}_{\varphi(\overline{\psi})}(\overline{\psi}) \leq \exp\left(\pi \sqrt{\frac{2}{3}} \sqrt{n}\right).$$

Therefore,

$$\tilde{p}(\psi_1^n) \geq \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\exp\left(\pi \sqrt{\frac{2}{3}} \sqrt{n}\right)},$$

which, of course, is the upper bound on $\hat{R}(\mathcal{I}_\Psi^n)$.

5.5 Sequential compression

The compression schemes considered so far operated on the whole block of symbols. In many applications the symbols arrive and must be encoded sequentially. Compression schemes for such applications are called *sequential* and associate with every pattern $\psi_1^n \in \Psi^n$, a probability distribution $q(x|\psi_1^n)$ over

$$[\max(\psi_1^n) + 1] = \{1, \ldots, \max(\psi_1^n) + 1\},$$

representing the probability that the encoder assigns to the possible values of ψ_{n+1} after seeing ψ_1^n. For example, $q(x|\Lambda) \stackrel{\text{def}}{=} q(x)$ is a distribution over $\{1\}$, namely, $q(1) = 1$, $q(x|1)$ and $q(x|11)$ are distributions over $\{1, 2\}$, while $q(x|12)$ is a distribution over $\{1, 2, 3\}$.

Let q be a sequential encoder. For each $n \in \mathbb{Z}^+$, q induces a probability distribution over Ψ^n given by

$$q(\psi_1^n) \overset{\text{def}}{=} \prod_{i=1}^{n} q(\psi_i|\psi_1^{i-1}).$$

Some simple algorithms along the lines of the add-constant rules were analyzed in [21] and shown to have diminishing per-symbol redundancy when the number of distinct symbols is small, but a constant per-symbol redundancy in general.

In this section, we describe two sequential encoders with diminishing per-symbol redundancy. The analysis of their redundancy is deferred to Chapter 8 The first encoder, $q_{1/2}$, has worst-case redundancy of at most

$$\frac{4\pi \log e}{\sqrt{3}(2-\sqrt{2})}\sqrt{n},$$

only slightly higher than the upper bound on $\hat{R}(\mathcal{I}_\Psi^n)$. However this encoder has high computational complexity, and in Section 5.5.2 we consider a sequential encoder, $q_{2/3}$, with linear computational complexity and redundancy less than

$$10n^{2/3},$$

which still grows sublinearly with n though not as slowly as the block redundancy.

5.5.1 A low redundancy encoder

We construct an encoder $q_{1/2}$ that for all n, and all patterns ψ_1^n achieves a redundancy

$$\hat{R}(\mathcal{I}_\Psi^n, q_{1/2}) \le \frac{4\pi \log e}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}.$$

The encoder uses distributions that are implicit in the block coding results.

Let

$$\hat{p}_{\psi_1^n}(\psi_1^n) \overset{\text{def}}{=} \max_{p \in \mathcal{I}_\Psi^n} p(\psi_1^n)$$

denote the maximum-likelihood probability assigned to a pattern $\psi_1^n \in \Psi^n$ by any *i.i.d.* distribution in $\mathcal{I}_\Psi^n$. Recall that $N(\overline{\varphi})$ is the number of patterns with profile $\overline{\varphi}$, and that, as in Lemma 21, every *i.i.d.* distribution assigns the same probability to all patterns of the same profile. We therefore obtain the following upper bound on the maximum pattern probabilities.

Lemma 27. For any pattern $\psi_1^n \in \Psi^n$ of profile $\overline{\varphi} \in \Phi^n$,

$$\hat{p}_{\psi_1^n}(\psi_1^n) \leq \frac{1}{N(\varphi(\psi_1^n))}. \qquad \square$$

Based on this upper bound, we can construct the following distribution over Ψ^n,

$$\tilde{p}(\psi_1^n) \overset{\text{def}}{=} \frac{\frac{1}{N\left(\varphi(\psi_1^n)\right)}}{\sum_{\overline{\psi}\in\Psi^n} \frac{1}{N\left(\varphi(\overline{\psi})\right)}} = \frac{1}{N(\varphi(\psi_1^n))\,|\Phi^n|}. \tag{5.6}$$

For $n \geq 1$, let

$$t_n = 2^{\lceil \log n \rceil}$$

be the smallest power of 2 that is at least n, *e.g.* $t_1 = 1$, $t_2 = 2$, and $t_3 = t_4 = 4$. Note that $\frac{t_n}{2} < n \leq t_n$.

For every $k \geq n$, and patterns ψ_1^n, let

$$\Psi^k(\psi_1^n) = \{\overline{y} \in \Psi^k : y_1 y_2 \dots y_n = \psi_1^n\}$$

be the set of all patterns that extend ψ_1^n in Ψ^k, and let

$$\tilde{p}^k(\psi_1^n) \overset{\text{def}}{=} \tilde{p}(\Psi^k(\psi_1^n)) = \sum_{\overline{y}\in\Psi^k(\psi_1^n)} \tilde{p}(\overline{y})$$

be the probability of the set $\Psi^k(\psi_1^n)$ under the distribution $\tilde{p}$.

The encoder assigns

$$q_{1/2}(1) = 1,$$

and for all $n > 1$ and $\psi_1^n \in \Psi^n$ it assigns the conditional probability

$$q_{1/2}(\psi_n|\psi_1^{n-1}) = \frac{\tilde{p}^{t_n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^{n-1})} \tag{5.7}$$

5.5.2 A low complexity encoder

The evaluation of $q_{1/2}(\psi_1^n|\psi_1^{n-1})$ in Equation (5.7) may take super polynomial time, hence implementing the encoder described in Section 5.5.1 may be impractical. Therefore we present a linear-complexity encoder q whose per-symbol redundancy diminishes to zero as the blocklength increases.

For notational convenience, let the prevelence of μ in ψ_1^n, $\varphi_\mu(\psi_1^n)$, be written as φ_μ^n, and let

$$r \stackrel{\text{def}}{=} \mu_{\psi_n}(\psi_1^{n-1}).$$

For $c \in \mathbb{Z}^+$, let

$$f_c(\varphi) \stackrel{\text{def}}{=} \max(\varphi, c) = \begin{cases} c, & 0 \le \varphi \le c-1 \\ \varphi, & \varphi \ge c, \end{cases}$$

let

$$g_c(\varphi) \stackrel{\text{def}}{=} \prod_{i=1}^{\varphi} f_c(i) = \begin{cases} c^\varphi, & 0 \le \varphi \le c-1 \\ \frac{c^c}{c!}\varphi!, & \varphi \ge c. \end{cases}$$

and let c be the sequence

$$c[n] \stackrel{\text{def}}{=} \lceil n^{\frac{1}{3}} \rceil.$$

The encoder assigns

$$q_{2/3}(1) = 1,$$

and for all $n > 1$, and $\psi_1^n \in \Psi^n$, it assigns the conditional probability

$$q_{2/3}(\psi_1^n | \psi_1^{n-1}) = \frac{1}{S_{c[n-1]}(\psi_1^{n-1})} \cdot \begin{cases} f_{c[n-1]}\big(\varphi_1^{n-1} + 1\big), & r = 0 \\ (r+1)\frac{f_{c[n-1]}\left(\varphi_{r+1}^{n-1}+1\right)}{f_{c[n-1]}\left(\varphi_r^{n-1}\right)}, & r > 0, \end{cases}$$

where

$$S_{c[n-1]}(\psi_1^{n-1}) \stackrel{\text{def}}{=} f_{c[n-1]}\big(\varphi_1^{n-1}+1\big) + \sum_{\mu=1}^{n} \varphi_\mu^{n-1} \cdot (\mu+1) \frac{f_{c[n-1]}\left(\varphi_{\mu+1}^{n-1}+1\right)}{f_{c[n-1]}\left(\varphi_\mu^{n-1}\right)}$$

is the normalization factor.

It follows by induction that for all $n > 1$ and all patterns $\psi_1^n \in \Psi^n$,

$$q_{2/3}(\psi_1^n) = \frac{\prod_{\mu=1}^{n} \left((\mu!)^{\varphi_\mu^n} g_{c[n]}(\varphi_\mu^n)\right)}{\prod_{i=1}^{n-1} S_{c[i]}(\psi_1^i)} \cdot \prod_{i=1}^{n-1} \left(\prod_{\mu=1}^{i} \frac{g_{c[i]}(\varphi_\mu^i)}{g_{c[i+1]}(\varphi_\mu^i)} \right).$$

Theorem 28. For all n,

$$\hat{R}(\mathcal{I}_\Psi^n, q) \le \mathcal{O}(n^{\frac{2}{3}}).$$

where the implied constant C is less than 10.

Proof See [2]. □

Acknowledgements

We would like to thank N. Jevtić, A. Noga, G. Shamir, W. Szpankowski, K. Viswanathan for helpful discussions. We also thank G. Shamir for pointing out previous work done on this topic. This chapter is adapted from 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. The author was a primary researcher of the sections reproduced here.

Chapter 6

Saddle point analysis bound on pattern redundancy

We saw that, irrespective of the alphabet size, patterns of *i.i.d.* distributed strings can be compressed with redundancy of at most

$$\hat{R}(\mathcal{I}_\Psi^n) \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}$$

bits. Hence as the blocklength n grows, the redundancy of patterns increases sublinearly with n, and the per-symbol redundancy diminishes to zero, even for infinite alphabets.

In this chapter we prove the improved lower bound on $\hat{R}(\mathcal{I}_\Psi^n)$ presented in the Theorem 26. To do so, we lower bound the highest probability of a pattern $\overline{\psi}$ by the highest probability of any single *i.i.d.* string whose pattern is $\overline{\psi}$. We obtain,

$$\hat{R}(\mathcal{I}_\Psi^n) \ge \hat{R}^-(\mathcal{I}_\Psi^n) \stackrel{\text{def}}{=} \log\left(\sum_{\overline{\psi}\in\Psi^n}\prod_{\mu=1}^{n}\left(\frac{\mu}{n}\right)^{\mu\varphi_\mu(\overline{\psi})}\right),$$

where $\varphi_\mu(\overline{\psi})$ is the number of symbols appearing μ times in $\overline{\psi}$. $\hat{R}^-(\mathcal{I}_\Psi^n)$ is of mathematical interest of its own and its simple formulation allows for a precise evaluation of its growth order.

To prove Theorem 26, we use Hayman's saddle point analysis on its generating function to show that

$$\hat{R}^-(\mathcal{I}_\Psi^n) = \left(\frac{3}{2}\log e\right)n^{\frac{1}{3}} - \frac{1}{3}\log n - \frac{2}{3}\log e - \frac{1}{2}\log 3 + o(1). \tag{6.1}$$

6.1 The generating function

As mentioned earlier, it is difficult to obtain the maximum probability of patterns. Instead, we lower bound these probabilities of patterns, and use Shtarkov's sum to derive a lower bound on redundancy.

Let

$$\Psi_p^{-1}(\overline{\psi}) = \{\overline{x} \in \mathcal{A}^* : \Psi(\overline{x}) = \overline{\psi} \text{ and } p(\overline{x}) > 0\}$$

be the *support* of a pattern $\overline{\psi}$ with respect to a distribution p. For every $\overline{\psi} \in \Psi^n$,

$$\max_{p \in \mathcal{I}_\Psi^n} p(\overline{\psi}) = \max_{p \in \mathcal{I}^n} p\big(\Psi_p^{-1}(\overline{\psi})\big) \geq \max_{p \in \mathcal{I}^n} \max_{\overline{x} \in \Psi_p^{-1}(\overline{\psi})} p(\overline{x}).$$

Let the number of symbols occuring μ times in $\overline{\psi}$ be φ_μ. Standard maximum-likelihood arguments imply that

$$\max_{p \in \mathcal{I}^n} \max_{\overline{x} \in \Psi_p^{-1}(\overline{\psi})} p(\overline{x}) = \prod_{\mu=1}^{n} \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu},$$

hence

$$\max_{p \in \mathcal{I}_\Psi^n} p(\overline{\psi}) \geq \prod_{\mu=1}^{n} \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}. \tag{6.2}$$

Let

$$\Phi^n = \{(\varphi_1, \ldots, \varphi_n) : \varphi_i \geq 0, \sum_{\mu=1}^{n} \mu\varphi_\mu = n\},$$

and

$$\Psi_{\overline{\varphi}} = \{\overline{\psi} : \varphi_\mu \text{ symbols appear } \mu \text{ times in pattern } \overline{\psi}\}.$$

Incorporating (6.2) into Shtarkov's sum (2.1), we obtain

$$\begin{aligned}
\hat{R}(\mathcal{I}_\Psi^n) &= \log\left(\sum_{\overline{\varphi} \in \Phi^n} \sum_{\overline{\psi} \in \Psi_{\overline{\varphi}}} \max_{p \in \mathcal{I}_\Psi^n} p(\overline{\psi})\right) \\
&\geq \log\left(\sum_{\overline{\varphi} \in \Phi^n} \sum_{\overline{\psi} \in \Psi_{\overline{\varphi}}} \prod_{\mu=1}^{n} \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}\right) \\
&= \log\left(\sum_{\overline{\varphi} \in \Phi^n} \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!} \prod_{\mu=1}^{n} \left(\frac{\mu}{n}\right)^{\mu\varphi_\mu}\right) \\
&\overset{\text{def}}{=} \log g(n).
\end{aligned} \tag{6.3}$$

Direct computation of $g(n)$ appears to be difficult. Instead, we evaluate a *generating function* of $g(n)$,

$$G(z) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} g(n) \frac{n^n}{n!} z^n, \tag{6.4}$$

from which the asymptotics of $g(n)$ can be obtained using *Hayman's* analysis [66].

To express the generating function $G(z)$ in a more explicit form, observe that

$$\begin{aligned} G(z) &= \sum_{n=0}^{\infty} \sum_{(\varphi_1,\ldots,\varphi_n)\in\Phi^n} \prod_{\mu\geq 1} \left(\frac{\mu^\mu z^\mu}{\mu!}\right)^{\varphi_\mu} \frac{1}{\varphi_\mu!} \\ &= \sum_{(\varphi_1,\ldots,\varphi_n)\in\Phi^n} \prod_{\mu\geq 1} \left(\frac{\mu^\mu z^\mu}{\mu!}\right)^{\varphi_\mu} \frac{1}{\varphi_\mu!} \\ &= \prod_{\mu\geq 1} \sum_{(\varphi_1,\ldots,\varphi_n)\in\Phi^n} \left(\frac{\mu^\mu z^\mu}{\mu!}\right)^{\varphi_\mu} \frac{1}{\varphi_\mu!}, \end{aligned}$$

thus yielding

$$G(z) = \exp\left(\sum_{k=1}^{\infty} \frac{k^k z^k}{k!}\right). \tag{6.5}$$

6.2 Hayman's analysis

In the last section, we lower bounded $\hat{R}(\mathcal{I}_\Psi^n)$ in terms of the coefficients of a generating function $G(z)$. Hayman [66] developed a technique to compute the asymptotics of the coefficients of power series that satisfy certain properties, which, as shown later, $G(z)$ also satisfies. In this section we describe Hayman's analysis. We follow the terminology used in [67].

Theorem 29. **[Hayman]** For

$$f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

let

$$a(z) \stackrel{\text{def}}{=} \frac{d \log f(z)}{d \log z} \quad \text{and} \quad b(z) \stackrel{\text{def}}{=} \frac{d^2 \log f(z)}{d(\log z)^2} = z a'(z), \tag{6.6}$$

and let the *saddle point* r_n be the solution of

$$a(r_n) = n.$$

If for some real R_1, the following three conditions hold:

Nonnegativity: $\exists R_0 < R_1$ such that for $R_0 < x < R_1$,

$$f(x) \geq 0;$$

Fast growth: As $x \to R_1 - 0$, namely, x approaches R_1 from below, $b(x) \to \infty$;

Basic split: $\exists \phi(x) > 0$, called the *basic split* such that

Local approximation: for $|\theta| \leq \phi(x)$, uniformly in θ as $x \to R_1$

$$f(xe^{i\theta}) \sim f(x) \exp(ia(x)\theta - \frac{\theta^2}{2} b(x));$$

Fast taper: for $\phi(x) < |\theta| < \pi$, uniformly in θ as $x \to R_1$

$$f(xe^{i\theta}) \sim \frac{o(f(x))}{\sqrt{b(x)}};$$

then,

$$a_n \sim \frac{f(r_n)}{r_n^n \sqrt{2\pi b(r_n)}}. \qquad \square$$

To understand Hayman's Theorem, note that by Cauchy's integral formula,

$$\int_{|z|=r} \frac{1}{z^n} dz = \begin{cases} 2\pi i & \text{if } n = 1, \\ 0 & \text{if } n \neq 1, \end{cases}$$

hence, for any r within the radius of convergence of $f(z)$, the coefficients of f can be expressed as

$$\begin{aligned} a_n &= \frac{1}{2\pi i} \int_{|z|=r} \frac{f(z)}{z^{n+1}} dz \\ &= \frac{1}{2\pi i} \int_{\substack{|z|=r \\ |\theta| \leq \phi(r)}} \frac{f(z)}{z^{n+1}} dz + \frac{1}{2\pi i} \int_{\substack{|z|=r \\ |\theta| \geq \phi(r)}} \frac{f(z)}{z^{n+1}} dz \\ &\overset{\text{def}}{=} I_1 + I_2. \end{aligned}$$

Hayman chose the radius such that the fast taper condition implies that I_2 is a negligible fraction of I_1.

To evaluate I_1, let $z = re^{i\theta}$. Using the McLaurin expansion of $\log f(re^{i\theta})$ with respect to θ,

$$\begin{aligned}
\log f(re^{i\theta}) &= \log f(r) + \sum_{k=1}^{\infty} \frac{\theta^k}{k!}\left(\frac{d}{d\theta}\right)^k \log f(re^{i\theta})\big|_{\theta=0} \\
&= \log f(r) + \sum_{k=1}^{\infty} \frac{i^k\theta^k}{k!}\left(\frac{d}{d(\log z)}\right)^k \log f(z)\big|_{z=r} \\
&= \log f(r) + ia(r)\theta - b(r)\frac{\theta^2}{2} + \sum_{k=3}^{\infty} \frac{i^k\theta^k}{k!}\left(\frac{d}{d(\log z)}\right)^k \log f(z)\big|_{z=r} \qquad (6.7)
\end{aligned}$$

where the second inequality follows because $d\theta = d(\log z)/i$. Hence,

$$f(re^{i\theta}) = f(r)\exp\left(ia(r)\theta - \frac{b(r)}{2}\theta^2 + \sum_{k=3}^{\infty} \frac{i^k\theta^k}{k!}\left(\frac{d}{d(\log z)}\right)^k \log f(z)\big|_{z=r}\right),$$

As $r \to R_1$, the local approximation condition for $|\theta| \leq \phi$ implies that the trailing sum is negligible. Therefore,

$$I_1 = \frac{1}{2\pi i}\int_{\substack{|z|=r \\ |\theta|\leq\phi(r)}} \frac{f(z)}{z^{n+1}}dz = \frac{f(r)}{2\pi r^n}\int_{\theta=-\phi(r)}^{\theta=\phi(r)} \exp\left(i(a(r)-n)\theta - b(r)\frac{\theta^2}{2}\right)d\theta.$$

Taking the radius to be r_n such that $a(r_n) = n$, we obtain

$$\begin{aligned}
I_1 &= \frac{f(r_n)}{2\pi r_n^n}\int_{\theta=-\phi(r_n)}^{\theta=\phi(r_n)} \exp\left(-b(r_n)\frac{\theta^2}{2}\right)d\theta \\
&= \frac{f(r_n)}{2\pi r_n^n\sqrt{b(r_n)}}\int_{y=-\sqrt{b(r_n)}\phi(r_n)}^{y=\sqrt{b(r_n)}\phi(r_n)} e^{-y^2/2}dy \\
&\overset{(a)}{\sim} \frac{f(r_n)}{2\pi r_n^n\sqrt{b(r_n)}}\int_{y=-\infty}^{y=\infty} e^{-y^2/2}dy \\
&= \frac{f(r_n)}{r_n^n\sqrt{2\pi b(r_n)}},
\end{aligned}$$

where (a) follows because it can be shown [66] that $b(r_n)\phi(r_n)^2 \to \infty$ as $n \to \infty$. It can be shown [66] that evaluating I_1 at any radius r yields

$$|I_1| = \frac{f(r)}{r^n\sqrt{2\pi b(r)}}\left(\exp\left(-\frac{(a(r)-n)^2}{b(r)}\right) + o_r(1)\right),$$

where the $o_r(1)$ is uniform over n, and diminishes to zero as $r \to R_1$.

As mentioned before, a_n is to be approximated by I_1, hence we bound the contribution of I_2 as follows,

$$|I_2| \leq \frac{1}{2\pi} \int_{\substack{|z|=r \\ |\theta| \geq \phi(r_n)}} \left| \frac{f(z)}{z^{n+1}} \right| dz$$
$$\overset{(a)}{=} o_r \left(\frac{f(r)}{\sqrt{b(r)} r^n} \right)$$

where (a) follows from the fast taper condition. Observe that at $r = r_n$, $I_2 = o(I_1)$, and that as $n \to \infty$, $r_n \to R_1$, implying

$$a_n \sim |I_1|_{r_n} \sim \frac{f(r_n)}{r_n^n \sqrt{2\pi b(r_n)}}.$$

Hayman's analysis essentially identifies a circle along which Cauchy's integral is captured by the contribution of an arc around the real line, small enough that the value of the integrand along the arc is well approximated by at most the quadratic terms of (6.7).

Hayman's analysis can also be viewed as a special case of the class of *saddle point approximations*. A *saddle point* of an analytic function is any point where the function does not vanish, but its derivative vanishes. The surface representing the modulus of any analytic function has no maxima and no minima but for isolated zeros of the function. Therefore, it can be shown *e.g.* [68], that any (simple) saddle point is the intersection of two level curves at right angles to each other. One of the lines bisecting the level curves is the direction of steepest descent from the saddle point, while the other line bisecting the level curve is the direction of steepest ascent. This leads to a saddle shape for the surface, hence the name.

The saddle point approximation uses a contour, in this case a circle centered at the origin, through a saddle point of the integrand in order to exploit the fact [68, 67] that a short arc that captures the contribution of the integral around the contour is likely to be around the saddle point of the integrand.

In the context of Hayman's analysis, for any admissible function f, the saddle point method's choice for the contour lets I_2 be a negligible fraction of I_1 near the point where the derivative of $f(z)/z^{n+1}$ vanishes, equivalently the solution of

$$(\ln f(z) - (n+1)\ln z)' = \frac{1}{z}(a(z) - (n+1)) = 0,$$

the saddle point on the surface $|f(z)/z^{n+1}|$. Note that for Hayman admissible functions, the saddle point is real and positive, and the saddle point is a maxima perpendicular to the real line, and a minima along the positive real line. For more details on the saddle point approximation and related results, see [66, 68, 67].

For the generating function G defined in Equation (6.5), the functions $a(z)$ and $b(z)$ of Equation (6.6) are

$$a(z) = \sum_{k=1}^{\infty} \frac{k^{k+1} z^k}{k!} \quad \text{and} \quad b(z) = \sum_{k=1}^{\infty} \frac{k^{k+2} z^k}{k!}. \tag{6.8}$$

We pick $R_1 = \frac{1}{e}$. The first two conditions are clearly satisfied for $G(z)$. For $\phi(x) = (1-ex)^{\frac{6}{5}}$, we show in Theorem 34, that the local approximation for G holds, and in Theorem 35 that $|G(z)|$ does drop rapidly for $|\theta| \geq \phi(x)$.

6.3 Preliminaries

We outline some results that will be extensively used in this chapter.

Observe that we can expand $G(xe^{i\theta})$ in θ as

$$G(xe^{i\theta}) = G(x)\exp\left(\sum_{l=1}^{\infty} \frac{(i\theta)^l}{l!} \frac{d^l \log G(z)}{d(\log z)^l}\bigg|_{z=x}\right) = G(x)\exp\left(\sum_{l=1}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!}\bigg|_{z=x}\right).$$

We first check for convergence of each of the summations over k. Indeed,

Lemma 30. For any l, $\sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!}$ converges for $x < \frac{1}{e}$.

Proof By the Cauchy ratio test, *e.g.* [69]. □

Therefore, in order to evaluate the n'th coefficient in the Taylor series, Hayman's theorem approximates the value of $G(z)$ in the complex integration over the circle $|z| = x$ by a correction over the value $G(x)$ for points on the circle near the positive real line, and by a term much smaller than $G(x)$ for points on the circle away from the positive real line.

Intuitively speaking it follows that at the basic split ϕ, the contribution of higher order terms is negligible and that the contribution of the second coefficient is large enough to satisfy fast taper. We choose a ϕ based on these criteria, and then prove that our choice indeed works.

Further, we shall denote by C positive constants that are, in particular, independent of x, θ and l.

6.4 Locating the basic split

We locate the basic split ϕ for

$$G(xe^{i\theta}) = G(x)\exp\left(\sum_{l=1}^{\infty}\frac{(i\theta)^l}{l!}\sum_{k=1}^{\infty}\frac{k^{k+l}x^k}{k!}\right).$$

To do so, we estimate the magnitude of the coefficients of θ, and ensure that at our choice of the ϕ, the second term is unbounded, and the contribution of any term beyond the second is negligible. In Theorems 34 and 35, we show that this choice works.

We upper bound the magnitude of the coefficients of θ as follows.

Lemma 31. For integers $l \geq 2$ and $x < \frac{1}{e}$,

$$\sum_{k=1}^{\infty}\frac{k^{k+l}x^k}{k!} \leq \frac{\sqrt{(2l)!}}{2^l(1-ex)^{l+\frac{1}{2}}}.$$

Proof From Feller's bounds (2.3),

$$\sum_{k=1}^{\infty}\frac{k^{k+l}x^k}{k!} \leq \frac{1}{\sqrt{2\pi}}\sum_{k=1}^{\infty}\frac{k^l x^k e^k}{\sqrt{k}}.$$

Squaring the right side,

$$\begin{aligned}\left(\sum_{k=1}^{\infty}\frac{k^l x^k e^k}{\sqrt{k}}\right)^2 &= \sum_{k=2}^{\infty}x^k e^k\sum_{m=1}^{k-1}((k-m)\cdot m)^{l-\frac{1}{2}}\\ &\leq \sum_{k=2}^{\infty}x^k e^k(k-1)\left(\frac{k}{2}\cdot\frac{k}{2}\right)^{l-\frac{1}{2}}\\ &\leq \frac{1}{2^{2l-1}}\sum_{k=0}^{\infty}k^{2l}x^k e^k\\ &\leq \frac{(2l)!}{2^{2l-1}}\sum_{k=0}^{\infty}\binom{k+2l}{2l}x^k e^k\\ &= \frac{(2l)!}{2^{2l-1}\cdot(1-ex)^{2l+1}}.\end{aligned}$$

Taking the positive square root proves the lemma. □

We lower bound the magnitude of the coefficient of θ^2 as follows.

Lemma 32. For $\frac{5}{6e} < x < \frac{1}{e}$,

$$\sum_{k=1}^{\infty} \frac{k^{k+2}x^k}{k!} \geq \frac{C}{(1-ex)^{\frac{5}{2}}}.$$

Proof From Feller's bounds (2.3),

$$\sum_{k=1}^{\infty} \frac{k^{k+2}x^k}{k!} \geq C_1 \sum_{k=1}^{\infty} \frac{k^2 x^k e^k}{\sqrt{k}}.$$

Squaring the right side,

$$\begin{aligned}
\left(\sum_{k=1}^{\infty} \frac{k^2 x^k e^k}{\sqrt{k}}\right)^2 &= \sum_{k=2}^{\infty} x^k e^k \sum_{m=1}^{k-1} ((k-m)\cdot m)^{\frac{3}{2}} \\
&\geq \sum_{k=2}^{\infty} x^k e^k \sum_{m=\lfloor \frac{k}{4} \rfloor}^{\lceil \frac{3k}{4} \rceil} ((k-m)\cdot m)^{\frac{3}{2}} \\
&\geq \sum_{k=2}^{\infty} x^k e^k \cdot \frac{k}{2} \left(\frac{3k}{4} \cdot \frac{k}{4}\right)^{\frac{3}{2}} \\
&= C_2 \sum_{k=2}^{\infty} k^4 x^k e^k \\
&\geq C_2 \cdot 4! \cdot \sum_{k=2}^{\infty} \binom{k}{4} x^k e^k \\
&= C_2 \cdot 4! \cdot (xe)^4 \cdot \sum_{k=4}^{\infty} \binom{k}{4} x^{k-4} e^{k-4} \\
&\geq \frac{C_3}{(1-ex)^5}.
\end{aligned}$$

In the last step we observed that $\frac{5}{6} < xe < 1$, and thus included it in the constant. Taking the positive square root proves the lemma. □

The following Lemma locates the basic split.

Lemma 33. $\exists \phi(x)$ so that

$$\lim_{x \to \frac{1}{e}} \phi(x)^2 \sum_{k=1}^{\infty} \frac{k^{k+2}x^k}{k!} \to \infty$$

and simultaneously for $l \geq 3$,

$$\lim_{x\to\frac{1}{e}} \phi(x)^l \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} = 0.$$

Proof Take $\phi(x) = (1-ex)^{\alpha}$ with $\frac{7}{6} < \alpha < \frac{5}{4}$. From Lemma 32,

$$\lim_{x\to\frac{1}{e}} \phi(x)^2 \sum_{k=1}^{\infty} \frac{k^{k+2}x^k}{k!} \geq \lim_{x\to\frac{1}{e}} \frac{C\cdot\phi(x)^2}{(1-ex)^{\frac{5}{2}}} \to \infty$$

because $\alpha < \frac{5}{4}$, and from Lemma 31,

$$\lim_{x\to\frac{1}{e}} \phi(x)^l \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} \leq \lim_{x\to\frac{1}{e}} \frac{\phi(x)^l\sqrt{(2l)!}}{2^l(1-ex)^{l+\frac{1}{2}}} = 0$$

because $\alpha > \max_{l\geq 3}(1+\frac{1}{2l}) = \frac{7}{6}$. Therefore all $\phi(x) = (1-ex)^{\alpha}$ with $\frac{7}{6} < \alpha < \frac{5}{4}$ satisfy the lemma. In particular we will be using $\phi(x) = (1-ex)^{\frac{6}{5}}$. □

6.5 Local approximation

We show that all points on the circle $|z| = x$ with argument $|\theta| \leq \phi(x) = (1-ex)^{\frac{6}{5}}$ can be approximated by a small correction over the value on the positive real line.

Theorem 34. Let $\phi(x) = (1-ex)^{\frac{6}{5}}$. Uniformly in θ, for $0 \leq |\theta| \leq \phi(x)$,

$$G(xe^{i\theta}) \sim G(x)\exp\left(i\theta a(x) - \frac{\theta^2}{2}b(x)\right),$$

i.e., for $0 \leq |\theta| \leq \phi(x)$, $\forall\ \epsilon > 0$, $\exists\ \delta(\epsilon)$ such that if $0 < |x - \frac{1}{e}| < \delta$,

$$\left|\frac{G(xe^{i\theta})}{G(x)\exp\left(i\theta a(x) - \frac{\theta^2}{2}b(x)\right)} - 1\right| < \epsilon.$$

Proof Observe that

$$\begin{aligned} G(xe^{i\theta}) &= \exp\left(\sum_{k=1}^{\infty} \frac{k^k x^k e^{ik\theta}}{k!}\right) \\ &= \exp\left(\sum_{k=1}^{\infty} \frac{k^k x^k}{k!} \sum_{l=0}^{\infty} \frac{(ik\theta)^l}{l!}\right) \\ &= \exp\left(\sum_{l=0}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!}\right). \end{aligned}$$

The rearrangement can be done for all $x < \frac{1}{e}$, as the original series is absolutely convergent for $x < \frac{1}{e}$. Split the term in the exponent as,

$$\begin{aligned}\sum_{l=0}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} &= \sum_{k=1}^{\infty} \frac{k^k x^k}{k!} + i\theta \sum_{k=1}^{\infty} \frac{k^{k+1}x^k}{k!} - \frac{\theta^2}{2} \sum_{k=1}^{\infty} \frac{k^{k+2}x^k}{k!} + \sum_{l=3}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} \\ &= \log(G(x)) + i\theta a(x) - \frac{\theta^2}{2} b(x) + \sum_{l=3}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!}.\end{aligned}$$

Observing that if $|t| < \epsilon \le 1$, $|e^t - 1| \le \left|e^{|t|} - 1\right| < (e-1)|t| < (e-1)\epsilon$, an equivalent statement for the local approximation would be that for $|\theta| < \phi(x)$, given $\epsilon > 0$, $\exists\ \delta(\epsilon)$ such that for $|x - \frac{1}{e}| < \delta$,

$$\left| \sum_{l=3}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} \right| < \epsilon.$$

To reduce the above expression note that each term, $\frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!}$, approaches 0 as $x \to \frac{1}{e}$, and that the summation converges by Cauchy's root test *e.g.* [70]. Therefore,

$$\begin{aligned}\left| \sum_{l=3}^{\infty} \frac{(i\theta)^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} \right| &\overset{(a)}{\le} \sum_{l=3}^{\infty} \frac{|\theta|^l}{l!} \sum_{k=1}^{\infty} \frac{k^{k+l}x^k}{k!} \\ &\overset{(b)}{\le} \sum_{l=3}^{\infty} \frac{\phi(x)^l}{l!} \frac{\sqrt{(2l)!}}{2^l (1-ex)^{l+\frac{1}{2}}} \\ &= \sum_{l=3}^{\infty} (1-ex)^{\frac{l}{5}-\frac{1}{2}} \frac{\sqrt{(2l)!}}{2^l l!} \\ &= \left(\sum_{m=0}^{\infty} \frac{(1-ex)^{\frac{m}{5}} \sqrt{(2m+6)!}}{2^{m+3}(m+3)!} \right) (1-ex)^{\frac{1}{10}}. \qquad (6.9)\end{aligned}$$

(a) is the mod-sum inequality and (b) from Lemma 31. Observe that the coefficient of $(1-ex)^{\frac{1}{10}}$ converges when $ex < 1$. To see this, observe that each term is finite when $x \to \frac{1}{e}$, and use Cauchy's root test for the convergence of the series. Since expression (6.9) can be made smaller than ϵ by taking x close enough to $\frac{1}{e}$, the theorem follows. □

6.6 Fast taper

We prove that our choice, $\phi(x) = (1-ex)^{\frac{6}{5}}$ from Lemma 33 is indeed a basic split.

Theorem 35. Let $\phi(x) = (1-ex)^{\frac{6}{5}}$. Uniformly in θ as $x \to \frac{1}{e}$

$$\left|G(xe^{i\theta})\right| \sim \frac{o(G(x))}{\sqrt{b(x)}} \qquad \forall\ \theta : 0 < \phi(x) \le |\theta| < \pi$$

i.e., for $\phi(x) \le |\theta| < \pi$, $\forall\ \epsilon > 0$, $\exists\ \delta(\epsilon)$ such that if $|x - \frac{1}{e}| < \delta$,

$$\left|\frac{G^2(xe^{i\theta})b(x)}{G^2(x)}\right| = \left|\frac{\sum_{k=1}^{\infty} \frac{k^{k+2}}{k!}x^k}{\exp\left(4\sum_{k=1}^{\infty} \frac{k^k}{k!}x^k \sin^2\left(\frac{k\theta}{2}\right)\right)}\right| < \epsilon$$

Proof We first upper bound $b(x)$ using Lemma 31.
We bound the denominator separately in the regions $(1-ex)^{\frac{6}{5}} \le |\theta| \le (1-ex)^{\frac{1}{8}}$ and $(1-ex)^{\frac{1}{8}} \le |\theta| \le \pi$. The bound for the second region will apply uniformly in any range lower bounded by $(1-ex)^{\alpha}$ with $\alpha < \frac{1}{4}$, in particular, we choose $\frac{1}{8}$.

We first consider the second region. Let $\phi'(x) = (1-ex)^{\frac{1}{8}}$. In the sum

$$\sum_{k=1}^{\infty} \frac{k^k}{k!}x^k \sin^2\left(\frac{k\theta}{2}\right),$$

reject all terms for which $\left|\frac{k\theta}{2}\right|$ is less than $\frac{1}{4}\phi'(x)$ or between $\pi \pm \frac{1}{4}\phi'(x)$. The sequence $\left|\frac{1}{2}\theta\right|, |\theta| \ldots \left|\frac{k}{2}\theta\right| \ldots$ will never have 2 consecutive terms $< \frac{1}{4}\phi'(x)$ or between $\pi \pm \frac{1}{4}\phi'(x)$ because $\phi'(x) \le |\theta| \le \pi$. Consequently, for any M consecutive terms, after this rejection process, we will have at least $\lfloor \frac{M}{2} \rfloor$ terms remaining. Lower bounding all remaining $\sin^2\left(\frac{k\theta}{2}\right)$ by $\sin^2\left(\frac{\phi'(x)}{4}\right)$ allows us to factor the $\sin^2(\frac{\phi'(x)}{4})$ term out of the summation. Call the sum of the remaining terms *residual* summation. The terms $\frac{k^k}{k!}x^k$ decrease monotonically with k for $x \le \frac{1}{e}$. So the lower bound for any residual summation is, using Lemma 36,

$$\sin^2\left(\frac{\phi'}{4}\right) \sum_{\substack{k=2 \\ k \text{ even}}}^{\infty} \frac{k^k}{k!}x^k \ge C\frac{\sin^2\left(\frac{\phi'(x)}{4}\right)}{\sqrt{1-ex}}$$

Define $v = \frac{1}{\sqrt[4]{1-ex}}$. Combining all that has been proved so far

$$\left|\frac{\sum_{k=1}^{\infty} \frac{k^{k+2}}{k!}x^k}{\exp\left(4\sum_{k=1}^{\infty} \frac{k^k}{k!}x^k \sin^2\left(\frac{k\theta}{2}\right)\right)}\right| \le \frac{Cv^{10}}{e^{cv}}$$

which can be made smaller than any $\epsilon > 0$, for all $|\theta| \ge (1-ex)^{\frac{1}{8}}$ by choosing $|x - \frac{1}{e}| < \delta_1(\epsilon)$. Note that the $\frac{\sin^2(\phi'(x))}{\sqrt[4]{1-ex}} \to \infty$ for $\beta < \frac{1}{8}$, and equals 1 for $\beta = \frac{1}{8}$.

To tackle the remaining region, *i.e.,* $(1-ex)^{\frac{6}{5}} \leq |\theta| \leq (1-ex)^{\frac{1}{8}}$, we use the following inequality for $\frac{k\theta}{2} \leq \frac{\pi}{2}$,

$$\left|\sin\left(\frac{k\theta}{2}\right)\right| \geq \frac{2}{\pi}\frac{k\theta}{2}.$$

In this region, we will have $\frac{\pi}{\theta}$ terms for which the inequality holds with both sides being positive.

We write $\theta = (1-ex)^{\alpha}$. Therefore $\frac{1}{8} \leq \alpha \leq \frac{6}{5}$. Squaring and substituting the above inequality into the left side of the Theorem,

$$\left|\frac{\sum_{k=1}^{\infty}\frac{k^{k+2}}{k!}x^k}{\exp\left(4\sum_{k=1}^{\infty}\frac{k^k}{k!}x^k\sin^2\left(\frac{k\theta}{2}\right)\right)}\right| \leq \left|\frac{\sum_{k=1}^{\infty}\frac{k^{k+2}}{k!}x^k}{\exp\left(\frac{16}{\pi^2}\sum_{k=1}^{\frac{\pi}{\theta}}\frac{k^k}{k!}x^k\frac{k^2\theta^2}{4}\right)}\right|.$$

We lower bound $(1-ex)^{2\alpha}\sum_{k=1}^{\frac{\pi}{(1-ex)^{\alpha}}}\frac{k^{k+2}x^k}{k!}$ using Lemma 37,

$$(1-ex)^{2\alpha}\sum_{k=1}^{\frac{\pi}{(1-ex)^{\alpha}}}\frac{k^{k+2}x^k}{k!} \geq \begin{cases}\frac{C}{(1-ex)^{\frac{\alpha}{2}}} & \frac{1}{8} \leq \alpha \leq 1\\ \frac{C}{(1-ex)^{\frac{5}{2}-2\alpha}} & 1 \leq \alpha \leq \frac{6}{5}.\end{cases}$$

Define $v = \frac{1}{\sqrt[16]{1-ex}}$. We conclude

$$\left|\frac{\sum_{k=1}^{\infty}\frac{k^{k+2}}{k!}x^k}{\exp\left(4\sum_{k=1}^{\infty}\frac{k^k}{k!}x^k\sin^2\left(\frac{k\theta}{2}\right)\right)}\right| \leq \frac{Cv^{40}}{e^{cv}}$$

which, for $\phi(x) \leq |\theta| \leq \sqrt[8]{1-ex}$, can be made smaller than $\epsilon > 0$, by taking $|x - \frac{1}{e}| \leq \delta_2(\epsilon)$. Picking $\delta = \min(\delta_1, \delta_2) = \delta_2$ concludes the proof for all $(1-ex)^{\frac{6}{5}} \leq \phi(x) \leq \pi$. □

We prove Lemma 36 and Lemma 37 used in Theorem 35.

Lemma 36. For $\frac{5}{6e} < x < \frac{1}{e}$,

$$\sum_{\substack{k=2\\ k\text{ even}}}^{\infty}\frac{k^k}{k!}x^k \geq \frac{C}{\sqrt{1-ex}}.$$

Proof From Feller's bounds (2.3),

$$\sum_{\substack{k=2\\ k\text{ even}}}^{\infty}\frac{k^k}{k!}x^k \geq \frac{e^{-\frac{1}{12}}}{\sqrt{2\pi}}\sum_{\substack{k=2\\ k\text{ even}}}^{\infty}\frac{1}{\sqrt{k}}(ex)^k.$$

To lower bound the sum on the right observe that,

$$\begin{aligned}\left(\sum_{\substack{k=2\\k \text{ even}}}^{\infty} \frac{(ex)^k}{\sqrt{k}}\right)^2 &= \sum_{\substack{k=4\\k \text{ even}}}^{\infty} (ex)^k \sum_{\substack{l=2\\l \text{ even}}}^{k-2} \frac{1}{\sqrt{l(k-l)}}\\ &\geq \sum_{\substack{k=4\\k \text{ even}}}^{\infty} (ex)^k \frac{k}{4}\cdot\frac{2}{k}\\ &= \frac{(ex)^4}{2(1-ex)(1+ex)}\\ &\geq \frac{C}{1-ex}.\end{aligned}$$

By observing that $\frac{5}{6} < ex < 1$ we incorporate ex and $1+ex$ into a constant. Taking the positive square root proves the lemma. □

Lemma 37. For $x > \frac{5}{6e}$,

$$(1-ex)^{2\alpha} \sum_{k=1}^{\frac{2}{(1-ex)^{\alpha}}} \frac{k^{k+2}x^k}{k!} \geq \begin{cases} \frac{C}{(1-ex)^{\frac{\alpha}{2}}} & \alpha \leq 1\\ \frac{C}{(1-ex)^{\frac{5}{2}-2\alpha}} & \alpha \geq 1.\end{cases}$$

Proof For any m, from Feller Bounds (2.3),

$$\sum_{k=1}^{m} \frac{k^{k+2}x^k}{k!} \geq C_1 \sum_{k=1}^{m} \frac{k^2x^ke^k}{\sqrt{k}}.$$

We first show that if $k < \frac{3ex}{2(1-ex)}$, the k'th term is less than the $k+1'$th term in the above summation.

To see that observe that the ratio of the $k+1'$th to the kth term is

$$\left(1+\frac{1}{k}\right)^{3/2} xe$$

and that

$$\left(1+\frac{1}{k}\right)^{3/2} xe \geq \left(1+\frac{3}{2k}\right) xe \geq 1.$$

where the second inequality holds if $k < \frac{3ex}{2(1-ex)}$. Since $ex > \frac{5}{6}$, the terms before $k = \frac{5}{4(1-ex)}$ in the summation are nondecreasing.

For $\alpha \leq 1$, observe that

$$\frac{1}{(1-ex)^{\alpha}} \leq \frac{1}{(1-ex)},$$

so that

$$\begin{aligned}\sum_{k=1}^{\frac{2}{(1-ex)^{\alpha}}} k^{3/2}x^k e^k &\geq \sum_{\frac{1}{(1-ex)^{\alpha}}}^{\frac{5}{4(1-ex)^{\alpha}}} k^{3/2}x^k e^k \\ &\overset{(a)}{\geq} \frac{1}{(1-ex)^{\frac{3}{2}\alpha}} \cdot (xe)^{\frac{1}{(1-ex)^{\alpha}}} \cdot \frac{1}{4(1-ex)^{\alpha}} \\ &\overset{(b)}{\geq} \frac{1}{4(1-ex)^{\frac{5}{2}\alpha}} \cdot \frac{1}{4},\end{aligned}$$

where (a) follows by replacing all terms of the summation with the first term in the summation and (b) because for all $\frac{1}{2} \leq y < 1$,

$$y^{\frac{1}{(1-y)^{\alpha}}} \geq y^{\frac{1}{1-y}} \geq \left(\frac{1}{2}\right)^2.$$

We complete the proof for $\alpha > 1$ by using the lemma for $\alpha = 1$, which we just proved. For $\alpha > 1$, observe that

$$\frac{1}{(1-ex)^{\alpha}} > \frac{1}{(1-ex)}.$$

Using the inequality,

$$(1-ex)^2 \sum_{k=1}^{\frac{1}{(1-ex)}} \frac{k^{k+2}x^k}{k!} \geq \frac{C}{(1-ex)^{\frac{1}{2}}},$$

observe that

$$\begin{aligned}(1-ex)^{2\alpha} \sum_{k=1}^{\frac{2}{(1-ex)^{\alpha}}} \frac{k^{k+2}x^k}{k!} &\geq (1-ex)^{2\alpha} \sum_{k=1}^{\frac{2}{(1-ex)}} \frac{k^{k+2}x^k}{k!} \\ &\geq \frac{C}{(1-ex)^{\frac{5}{2}-2\alpha}}.\end{aligned}$$

□

6.7 Evaluation of coefficients

Using Hayman's analysis, we evaluate the lower bound on $\hat{R}(\mathcal{I}_{\Psi}^n)$, namely $\frac{n!}{n^n}$ times the n'th coefficient of the expansion of $G(z)$.

Theorem 38.

$$\hat{R}^-(\mathcal{I}^n_\Psi) = \left(\frac{3}{2}\log e\right)n^{\frac{1}{3}} - \frac{1}{3}\log n - \frac{2}{3}\log e - \frac{1}{2}\log 3 + o(1).$$

Proof From (6.3), (6.4) and (6.5), we have that

$$\sum_{n=0}^{\infty} 2^{\hat{R}^-(\mathcal{I}^n_\Psi)}\frac{n^n}{n!}z^n = G(z) = \exp\left(\sum_{k=1}^{\infty}\frac{k^k z^k}{k!}\right).$$

From the observations following (6.8) and Theorems 34 and 35, we conclude that $G(z)$ satisfies the conditions of Theorem 29.

To use (6.6), we need to evaluate the function $a(z)$ shown in (6.8) to be

$$\sum_{k=1}^{\infty}\frac{k^{k+1}z^k}{k!}.$$

We do so using the related “tree function” [67]

$$T(z) = \sum_{k=1}^{\infty}\frac{k^{k-1}z^k}{k!},$$

which satisfies [67] the equation

$$T(z) = ze^{T(z)}. \tag{6.10}$$

Therefore,

$$\sum_{k=1}^{\infty}\frac{k^k z^k}{k!} = \frac{T(z)}{1-T(z)}. \tag{6.11}$$

By differentiating Equations (6.10) and (6.11) and using the absolute convergence of the series, we obtain

$$a(z) = \frac{T(z)}{(1-T(z))^3}, \text{ and } b(z) = \frac{2T(z)^2 + T(z)}{(1-T(z))^5}.$$

At $z = \frac{1}{e}$, we have the following singular expansion [67],

$$\frac{1}{1-T(z)} = \frac{1}{\sqrt{2(1-ez)}} + \frac{1}{3} - \frac{\sqrt{2}}{24}\sqrt{1-ez} + \mathcal{O}(1-ez).$$

Consequently, it can be verified that

$$a(z) = \frac{1}{(2(1-ez))^{\frac{3}{2}}} + \mathcal{O}\left(\frac{1}{\sqrt{1-ez}}\right),$$

and the solution to $a(r_n) = n$ is

$$r_n = \frac{1}{e}\left(1 - \frac{1}{2n^{\frac{2}{3}}}\right) + \mathcal{O}\left(\frac{1}{n^{\frac{4}{3}}}\right).$$

The n'th coefficient of the $G(z)$ therefore equals

$$\frac{G(r_n)}{r_n^n\sqrt{2\pi b(r_n)}}(1 + o(1)). \tag{6.12}$$

We evaluate the terms to be

$$\begin{aligned} G(r_n) &= \exp(n^{\frac{1}{3}} - \frac{2}{3} + \mathcal{O}(n^{-\frac{1}{3}})), \\ r_n^n &= \exp(-n - \frac{1}{2}n^{\frac{1}{3}} + \mathcal{O}(n^{-\frac{1}{3}}))(1 + \mathcal{O}(n^{-\frac{2}{3}})), \text{ and} \\ b(r_n) &= 3n^{\frac{5}{3}} + \mathcal{O}(n), \end{aligned}$$

and use them to evaluate $\hat{R}^-(\mathcal{I}_\Psi^n)$,

$$\hat{R}^-(\mathcal{I}_\Psi^n) = \left(\frac{3}{2}\log e\right)n^{\frac{1}{3}} - \frac{1}{3}\log n - \frac{2}{3}\log e - \frac{1}{2}\log 3 + o(1). \qquad \square$$

We note that this is the highest accuracy of the asymptotic expansion allowed by the Hayman's theorem, limited by the form of the Equation (6.12).

Corollary 39.

$$\hat{R}(\mathcal{I}_\Psi^n) \geq \left(\frac{3}{2}\log e\right)n^{\frac{1}{3}} - \frac{1}{3}\log n - \frac{2}{3}\log e - \frac{1}{2}\log 3 + o(1). \qquad \square$$

Acknowledgements

We thank Wojciech Szpankowski for sharing his intuition behind saddle point approximations. This chapter is adapted from N. Jevtić, A. Orlitsky, and N.P. Santhanam. A lower bound on compression of unknown alphabets. Theoretical Computer Science, Feb 2005. The author was a primary researcher of the sections reproduced here.

Chapter 7

Shapes, patterns and strings: summary of properties

Throughout the chapters so far, we related sequences, shapes, patterns, types and profiles to partitions of sets and integers. In this section we summarize these relations, and use this unified framework to explain several known combinatorial identities relating the Bell numbers, the Fubini numbers, the Stirling numbers of the second kind, and two additional quantities, the number of length-n, k-ary sequences where exactly m of the k symbols appear, and the number of length-n shapes with m symbols.

Table 7.1 describes the various quantities defined in the report and some of the upper bounds obtained. It consists of three columns, corresponding to sequences, shapes, and patterns, which we collectively call *strings*. It also comprises four main horizontal sections describing the strings, their classification into types, the classification of types into profiles, and the resulting bounds on the redundancy.

We now describe the various table entries in row order. In both the table and the description below, all strings are of length n. In addition, sequences are over the initial segment $[k] = \{1, \ldots, k\}$.

Every sequence can be written as $x_1 \ldots x_n$ where $x_i \in [k]$. A shape is a string $\mathsf{s}_1 \ldots \mathsf{s}_n$ where $\{\mathsf{s}_1, \ldots, \mathsf{s}_n\}$ is an initial segment, namely, $\{\mathsf{s}_1, \ldots, \mathsf{s}_n\} = [\max\{\mathsf{s}_1, \ldots, \mathsf{s}_n\}]$. A pattern is a string $\psi_1 \ldots \psi_n$ where $\{\psi_1, \ldots, \psi_i\}$ is an initial segment for every $1 \leq i \leq n$. Consider for example the sequence 51535 over, say, [6]. Its shape is 31323 and its pattern is 12131. Note that $\{3,1,3,2,3\} = [3]$, and that $\{1\} = [1]$, $\{1,2\} = [2]$, $\{1,2,1\} = [2]$, etc.

The total number of sequences is k^n. Since every shape corresponds to an ordered partition of $[n]$, the number of shapes is the nth Fubini number $F(n)$. Since every pattern corresponds to an unordered partition of $[n]$, the number of patterns is the nth Bell number $B(n)$.

We denote the number of sequences with exactly m symbols by $N(n,m,k)$. For example, $N(n,1,k) = |\{1\ldots1,\ldots,k\ldots k\}| = k$, $N(n,2,k) = \binom{k}{2}\cdot(2^n-2)$, and $N(n,n,k) = k^{\underline{n}}$, where $k^{\underline{n}} = k\cdot(k-1)\cdot(k-n+1)$ is the nth *falling power* of k. Note that when $m > \min\{n,k\}$, $N(n,m,k) = 0$. We denote the number of shapes with exactly m symbols by $F(n,m)$. For example, $F(n,1) = |\{1\ldots1\}| = 1$, $F(n,2) = 2^n - 2$, and $F(n,n) = n!$. The number of patterns with exactly m symbols is the second-type Stirling number $B(n,m)$, the number of unordered partitions of $[n]$ into m parts. For example, $B(n,1) = |\{1\ldots1\}| = 1$ and similarly $B(n,2) = 2^{n-1}-1$, and $B(n,n) = 1$.

The *multiplicity* μ of a symbol is the number of times it appears in a string. We classify strings by their *type*—the multiplicity of all symbols in the string. The type of a sequence is therefore a k-tuple $(\mu_1,\ldots,\mu_k)$ where the multiplicities μ_i are non-negative integers summing to n. The type of a shape and a pattern is an m-tuple $(\mu_1,\ldots,\mu_m)$ where m is the number symbols appearing and the multiplicities μ_i are positive integers summing to n. For example, the type of 51535 is $(1,0,1,0,3)$ as 1 and 3 appear once and 5 appears thrice. Similarly, the type of the shape 31323 is $(1,1,3)$ and that of the pattern 12131 is $(3,1,1)$.

The type of a sequence corresponds to an ordered partition of n into k non-negative parts, hence sequences fall into $\binom{n+k-1}{k-1}$ types. The type of a shape or a pattern corresponds to an ordered partition of n hence shapes and patterns fall into 2^{n-1} types.

The type of a sequence with m symbols can be described by specifying these symbols and an ordered partition of n into m parts. Hence there are $\binom{k}{m}\binom{n-1}{m-1}$ such types. The type of a shape or a pattern with m symbols is an ordered partition of n into m parts, hence there are $\binom{n-1}{m-1}$ such types.

A sequence has type $(\mu_1,\ldots,\mu_k)$ if for all $1 \le i \le k$, symbol i appears μ_i times. The number of such sequences is $\binom{n}{\mu_1,\ldots,\mu_k}$. Similarly, the number of shapes of type

$(\mu_1,\ldots,\mu_m)$ is $\binom{n}{\mu_1,\ldots,\mu_m}$. Every m-symbol profile corresponds to $m!$ shapes, hence the number of profiles of type $(\mu_1,\ldots,\mu_m)$ is $\binom{n}{\mu_1,\ldots,\mu_m}/m!$.

Next we describe patterns. The *prevalence* φ_μ is the number of symbols with multiplicity μ appearing in a string. We classify strings by their *profile*—the prevalence of all possible multiplicities. The profile of any string is therefore an n-tuple $(\varphi_n,\ldots,\varphi_1)$ where all the prevalences φ_μ are non-negative and $\sum_\mu \mu\varphi_\mu = n$. For example, the profile of the sequence 51535 is $(0,0,1,0,2)$ as two symbols (1 and 3) appear once and one symbol (5) appears thrice. It is easy to see that $(0,0,1,0,2)$ is also the profile of the sequence's shape, 31323, and of its pattern 12131.

The profile of a string can be identified with an unordered partition of n, hence the number of profiles of strings is $p(n)$, the number of unordered partitions of n.

Similarly, the profile of a string with m symbols is an unordered partition of n into m parts, hence the number of such profiles is $p_m(n)$, the number of unordered partitions of n into m parts.

As shown in, *e.g.* [2], the number of patterns with profile $(\varphi_n,\ldots,\varphi_1)$ is

$$\frac{n!}{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!}.$$

The number of symbols in the string is $\sum_\mu \varphi_\mu \overset{\text{def}}{=} m$. Since every pattern corresponds to $m!$ shapes and to $k^{\underline{m}}$ sequences, there are

$$m!\cdot\frac{n!}{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!}$$

shapes, and

$$k^{\underline{m}}\cdot\frac{n!}{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!}$$

sequences with profile $(\varphi_n,\ldots,\varphi_1)$.

This framework of viewing the Bell, Fubini, and Stirling numbers as the number of sequences, shapes, and patterns, can be used to describe several combinatorial identities relating them [62, 71]. Every sequence can be specified by its shape and the set of symbols appearing in it. The shape, in turn, can be specified by its pattern and a permutation reflecting the order in which the symbols appear. Hence

$$N(n,m,k) = \binom{k}{m}F(n,m) = k^{\underline{m}}B(n,m). \tag{7.1}$$

For all $n, m \geq 1$ the number of shapes with a given number of symbols satisfies the recursion

$$F(n, m) = mF(n-1, m) + mF(n-1, m-1).$$

Equation (7.1) therefore implies for all $n, m \geq 1$, the analogous recursions for patterns

$$B(n, m) = mB(n-1, m) + B(n-1, m-1)$$

and sequences

$$N(n, m, k) = mN(n-1, m, k) + (k-m+1)N(n-1, m-1, k).$$

While there are no closed-form expressions for the Fubini or Bell numbers, or for the number of strings with a given number of symbols, it is possible to express them as sums. Consider the number of shapes with m symbols. A m-symbol shape is a sequence over $[m]$ containing all integers from 1 to m. For $1 \leq j \leq m$, let $A_j \stackrel{\text{def}}{=} [m] - \{j\}$ be the set of all integers from 1 to m, excluding j. Then A_j^n is the collection of length-n sequences over $[m]$ that do not include j. Therefore

$$\mathsf{S}_m^n = [m]^n - \cup_{j=1}^m A_j^n,$$

hence

$$F(n, m) = m^n - \left|\cup_{j=1}^m A_j^n\right|.$$

Let $\binom{[m]}{l}$ be the collection of all l-sized subsets of $[m]$. For all $J_l \in \binom{[m]}{l}$

$$\left|\bigcap_{j \in J_l} A_j^n\right| = \left|\bigcap_{j \in J_l} A_j\right|^n = (m-l)^n$$

where the first equality holds because for all sets $B_1, \ldots, B_l$

$$\bigcap_{i=1}^l B_i^n = \left(\bigcap_{i=1}^l B_i\right)^n$$

and the second equality holds as

$$\left|\bigcap_{j \in J_l} A_j\right| = \left|[m] - J_l\right| = m - l.$$

By the inclusion/exclusion principle

$$\begin{aligned}\left|\cup_{j=1}^m A_j^n\right| &= \sum_{l=1}^m (-1)^{l-1} \sum_{J_l \in \binom{[m]}{l}} \left|\bigcap_{j\in J_l} A_j^n\right| \\ &= \sum_{l=1}^m (-1)^{l-1}\binom{m}{l}(m-l)^n.\end{aligned}$$

Hence for all $n, m \geq 1$

$$\begin{aligned}F(n,m) &= m^n - \sum_{l=1}^m (-1)^{l-1}\binom{m}{l}(m-l)^n \\ &= (-1)^m \sum_{l=0}^m (-1)^l \binom{m}{l} l^n.\end{aligned}$$

For example, letting $m = n$, we obtain

$$(-1)^n \sum_{i=0}^n (-1)^l \binom{n}{l} l^n = F(n,n) = n!$$

and letting $m > n$, we obtain

$$\sum_{i=0}^m (-1)^l \binom{n}{l} l^n = 0.$$

Equation (7.1) then implies that for patterns

$$B(n,m) = \frac{(-1)^m}{m!} \sum_{l=0}^m (-1)^l \binom{m}{l} l^n \tag{7.2}$$

and that for sequences over $[k]$

$$N(n,m,k) = (-1)^m \binom{k}{m} \sum_{l=0}^m (-1)^l \binom{m}{l} l^n.$$

Expressing the nth Bell number as the sum of all $B(n,m)$ and incorporating (7.2), we obtain

$$\begin{aligned}B(n) = \sum_{m=1}^\infty B(n,m) &= \sum_{m=1}^\infty (-1)^m \sum_{l=1}^m \frac{(-1)^l}{l!(m-l)!} l^n \\ &= \sum_{l=1}^\infty \frac{(-1)^l l^n}{l!} \sum_{m=l}^\infty \frac{(-1)^m}{(m-l)!} \\ &= \frac{1}{e}\sum_{l=1}^\infty \frac{l^n}{l!}.\end{aligned}$$

Using other techniques, a corresponding formula for shapes can be obtained:

$$F(n) = \sum_{l=1}^{\infty} \frac{l^n}{2^{l+1}}.$$

Acknowlegdements

We thank N. Alon, N. Jevtić, N. Kashyap, G. Shamir, W. Szpankowski, K. Viswanathan, and J. Zhang for helpful discussions and comments. This chapter is adapted from A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004. The author was a primary researcher of the paper.

Table 7.1: Summary of terms and upper bounds

		Sequences over $[k]$	Shapes	Patterns
S t r	Description	$x_1 \dots x_n$: $x_i \in [k]$	$\mathsf{s}_1 \dots \mathsf{s}_n$: $\{\mathsf{s}_1, .., \mathsf{s}_n\} =$ $[\max\{\mathsf{s}_1, .., \mathsf{s}_n\}]$	$\psi_1 \dots \psi_n$: For all $j \le n$ $\{\psi_1, .., \psi_j\} =$ $[\max\{\psi_1, .., \psi_j\}]$
	Example	51535	31323	12131
	#, All	k^n	$F(n)$	$B(n)$
	#, $m-$symbols	$N(n,m,k)$	$F(n,m)$	$B(n,m)$
T y p e s	Description	$(\mu_1, \dots, \mu_k)$: $\mu_i \ge 0$, $\sum_i \mu_i = n$	$(\mu_1, \dots, \mu_m)$: $1 \le m \le n$, $\mu_i \ge 1$, $\sum_i \mu_i = n$	
	Example	(1,0,1,0,3)	(1,1,3)	(3,1,1)
	All	$\binom{n+k-1}{k-1}$	2^{n-1}	
	$m-$symbols	$\binom{k}{m}\binom{n-1}{m-1}$	$\binom{n-1}{m-1}$	
	# strings/type	$\binom{n}{\mu_1,\dots,\mu_k}$	$\binom{n}{\mu_1,\dots,\mu_m}$	$\frac{1}{m!}\binom{n}{\mu_1,\dots,\mu_m}$
φ	Description	$(\varphi_n, \dots, \varphi_1)$, $\varphi_\mu \ge 0$, $\sum_\mu \mu\varphi_\mu = n$		
	Example	(0,0,1,0,2)		
	All	$p(n)$		
	$m-$symbols	$p_m(n)$		
	#strings/profile	$k^{\sum_\mu \varphi_\mu} \cdot \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!}$	$\left(\sum_\mu \varphi_\mu\right)! \cdot \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!}$	$\frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!}$
R d n	Equiprobable strings	Type	Type	Profile
	All Upper bound	$\log\binom{n+k-1}{k-1}$	$n-1$	$\log p(n)$
	m-symbols	$\log\left(\binom{k}{m}\binom{n-1}{m-1}\right)$	$\log\binom{n-1}{m-1}$	$\log p_m(n)$
	maximizing m	$\lceil \frac{nk}{n+k+1} \rceil$	$\lceil \frac{n}{2} \rceil$ and $\lfloor \frac{n}{2} \rfloor$	$\frac{\sqrt{6}}{\pi}\sqrt{n}\log n \cdot$ $(1+o(1))$

Chapter 8

Good Turing estimators

The focus of this chapter will be large alphabet distribution estimation.

In the large alphabet setting studied here, estimators assign probabilities to the events that the next element is one of the elements that has appeared before, and assign a probability that the next element is hitherto unseen. This represents a simplification from estimating the distribution over the whole support at each step, and we will see that this approach is analogous to doing sequential pattern compression.

We define the Good-Turing estimators and show that some common variants perform well for large alphabet distribution estimation problems. However, they are not diminishing attenuation estimators.

We derive two diminishing-attenuation estimators. The first is computationally more efficient and requires only a constant number of operations per symbol. Its sequence attenuation is at most $2^{\mathcal{O}(n^{2/3})}$, hence its symbol attenuation converges to 1 as $2^{\mathcal{O}(n^{-1/3})}$. The second estimator requires a super-polynomial number of calculations, however its sequence attenuation is lower, at most $2^{\mathcal{O}(n^{1/2})}$, hence its symbol attenuation converges to 1 at the faster rate of $2^{\mathcal{O}(n^{-1/2})}$.

All constants involved in the asymptotic terms are small. The proofs of the attenuations of the two estimators are rather different. The proof for the low complexity estimator uses potential functions, while the proof for the higher complexity estimator uses results on set partitions and celebrated results of Hardy and Ramanujan [50] on the number of partitions of an integer.

To better understand the behavior of the estimator, we study the probability it assigns to some simple sequences, and show that while it often behaves as our intuition would indicate, sometimes its estimates are surprising. For example, as we would intuitively guess, after observing a long sequence of identical symbols, the estimator predicts that the next symbol will be the same too, and after seeing a long sequence whose symbols are all different, it predicts that the next symbol will be new too. However if every symbol in the sequence appears twice, then our intuition would say that since roughly every other symbol is new, the probability of the next symbol being new is half. Yet the probability that the estimator assigns to a new symbol is lower.

Estimators for the Good-Turing problem

An estimator associates with every sequence of observations a probability distribution over the set of elements in the sample, and "new". For example, after observing the sample

giraffe, hippopotamus, giraffe, elephant, elephant, giraffe,

an estimator postulates a distribution over the set {giraffe, hippopotamus, elephant, "new"}, reflecting the probability that a randomly chosen element is any one of these animals, or new.

Note that the estimator is not required to distinguish between unseen elements. As mentioned in the introduction, this simplification is equivalent to universally compressing patterns of strings instead of the strings themselves.

We assume no a priori knowledge on the elements in the sample, a giraffe is no different to us from an elephant, hence we replace the name of each animal by the order in which it appears. For example, in the sequence above, we denote giraffes by 1, hippopotami by 2, and elephants by 3. The sequence of animals then turns into the integer sequence $1, 2, 1, 3, 3, 1$, which we often abbreviate as 121331. Recall that this is just the pattern of the original sequence.

This representation abstracts the names of the elements, always referring to the numbers $1, 2, \ldots, k$, thereby allowing us to enumerate, and hence assign probabilities, to sequences of arbitrary elements. Note that a "new" element is represented by a number one more than the number of elements hitherto seen.

The estimator is sequential, namely for every pattern ψ_1^n, $n \in \mathbb{N}$, it corresponds to a probability distribution $q(x|\psi_1^n)$ over $[m(\psi_1^n)+1] = \{1,\ldots,m(\psi_1^n)+1\}$, representing the probability that the estimator assigns to the possible values of ψ_{n+1}, after seeing ψ_1^n. For example, $q(x) \overset{\text{def}}{=} q(x|\Lambda)$ is a distribution over $\{1\}$, namely, $q(1|\Lambda) = 1$, while $q(x|121)$ is a distribution over $\{1,2,3\}$.

For a simple example, consider the add-one estimator alluded to in Chapter 1, and henceforth denoted q_{+1}. After observing the pattern ψ_1^n it assigns to any $x \in [m(\psi_1^n)+1]$ a probability proportional to one more than the number of times it appeared in ψ_1^n. For example, after observing the pattern 1, it estimates $q_{+1}(1|1) = (1+1)/3 = 2/3$ and $q_{+1}(2|1) = (0+1)/3 = 1/3$.

For each $n \in \mathbb{Z}^+$, an estimator q induces a probability distribution over Ψ^n given by

$$q(\psi_1^n) \overset{\text{def}}{=} \prod_{i=1}^{n} q(\psi_i|\psi_1^{i-1}).$$

For example, the probability that the add-one estimator ascribes to the pattern 1213 is

$$q_{+1}(1213) = q_{+1}(1|\Lambda)\cdot q_{+1}(2|1)\cdot q_{+1}(1|12)\cdot q_{+1}(3|121) = \frac{1}{1}\cdot\frac{1}{3}\cdot\frac{2}{5}\cdot\frac{1}{6} = \frac{1}{45}.$$

8.1 Unbounded- and constant-attenuation estimators

We show that the add-one estimator has unbounded attenuation and that the Good-Turing and a modified version of the add-one estimator have constant attenuations, though these constants are larger than 1.

8.1.1 The add-one estimator and a variation

It is easy to see that add-constant estimators have unbounded attenuation. Consider for example the add-one estimator. To the pattern $123\ldots n$ it assigns probability

$$\frac{1}{1}\cdot\frac{1}{3}\cdot\ldots\cdot\frac{1}{2n+1} = \frac{2^n\cdot n!}{(2n+1)!}.$$

Since, as we saw in the introduction, $\hat{p}(12\ldots n) = 1$, we obtain that q_{+1} has symbol attenuation of roughly $2n/e$, hence its attenuation is unbounded.

By applying the add-one estimator in two steps, we obtain an estimator $q_{+1'}$ with attenuation of between 2.65 and 2.85. The estimator $q_{+1'}$ uses the add-1 rule to estimate the probability of the next symbol being new or repeated. For repeated symbols, $q_{+1'}$ assigns a probability proportional to the number of occurrences of the symbol. Formally, given ψ_1^n, let m be the number of distinct symbols appearing in ψ_1^n and for $1 \le \psi \le m$ let μ_ψ be the number of times the symbol ψ appeared in ψ_1^n. Then $q_{+1'}$ assigns to each $1 \le \psi \le m+1$ the probability

$$q_{+1'}(\psi|\psi_1^n) \stackrel{\text{def}}{=} \begin{cases} \frac{m+1}{n+2} & \psi = m+1 \\ \frac{n-m+1}{n+2} \cdot \frac{\mu_\psi}{n} & 1 \le \psi \le m. \end{cases}$$

The following can be proved.

Theorem 40.

$$2.65 \le \hat{A}^*(q_{+1'}) \le 2.85. \qquad \square$$

It can be shown however that for sequences with $m = o(n)$, the estimator $q_{+1'}$ has subexponential sequence attenuation, hence diminishing symbol attenuation.

8.1.2 The Good-Turing estimator

We show that the attenuation of the Good-Turing estimator is a constant between 1.39 and 2. First we need a few definitions.

The *multiplicity* of $\psi \in \mathbb{Z}^+$ in ψ_1^n is

$$\mu_\psi \stackrel{\text{def}}{=} \mu_\psi(\psi_1^n) \stackrel{\text{def}}{=} |\{1 \le i \le n : \psi_i = \psi\}|,$$

the number of times ψ appears in ψ_1^n. The *prevalence* of the multiplicity $\mu \in \mathbb{N}$ in $\overline{\psi}$ is

$$\varphi_\mu \stackrel{\text{def}}{=} \varphi_\mu(\overline{\psi}) \stackrel{\text{def}}{=} |\{\psi : \mu_\psi = \mu\}|,$$

the number of symbols appearing μ times in ψ_1^n. Given ψ_1^{n+1}, let

$$r \stackrel{\text{def}}{=} \mu_{\psi_{n+1}}(\psi_1^n).$$

The Good Turing estimator [40] is then

$$q(\psi_{n+1}|\psi_1^n) = \begin{cases} \frac{\varphi'_1}{n}, & r = 0 \\ \frac{r+1}{n}\frac{\varphi'_{r+1}}{\varphi'_r}, & r \ge 1. \end{cases}$$

where φ'_μ is a smoothed value of φ_μ. Smoothing is needed for a variety of reasons. One of them is that if $\varphi_\mu(\psi_1^n) = 0$ for some $\mu > 0$, then, without smoothing the estimator would assign $q(\psi_{n+1}|\psi_1^n) = 0$ for the symbols appearing $\mu - 1$ times in ψ_1^n. Many smoothing methods have been proposed, some seem too difficult to analyze. All those we analyzed yield attenuation > 1 and all will result with a constant > 1. Here we consider only one of the simplest smoothing techniques

$$\varphi'_\mu = \max(\varphi_\mu, 1)$$

which ensures nonzero probabilities for all symbols in $[1, m(\psi_1^n) + 1]$. This smoothing method results in the estimator

$$q_{\text{GT1}}(\psi_{n+1}|\psi_1^n) \stackrel{\text{def}}{=} \begin{cases} \frac{\max(\varphi_1,1)}{S_{\text{GT1}}(\psi_1^n)}, & r = 0 \\ \frac{r+1}{S_{\text{GT1}}(\psi_1^n)} \frac{\max(\varphi_{r+1},1)}{\varphi_r}, & r \geq 1, \end{cases}$$

where

$$S_{\text{GT1}}(\psi_1^n) \stackrel{\text{def}}{=} \max(\varphi_1, 1) + \sum_{\mu:\varphi_\mu > 0} \varphi_\mu \cdot (\mu + 1) \frac{\max(\varphi_{\mu+1}, 1)}{\varphi_\mu} = \sum_{\mu=0}^{n} (\mu + 1) \max(\varphi_{\mu+1}, 1)$$

is a normalization factor. The attenuation of q_{GT1} can be bounded as follows.

Theorem 41.

$$1.39 \leq \hat{A}^*(q_{\text{GT1}}) \leq 2.$$

Proof outline The lower bound is proved by considering the pattern

$$1, 2, (1, 3, 2,)^{n/3} \stackrel{\text{def}}{=} 1, 2, 1, 3, 2, 1, 3, 2, \ldots, 1, 3, 2.$$

The estimator q_{GT1} assigns to this sequence probability of $\Theta(72^{-n/3})$ while its maximum likelihood probability is $\Theta(3^{-n/3})$. This bound can be improved using more complex patterns.

To prove the upper bound, let $r(i) \stackrel{\text{def}}{=} \mu_{\psi_{i+1}}(\psi_1^i)$, and $\varphi_\mu^i \stackrel{\text{def}}{=} \varphi_\mu(\psi_1^i)$. It can be shown by induction that

$$q_{\text{GT1}}(\psi_1^n) = \frac{\prod_{\mu=1}^n (\mu!)^{\varphi_\mu}}{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)} \cdot \prod_{i=1}^{n-1} \frac{\max(\varphi_{r(i)+1}^i, 1)}{\varphi_{r(i)}^i}.$$

This implies,

$$\hat{A}^n(q_{\text{GT1}}) \leq \left(\max_{\psi_1^n \in \Psi^n} \frac{\prod_{\mu=1}^n \varphi_\mu^n!}{\prod_{i=1}^{n-1} \max(\varphi_{r(i)+1}^i, 1)/\varphi_{r(i)}^i} \right) \cdot \left(\max_{\psi_1^n \in \Psi^n} \frac{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)}{n!} \right) \stackrel{\text{def}}{=} \hat{A}_G^n \cdot \hat{A}_S^n.$$

To prove the theorem, we bound each of $\hat{A}_G^n$ and $\hat{A}_S^n$ individually. □

8.2 Diminishing-attenuation estimators

We describe two diminishing-attenuation estimators. The first is computationally efficient and uses just a constant number of operations per symbol, hence has linear complexity for the whole sequence. The second has super polynomial, though subexponential, complexity, but its attenuation diminishes to 1 faster.

8.2.1 A low complexity estimator

For $c \in \mathbb{Z}^+$, let

$$f_c(\varphi) \overset{\text{def}}{=} \max(\varphi, c)$$

and let

$$g_c(\varphi) \overset{\text{def}}{=} \prod_{i=1}^{\varphi} f_c(i) = \begin{cases} c^{\varphi}, & 0 \le \varphi \le c-1 \\ \frac{c^c}{c!}\varphi!, & \varphi \ge c. \end{cases}$$

Define also the sequence

$$c[n] = \lceil n^{\frac{1}{3}} \rceil.$$

The estimator assigns $q_{2/3}(1) = 1$, and for all $n > 1$, and $\psi_1^n \in \Psi^n$, it assigns the conditional probability

$$q_{2/3}(\psi_{n+1}|\psi_1^n) = \frac{1}{S_{c[n]}(\psi_1^n)} \cdot \begin{cases} f_{c[n]}(\varphi_1 + 1), & r = 0 \\ (r+1)\frac{f_{c[n]}(\varphi_{r+1}+1)}{f_{c[n]}(\varphi_r)}, & r > 0, \end{cases}$$

where

$$S_{c[n]}(\psi_1^n) \overset{\text{def}}{=} f_{c[n]}(\varphi_1 + 1) + \sum_{\mu=1}^{n} \varphi_\mu \cdot (\mu+1)\frac{f_{c[n]}(\varphi_{\mu+1}+1)}{f_{c[n]}(\varphi_\mu)}$$

is a normalization factor, and, as before, μ_ψ is the multiplicity of ψ, φ_μ is the prevalence of μ, and $r \overset{\text{def}}{=} \mu_{\psi_{n+1}}(\psi_1^n)$.

Theorem 42. For all n,

$$\hat{A}^n(q_{2/3}) \le 2^{\mathcal{O}(n^{\frac{2}{3}})}$$

where the implied constant is at most 10.

Proof outline The theorem holds trivially for $n = 1$. For $n \ge 2$, it can be shown that for all $\psi_1^n \in \Psi^n$,

$$q_{2/3}(\psi_1^n) = \frac{\prod_{\mu=1}^{n}\left((\mu!)^{\varphi_\mu^n} g_{c[n]}(\varphi_\mu^n)\right)}{\prod_{i=1}^{n-1} S_{c[i]}(\psi_1^i)} \cdot \prod_{i=1}^{n-1}\left(\prod_{\mu=1}^{i} \frac{g_{c[i]}(\varphi_\mu^i)}{g_{c[i+1]}(\varphi_\mu^i)}\right)$$

where we used the abbreviation $\varphi_\mu^i \overset{\text{def}}{=} \varphi_\mu(\psi_1^i)$. Therefore,

$$\begin{aligned}\hat{A}^n(q_{2/3}) &\leq \max_{\psi_1^n \in \Psi^n} \prod_{\mu=1}^{n} \frac{\varphi_\mu^n!}{g_{c[n]}(\varphi_\mu^n)} \cdot \max_{\psi_1^n \in \Psi^n} \frac{\prod_{i=1}^{n-1} S_{c[i]}(\psi_1^i)}{n!} \cdot \max_{\psi_1^n \in \Psi^n} \prod_{i=1}^{n-1} \left(\prod_{\mu=1}^{i} \frac{g_{c[i+1]}(\varphi_\mu^i)}{g_{c[i]}(\varphi_\mu^i)} \right) \\ &\overset{\text{def}}{=} \hat{A}_G^n \cdot \hat{A}_S^n \cdot \hat{A}_L^n. \end{aligned} \tag{8.1}$$

Observing that for all $c \in \mathbb{Z}^+$ and $\varphi \in \mathbb{N}$, $g_c(\varphi) \geq \varphi!$, we obtain

$$\hat{A}_G^n \leq 1.$$

Let $c[n] = \gamma \in \mathbb{Z}^+$, then it can shown that for all $\psi_1^n \in \Psi^n$,

$$S_{c[n]}(\psi_1^n) \leq (1 + \frac{1}{\gamma})n + \sqrt{\frac{2n(2\gamma+1)^2}{\gamma}},$$

implying

$$\hat{A}_G^n \leq \left(\frac{1}{n-1} \sum_{i=1}^{n-1} \left(1 + \frac{1}{c[i]} + \sqrt{\frac{2(2c[i]+1)^2}{ic[i]}} \right) \right)^{n-1} \cdot \frac{1}{n}.$$

It can also be shown that

$$\hat{A}_L^n \leq \prod_{i=1}^{n-1} \left(\frac{c_{i+1}}{c_i} \right)^{\sqrt{2ic[i+1]}}.$$

Incorporating these inequalities into Equation (8.1), we obtain

$$\hat{A}^n(q_{2/3}) \leq \prod_{i=1}^{n-1} \left(\frac{c_{i+1}}{c_i} \right)^{\sqrt{2ic[i+1]}} \cdot \left(\frac{1}{n-1} \sum_{i=1}^{n-1} \left(1 + \frac{1}{c[i]} + \sqrt{\frac{2(2c[i]+1)^2}{ic[i]}} \right) \right)^{n-1} \cdot \frac{1}{n}.$$

The theorem follows from the definition of $c[n] = \lceil n^{1/3} \rceil$. □

8.2.2 A low attenuation estimator

Building on an equivalence between set partitions and patterns [2], we obtain an estimator $q_{1/2}$ achieving a sequence attenuation of $2^{\mathcal{O}(\sqrt{n})}$. The estimator assigns $q_{1/2}(1) = 1$, and for all $n > 1$ and $\psi_1^n \in \Psi^n$, it assigns the conditional probability

$$q_{1/2}(\psi_{n+1}|\psi_1^n) = \frac{\sum_{\overline{y} \in \Psi^{2^{\frac{t_n}{2}}}(\psi_1^n.\psi_{n+1})} \tilde{p}(\overline{y})}{\sum_{\overline{y} \in \Psi^{2^{\frac{t_n}{2}}}(\psi_1^n)} \tilde{p}(\overline{y})}. \tag{8.2}$$

where $\frac{t_n}{2} \overset{\text{def}}{=} 2^{\lceil \log n+1 \rceil -1}$ is the largest power of 2 smaller than $n+1$, and

$$\Psi^{2\frac{t_n}{2}}(\psi_1^n) \overset{\text{def}}{=} \{y_1^{2\frac{t_n}{2}} \in \Psi^{2\frac{t_n}{2}} : y_1^n = \psi_1^n\}$$

is the set of patterns of length $2\frac{t_n}{2}$ with prefix ψ_1^n. It follows that for all $n > 1$ and all ψ_1^n

$$q_{1/2}(\psi_1^n) = q_{1/2}(\psi_1^{\frac{t_n}{2}}) \frac{\sum_{\overline{y} \in \Psi^{2\frac{t_n}{2}}(\psi_1^n)} \tilde{p}(\overline{y})}{\sum_{\overline{y} \in \Psi^{2\frac{t_n}{2}}(\psi_1^{\frac{t_n}{2}})} \tilde{p}(\overline{y})}.$$

While this estimator is computationally complex, it achieves a lower attenuation.

We now analyze the attenuation of this estimator.

We now bound the redundancy of $q_{1/2}$.

Theorem 43. For all n,

$$\hat{A}^n(q_{1/2}) \le \exp\left(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}\right).$$

Proof Recall that

$$\hat{A}^n(q_{1/2}) = \max_{\psi_1^n} \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{q_{1/2}(\psi_1^n)}.$$

The theorem holds trivially for $n = 1$. For $n > 1$, rewrite

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} = \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \cdot \frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)}.$$

For all $\psi_1^n \in \Psi^n$, Lemma 44 shows that

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right),$$

and Lemma 45 that

$$\frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{t_n}}{\sqrt{2}-1}\right),$$

and the theorem follows. □

Lemma 44. For all n and ψ_1^n,

$$\frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\tilde{p}^{t_n}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right).$$

Proof Let $\hat{p}_{\psi_1^i}(\psi_1^k)$ denote any probability induced on ψ_1^k by a distribution $\hat{p}_{\psi_1^i}$ maximizing $p(\psi_1^i)$. Observe that

$$\begin{aligned}\hat{p}_{\psi_1^n}(\psi_1^n) &\overset{(a)}{=} \sum_{\overline{y}\in\Psi^{t_n}(\psi_1^n)} \hat{p}_{\psi_1^n}(\overline{y})\\ &\overset{(b)}{\le} \sum_{\overline{y}\in\Psi^{t_n}(\psi_1^n)} \hat{p}_{\overline{y}}(\overline{y})\\ &\overset{(c)}{\le} \left(\sum_{\overline{y}\in\Psi^{t_n}(\psi_1^n)} \tilde{p}(\overline{y})\right) \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right)\\ &= \tilde{p}^{t_n}(\psi_1^n) \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{t_n}\right),\end{aligned}$$

where (a) follows since for all $k \ge n$ and any *i.i.d.* induced distribution,

$$\sum_{\overline{y}\in\Psi^{k}(\psi_1^n)} p(\overline{y}) = p(\psi_1^n),$$

(b) from the definition of maximum-likelihood pattern probabilities, and (c) because Equation (5.6), together with Lemmas 23 and 27 imply that for all n,

$$\tilde{p}(\psi_1^n) = \frac{1}{N(\varphi(\psi_1^n))\,|\Phi^n|} \ge \frac{\hat{p}_{\psi_1^n}(\psi_1^n)}{\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right)}.$$

Note that this inequality corresponds to the upper bound on $\hat{R}(\mathcal{I}_\Psi^n)$ in Section 5.4.2. □

Lemma 45. For all $n \ge 2$ and all ψ_1^n,

$$\frac{\tilde{p}^{t_n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{t_n}}{\sqrt{2}-1}\right).$$

Proof We prove by induction on $i \ge 0$ that for all $2^i < n \le 2^{i+1}$ and all ψ_1^n,

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \le \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^{i+1}}}{\sqrt{2}-1}\right). \tag{8.3}$$

The lemma will follow since for every n, t_n is a power of two.

The basis holds since for $i = 0$, $n = 2$, and all ψ_1^2 satisfy

$$q_{1/2}(\psi_1^2) = \tilde{p}(\psi_1^2) = \frac{1}{2}.$$

To prove the step, note from (5.7) that for $i \geq 1$, all $2^i < n \leq 2^{i+1}$ and ψ_1^n satisfy

$$q_{1/2}(\psi_1^n) = q_{1/2}(\psi_1^{2^i})\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})},$$

hence

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^n)}{q_{1/2}(\psi_1^n)} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} = \frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} \cdot \frac{\tilde{p}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})}. \tag{8.4}$$

By the induction hypothesis,

$$\frac{\tilde{p}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} = \frac{\tilde{p}^{2^i}(\psi_1^{2^i})}{q_{1/2}(\psi_1^{2^i})} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{2^i}}{\sqrt{2}-1}\right). \tag{8.5}$$

By definition (5.6) and Lemma 23,

$$\tilde{p}(\psi_1^{2^i}) = \frac{1}{N(\varphi(\psi_1^{2^i}))\left|\Phi^{2^i}\right|} \geq \frac{1}{N(\varphi(\psi_1^{2^i}))\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}\right)}.$$

On the other hand, distinct patterns $\psi_1^{2^i}$ have disjoint sets $\Psi^{2^{i+1}}(\psi_1^{2^i})$, while patterns of the same profile have the same probability $\tilde{p}^{2^{i+1}}(\psi_1^{2^i})$, hence

$$N\left(\varphi(\psi_1^{2^i})\right) \cdot \tilde{p}^{2^{i+1}}(\psi_1^{2^i}) \leq \sum_{\overline{y}\in\Psi^{2^{i+1}}(\psi_1^{2^i})} \tilde{p}(\overline{y}) = 1,$$

and thus

$$\tilde{p}^{2^{i+1}}(\psi_1^{2^i}) \leq \frac{1}{N(\varphi(\psi_1^{2^i}))}.$$

It follows that

$$\frac{\tilde{p}^{2^{i+1}}(\psi_1^{2^i})}{\tilde{p}(\psi_1^{2^i})} \leq \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{2^i}\right).$$

Incorporating this inequality and (8.5) in (8.4), we get (8.3). □

8.3 Lower bound on compression of patterns

A lower bound the attenuation of any estimator over patterns follows from Theorem 26,

$$\exp\left(\frac{3}{2}n^{\frac{1}{3}}(1+o(1))\right). \qquad \square$$

8.4 Examples

To better understand the behavior of the diminishing-attenuation estimators, we consider the conditional probabilities assigned to some simple sequences by the low-complexity estimator $q_{\frac{1}{3}}$, and compare it to what one would logically expect.

Consider first the sequence $aaa\ldots$. Since the same symbol always repeats, after observing a large portion of this sequence, one would guess that the next symbol would be 'a' as well. Indeed after observing n elements, the estimator assigns probability $1-\Theta(\frac{1}{n})$ for the next symbol being 'a' and probability $\Theta(\frac{1}{n})$ to a new symbol.

For the alternating sequence $abab\ldots$, one would predict probability half for the next symbol being each of 'a' and 'b. Similarly, the estimator assigns probability $\Theta(\frac{1}{n})$ to a new symbol and splits the remaining probability evenly between 'a' and 'b'.

Of course, we are more interested in the behavior of the estimator when the number of symbols appearing is large. In the extreme case where all symbols are different, for example, after observing the sequence $abc\ldots$, we would expect the next symbol to be new. Indeed the estimator assigns probability $1-\Theta(\frac{1}{n^{5/3}})$ that the next symbol will be new.

But for large-alphabet sequences where the probability of new is not 1, intuition may not serve well. Consider perhaps the simplest such case, the sequence $aabbcc\ldots$. After observing an even number n of symbols, e.g., $aabbcc$, the estimator assigns probability 1/4 to the next symbol being new and $3/(2n)$ to each of the preceding symbols, and after observing an odd number n of symbols, e.g., $aabbc$, the estimator assigns probability approaching 1 to the next symbol being the same as the last one, e.g., 'c' in this example.

These estimations may be at odds with the intuition saying that since every other element so far was new, the next symbol will be new with probability 1/2. One possible explanation for the lower probability of new assigned by the estimator is that it can be shown [72] that after seeing n symbols of the sequence, the most likely alphabet is of size $0.62n$, hence, roughly speaking, the probability of seeing a new one is about $(0.12n)/(0.62n) \approx 0.2$.

Acknowledgements

We would like to thank N. Jevtić, A. Noga, G. Shamir, W. Szpankowski, K. Viswanathan for helpful discussions. This chapter is adapted from A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427—431, October 17 2003. (See also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003). The author was a primary researcher of the sections reproduced here.

Chapter 9

Entropy rate of patterns

In this chapter we determine the entropy rate of patterns of certain processes, and for *i.i.d.* processes, we bound the speed at which the per-symbol pattern entropy converges to this rate, and show that patterns satisfy an asymptotic equipartition property. To derive some of these results, we upper bound the probability that the n'th variable in a random process differs from all preceding ones. We note that related entropy-rate results were independently derived by Gemelos and Weissman [52, 53], and that subsequent results appeared in [54, 55, 56].

We denote a random n-symbol sequence by $\overline{X} = X_1, \ldots, X_n$ and its pattern by $\overline{\Psi} = \Psi_1, \ldots, \Psi_n$. The entropy of the sequence is

$$H(\overline{X}) = \sum_{\overline{x}} p(\overline{x}) \log \frac{1}{p(\overline{x})},$$

and its entropy rate is the asymptotic per-symbol entropy

$$\mathcal{H}_X = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}).$$

Similarly, the *pattern entropy* is

$$H(\overline{\Psi}) = \sum_{\overline{\psi}} p(\overline{\psi}) \log \frac{1}{p(\overline{\psi})},$$

and the *pattern entropy rate* is the asymptotic per-symbol entropy

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}).$$

These concepts are illustrated by the following examples.

Example 2. Consider the process $X_1, X_2, \ldots$ where $X_1 = 1$ and for $n = 2, 3, \ldots$, X_n is distributed uniformly over $\{X_{n-1}+1, \ldots, X_{n-1}+n\}$. For example, 1,2,3,4 and 1,3,6,10 are two equally-likely realizations of $X_1, \ldots, X_4$. Since $X_1, \ldots, X_n$ can assume $n!$ equally likely realizations, the sequence entropy is

$$H(\overline{X}) = \log n!,$$

and its entropy rate is

$$\mathcal{H}_X = \lim_{n\to\infty} \frac{1}{n} \log n! = \infty.$$

On the other hand, $X_n > X_i$ for all $i = 1, \ldots, n-1$, hence the pattern is always $\overline{\Psi} = 12\ldots n$, implying zero pattern entropy rate,

$$\mathcal{H}_\Psi = 0. \qquad \square$$

Example 3. Consider independent Bernoulli-half trials $X_1, X_2, \ldots$. As with all *i.i.d.* distributions,

$$H(\overline{X}) = nH(X_1),$$

hence the sequence entropy rate is

$$\mathcal{H}_X = H(X_1) = 1.$$

It is easy to verify that the resulting patterns are all 2^{n-1} sequences over $\{1, 2\}$ starting with 1. Each pattern corresponds to two possible trial sequences hence has probability $2^{-(n-1)}$. If follows that

$$H(\overline{\Psi}) = n - 1,$$

and the pattern entropy rate is

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{n-1}{n} = 1. \qquad \square$$

Note that in the last example, $\mathcal{H}_\Psi = \mathcal{H}_X$. We show that for all finite entropy discrete stationary processes,

$$\mathcal{H}_\Psi = \mathcal{H}_X. \tag{9.1}$$

Recall that Kieffer [11] showed that *i.i.d.* distributions over infinite alphabets entail an infinite per-symbol redundancy,

$$\mathcal{R}_X = \infty, \tag{9.2}$$

while, as shown in [2], the patterns of such processes incur asymptotically zero per-symbol redundancy,

$$\mathcal{R}_{\Psi} = 0. \tag{9.3}$$

These results suggest conveying a sequence $\overline{X}$ by first describing its pattern $\overline{\Psi}$ and then the dictionary Δ that maps $\{1,\ldots,\Psi_n\}$ to $\{X_1,\ldots,X_n\}$. For example, if $\overline{X}$ =“*abracadabra*”, we can convey the pattern $\overline{\Psi} = 12314151231$ and the dictionary $\Delta(1) = a$, $\Delta(2) = b$, $\Delta(3) = r$, $\Delta(4) = c$, and $\Delta(5) = d$.

Since the pattern and dictionary determine the sequence, it is easy to see that

$$\mathcal{R}_{\Psi} + \mathcal{R}_{\Delta|\Psi} \geq \mathcal{R}_X.$$

Hence

$$\mathcal{R}_{\Delta|\Psi} = \infty. \tag{9.4}$$

Together, these results imply that for *i.i.d.* distributions over arbitrary alphabets, not knowing the underlying distribution results in infinite redundancy (9.2). Yet all the redundancy is associated with describing the dictionary (9.4), and none with the pattern (9.3).

Two comparisons between these and existing pattern-compression results are in order. For simplicity, we describe them using discrete *i.i.d.* distributions.

First, while the original sequence and its pattern have the same (asymptotic per-symbol) entropy (9.1), the (asymptotic per-symbol) redundancy of the sequence is infinite (9.2) whereas that of the pattern diminishes to zero (9.3). Hence, when the distribution is known, describing the sequence and its pattern require the same number of bits, but when the distribution is not known, the sequence may require infinitely many additional bits whereas the pattern requires none.

Additionally, since (a) $\overline{X}$ determines $\overline{\Psi}$, and (b) given $\overline{\Psi}$ there is a 1-1 correspondence between $\overline{X}$ and Δ, we obtain

$$H(\overline{X}) \overset{(a)}{=} H(\overline{\Psi}) + H(\overline{X}|\overline{\Psi}) \overset{(b)}{=} H(\overline{\Psi}) + H(\Delta|\overline{\Psi}).$$

It follows that

$$\mathcal{H}_{\Delta|\Psi} \overset{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} H(\Delta|\overline{\Psi}) = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}) - \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}) = \mathcal{H}_X - \mathcal{H}_{\Psi} = 0. \tag{9.5}$$

Hence, while when the distribution is not known, essentially all the redundancy in describing a sequence derives from describing the dictionary (9.4) and none from the pattern (9.3), when the distribution is known, essentially all the bits go towards describing the pattern (9.1), and none towards the dictionary (9.5).

9.1 The probability of innovation

The essential difference between a sequence and its pattern is that the latter groups all hitherto unseen symbols into a single *new* element. For symbols that have been observed, the symbols and their indices in the pattern have 1-1 correspondence given the past sequence. To relate sequence and pattern entropy, we therefore show that for any discrete stationary distribution the probability of observing new elements decreases to zero with time. We begin with some definitions.

For $n \geq 1$, let $x^{n-1} = x_1, \ldots, x_{n-1}$ and let $\mathcal{A}(x^{n-1}) = \{x_1, \ldots, x_{n-1}\}$ be the set of elements observed in x^{n-1}. For a random process $X_1, X_2, \ldots$, let

$$I_n = \begin{cases} 1 & X_n \notin \mathcal{A}(X^{n-1}), \\ 0 & \text{otherwise.} \end{cases}$$

indicate whether the n'th symbol is new, and let

$$M_n \stackrel{\text{def}}{=} |\mathcal{A}(X^n)| = \sum_{i=1}^{n} I_i,$$

be the number of distinct symbols in $X_1, \ldots, X_n$. Finally. the *innovation probability* of the process at time n is

$$\nu_n \stackrel{\text{def}}{=} p(I_n = 1) = EI_n,$$

the probability that the nth symbol differs from all previous ones.

Since this section concerns only discrete distributions, assume without loss of generality that these strings are drawn from $\mathbb{N} = \{1, 2, \ldots\}$. For a stationary distribution, let $p_j \stackrel{\text{def}}{=} p(X_n = j)$ denote the marginal probability that the nth random variable is j. The distribution's *marginal entropy*,

$$H \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} p_j \log \frac{1}{p_j}$$

is the entropy of each X_n.

The next lemma shows that for any stationary distribution the expected number of symbols grows sublinearly with n and provides a stronger bound for distributions with finite marginal entropy.

Lemma 46. For all discrete stationary distributions,

$$EM_n = o(n)$$

and if, in addition, the distribution has finite marginal entropy H, then,

$$EM_n \le \frac{nH}{\log n}(1 + o(1)).$$

Proof For $j \in \mathbb{N}$, let

$$I_{n,j} = \begin{cases} 1 & X_n = j \notin \mathcal{A}(X^{n-1}) \\ 0 & \text{else,} \end{cases}$$

indicate whether X_n is new and equals j. Then,

$$I_n = \sum_{j=1}^{\infty} I_{n,j}.$$

For any function k_n of n,

$$M_n = \sum_{i=1}^{n}\sum_{j=1}^{k_n} I_{i,j} + \sum_{i=1}^{n}\sum_{j=k_n+1}^{\infty} I_{i,j}.$$

Since any element j can be new at most once,

$$\sum_{i=1}^{n}\sum_{j=1}^{k_n} I_{i,j} = \sum_{j=1}^{k_n}\sum_{i=1}^{n} I_{i,j} \le \sum_{j=1}^{k_n} 1 = k_n,$$

and, since p_j denotes the probability that $X_n = j$,

$$E\left(\sum_{i=1}^{n}\sum_{j=k_n+1}^{\infty} I_{i,j}\right) = \sum_{i=1}^{n}\sum_{j=k_n+1}^{\infty} p(X_n = j, I_n = 1) \le \sum_{i=1}^{n}\sum_{j=k_n+1}^{\infty} p_j = n \cdot \sum_{j=k_n+1}^{\infty} p_j.$$

Letting k_n increase to infinity as $o(n)$, we obtain

$$EM_n \le k_n + n \cdot \sum_{j=k_n+1}^{\infty} p_j = o(n),$$

where the equality follows since $\sum_{j=k_n+1}^{\infty} p_j = o(1)$.

To prove the second part of the lemma, assume without loss of generality that the probabilities p_j are non-increasing. Then $p_j \le \frac{1}{j}$ for all $j \ge 1$, and

$$\sum_{j=k_n+1}^{\infty} p_j < \frac{1}{\log k_n} \cdot \sum_{j=k_n+1}^{\infty} p_j \log j \le \frac{1}{\log k_n} \cdot \sum_{j=k_n+1}^{\infty} p_j \log \frac{1}{p_j} \le \frac{H}{\log k_n}.$$

Hence

$$EM_n \le k_n + n \cdot \frac{H}{\log k_n},$$

and the lemma follows by letting

$$k_n = \frac{nH}{\log^2(nH)}. \qquad \square$$

In Corollary 48, we apply this lemma to show that the innovation probability of any stationary distribution diminishes with time, a result used in the next section to determine the entropy rate of patterns. We first show that the innovation probability of any stationary process decreases monotonically.

Lemma 47. For any stationary process,

$$\nu_n \ge \nu_{n+1}.$$

Proof For every stationary process and every n,

$$\begin{aligned}\nu_n = p\Big(X_n \notin \mathcal{A}(\{X_1,\dots,X_{n-1}\}\Big) &= p\Big(X_{n+1} \notin \mathcal{A}(\{X_2,\dots,X_n\})\Big) \\ &\ge p\Big(X_{n+1} \notin \mathcal{A}(\{X_1,\dots,X_n\})\Big) = \nu_{n+1}. \qquad \square\end{aligned}$$

Corollary 48. For any discrete stationary process,

$$\lim_{n\to\infty} \nu_n = 0,$$

and if, in addition, the distribution has finite marginal entropy H, then for all n,

$$\nu_n \le \frac{H}{\log n}(1 + o(1)).$$

Proof From Lemmas 46 and 47,

$$n\nu_n \le \sum_{i=1}^{n} \nu_i = \sum_{i=1}^{n} EI_i = E\sum_{i=1}^{n} I_i = EM_n = o(n),$$

and if the distribution has finite marginal entropy H, then

$$n\nu_n \leq EM_n \leq \frac{nH}{\log n}(1+o(1)). \qquad \square$$

For *i.i.d.* distributions, the last bound can be slightly improved.

Lemma 49. For all discrete *i.i.d.* distributions with finite entropy H and all n,

$$\nu_n \leq \frac{H}{\log n}.$$

Proof One simple proof notes that for every $0 < p < 1$ and $n \geq 1$, the Taylor series expansion of $\ln(1-x)$ yields

$$\ln\frac{1}{p} = -\ln(1-(1-p)) \geq \sum_{i=1}^{n-1}\frac{(1-p)^i}{i} \geq (1-p)^{n-1}\sum_{i=1}^{n-1}\frac{1}{i} \geq (1-p)^{n-1}\ln n.$$

Therefore,

$$\nu_n = \sum_{x\in\mathcal{A}} p(x)(1-p(x))^{n-1} \leq \frac{1}{\log n}\sum_{x\in\mathcal{A}} p(x)\log\frac{1}{p(x)} = \frac{H}{\log n}.$$

An alternate proof uses the following relation between innovation and the entropy. Observe that

$$\begin{aligned}\sum_{i\geq 1}\frac{\nu_{i+1}}{i} &= \sum_{i\geq 1}\frac{1}{i}\sum_{j\geq 1}p_j(1-p_j)^i\\ &= \sum_{j\geq 1}p_j\sum_{i\geq 1}\frac{1}{i}(1-p_j)^i\\ &= \sum_{j\geq 1}p_j\ln\frac{1}{p_j}.\end{aligned}$$

The lemma then follows by observing that the innovation is strictly decreasing, hence

$$\sum_{j\geq 1}p_j\ln\frac{1}{p_j} = \sum_{i\geq 1}\frac{\nu_i}{i} > \sum_{i=1}^{n}\frac{\nu_i}{i} \geq \nu_n\sum_{i=1}^{n}\frac{1}{i} \geq \nu_n\ln n. \qquad \square$$

Note that while this bound is not tight for all *i.i.d.* distributions, for example the independent Bernoulli-half process has $\nu_n = H/2^{n-1}$, it is tight in the following sense.

Lemma 50. For all positive H and ϵ there is a distribution with entropy H and innovation probability

$$\nu_n = \Omega\left(\frac{1}{(\log n)^{1+\epsilon}}\right).$$

Proof We first show that for any $\epsilon > 0$, there is a finite entropy distribution with $\nu_n = \Omega\left(\frac{1}{(\log n)^{1+\epsilon}}\right)$. Define the probability distribution $(p_2, p_3, \ldots)$ by

$$p_i = \frac{1}{S\, i\, (\log i)^{2+\epsilon}},$$

where

$$S = \sum_{i \geq 2} \frac{1}{i(\log i)^{2+\epsilon}} < \infty \tag{9.6}$$

is a normalization factor. The distribution's entropy is

$$H^\epsilon = \sum_{i \geq 2} \frac{\log i + (2+\epsilon)\log\log i + \log S}{S\, i(\log i)^{2+\epsilon}} < \infty, \tag{9.7}$$

and, observing that $S > 1/2$, we obtain that for all $n \geq 4$,

$$\begin{aligned}
\nu_n &> \sum_{i \geq 2} (1 - (n-1)p_i)p_i \\
&> \sum_{i \geq n} \frac{1}{S\, i(\log i)^{2+\epsilon}} - \sum_{i \geq n} \frac{n-1}{S^2\, i^2(\log i)^{4+2\epsilon}} \\
&> \sum_{i \geq n} \frac{1}{S\, i(\log i)^{2+\epsilon}} - \sum_{i \geq n} \frac{1}{2S\, i(\log i)^{2+\epsilon}} \\
&= \Theta\left(\frac{1}{(\log n)^{1+\epsilon}}\right).
\end{aligned}$$

The distribution therefore has the desired innovation probability, and we now modify it to also have the required entropy H. The modification depends on whether H is larger or smaller than H^ϵ.

If $H > H^\epsilon$, consider the distribution $(p'_2, p'_3, \ldots)$ defined by

$$p'_i = \frac{1}{S\, i\, (\log i)^{2+\delta}}$$

where $0 < \delta \leq \epsilon$ and S is a normalization factor defined as before. Its entropy can be made arbitrarily large by decreasing δ and its innovation is

$$\nu_n = \Omega\left(\frac{1}{(\log n)^{1+\delta}}\right) = \Omega\left(\frac{1}{(\log n)^{1+\epsilon}}\right).$$

If $H < H^\epsilon$, consider the distribution $(p''_1, p''_2, \ldots)$ with $p''_1 = 1 - q$ for some $0 < q < 1$ and

$$p''_i = \frac{q}{S\, i\, (\log i)^{2+\epsilon}}$$

for $i \geq 2$, where S is defined by (9.6). Its entropy is

$$h(q) + qH^\epsilon,$$

where H^ϵ is defined in (9.7). This entropy can be made equal to any value $0 < H < H^\epsilon$ by an appropriate choice of q. Clearly the new distribution also satisfies

$$\nu_n = \Omega\biggl(\frac{1}{(\log n)^{1+\epsilon}}\biggr). \qquad \square$$

9.2 The entropy rate of patterns

We determine the entropy rate of patterns of certain processes. We observe that when the alphabet is finite, the entropy rates of the process and its pattern coincide, and extend this result to all discrete processes that are either *i.i.d.*, or finite-entropy stationary. For *i.i.d.* distributions with a continuous component we show that the pattern entropy rate equals that of a modified process where the continuous probability is assigned to a new discrete element. We note that similar results were independently obtained by Gemelos and Weissman [52, 53].

It is easy to see that whenever the alphabet $\mathcal{A}$ is finite, the process and pattern entropy rates coincide. Observe that

$$H(\overline{X}) - \log|\mathcal{A}|! \leq H(\overline{\Psi}) \leq H(\overline{X}), \tag{9.8}$$

where the upper bound follows as the sequence determines the pattern, and the lower bound follows as, for the same reason,

$$H(\overline{X}) = H(\overline{\Psi}) + H(\overline{X}|\overline{\Psi}) \tag{9.9}$$

and every pattern can derive from at most $|\mathcal{A}|!$ sequences, hence

$$H(\overline{X}|\overline{\Psi}) \leq \log|\mathcal{A}|!.$$

Taking limits in (9.8) we see that for all distributions over finite alphabets,

$$\mathcal{H}_\Psi = \lim_{n\to\infty} \frac{1}{n} H(\overline{\Psi}) = \lim_{n\to\infty} \frac{1}{n} H(\overline{X}) = \mathcal{H}_X. \tag{9.10}$$

Note that for *i.i.d.* processes, bounds similar to (9.8) appeared in [22, 51, 54, 55].

The rest of the section extends (9.10) to distributions over infinite alphabets. We use the following lemma relating conditional pattern entropy and the pattern entropy rate.

Lemma 51. For any process, if

$$H(\Psi_n|\Psi^{n-1}) \geq h_n.$$

and

$$\lim_{n\to\infty} h_n = \mathcal{H}_X,$$

then

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof Since X^n determines Ψ^n and $H(\Psi^n) = \sum_{i=1}^n H(\Psi_i|\Psi^{i-1})$,

$$\frac{1}{n} H(X^n) \geq \frac{1}{n} H(\Psi^n) \geq \frac{1}{n} \sum_{i=1}^n h_i.$$

Taking limits as $n \to \infty$, the lemma follows because Cesáro's mean theorem implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n h_j = \mathcal{H}_X. \qquad \square$$

We begin with *i.i.d.* distributions, and among them start with those over discrete alphabets. We show that a random sequence is likely to contain all high-probability elements, and that when this happens, the conditional entropy of the pattern approaches that of the sequence.

As in Section 9.1 we assume without loss of generality that the alphabet is $\mathbb{N} = \{1, 2, \ldots\}$ and let $p_i \overset{\text{def}}{=} p(X_n = i)$. For $\epsilon \geq 0$, we let

$$A_\epsilon \overset{\text{def}}{=} \{i : p_i > \epsilon\}$$

be the set of all elements whose probability exceeds ϵ.

Theorem 52. For all discrete *i.i.d.* distributions,

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof We first show that a random sequence is likely to contain all elements of sufficiently high probability. More precisely, recall that $\mathcal{A}(X^n)$ is the set of all elements in X^n, and that $A_{\frac{\ln n}{n}}$ is the set of all elements whose probability exceeds $\frac{\ln n}{n}$. Clearly, $|A_{\frac{\ln n}{n}}| \le \frac{n}{\ln n}$, hence

$$p\Big(A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(X^n)\Big) > 1 - \frac{n}{\ln n}\left(1 - \frac{\ln n}{n}\right)^n > 1 - \frac{1}{\ln n}.$$

Let

$$J_n = \begin{cases} 1 & A_{\frac{\ln n}{n}} \subseteq \mathcal{A}(X^n) \\ 0 & \text{otherwise} \end{cases}$$

indicate whether X^n contains all high-probability elements. Then

$$\begin{aligned} H(\Psi_{n+1}|\Psi^n) &\ge H(\Psi_{n+1}|X^n) \\ &\ge H(\Psi_{n+1}|X^n, J_n) \\ &\ge p(J_n = 1) H(\Psi_{n+1}|X^n, J_n = 1) \\ &\ge \left(1 - \frac{1}{\ln n}\right) \sum_{i \in A_{\frac{\ln n}{n}}} p_i \log \frac{1}{p_i} \\ &\overset{\text{def}}{=} \left(1 - \frac{1}{\ln n}\right) H(A_{\frac{\ln n}{n}}). \end{aligned}$$

The theorem follows from Lemma 51 as

$$\lim_{n \to \infty} H(A_{\frac{\ln n}{n}}) = \mathcal{H}_X. \qquad \square$$

For mixed *i.i.d.* distributions we show that the entropy rate of the pattern equals that of a slightly modified process. Let X be a random variable drawn from a mixed distribution p with discrete support A_0 and continuous probability q. Define $\tilde{X}$ to be the discrete random variable obtained from X by replacing all elements present in the continuous support with a single new discrete element. Then

$$\mathcal{H}_{\tilde{X}} = \sum_{i \in A_0} p_i \log \frac{1}{p_i} + q \log \frac{1}{q} = H(A_0) + q \log \frac{1}{q}.$$

Theorem 53. For all *i.i.d.* distributions,

$$\mathcal{H}_\Psi = \mathcal{H}_{\tilde{X}}.$$

Proof Since $\tilde{X}^n$ determines Ψ^n with probability 1, we proceed as in Theorem 52. Recall the definitions of J_n, $A_{\frac{\ln n}{n}}$, and $H(A_{\frac{\ln n}{n}})$, and let $\mathcal{X}(\overline{x})^c$ denote the set of symbols not in $\overline{x}$. Proceeding similarly to the proof of Theorem 52, we obtain

$$\begin{aligned} H(\Psi_{n+1}|\Psi^n) &\geq p(J_n = 1)H(\Psi_{n+1}|\tilde{X}^n, J_n = 1) \\ &\geq \left(1 - \frac{1}{\ln n}\right)\left(H(A_{\frac{\ln n}{n}}) + \min_{\overline{x}:A_{\frac{\ln n}{n}} \subseteq \mathcal{X}(\overline{x})} p(\mathcal{X}(\overline{x})^c)\log\frac{1}{p(\mathcal{X}(\overline{x})^c)}\right). \end{aligned}$$

The theorem follows by applying Lemma 51 to $\tilde{X}^n$ as

$$\lim_{n\to\infty} H(A_{\frac{\ln n}{n}}) = H(A_0),$$

and

$$\lim_{n\to\infty} \min_{\overline{x}:A_{\frac{\ln n}{n}} \subseteq \mathcal{X}(\overline{x})} p(\mathcal{X}(\overline{x})^c) = q. \qquad \square$$

We now address stationary processes. Note that while Theorem 52 shows that $\mathcal{H}_\Psi = \mathcal{H}_X$ for all discrete *i.i.d.* processes, even those with infinite entropy, as the next example indicates, this equality cannot hold for all discrete stationary processes with infinite entropy.

Example 4. Consider the constant stationary process $X_1 = X_2 = \ldots$ defined by

$$p_j = p(X_n = j) = \frac{1}{S}\frac{1}{j\log^2 j},$$

where S is a normalization factor. Then,

$$H(X_1) = \sum_{j=1}^{\infty} p_j \log\frac{1}{p_j} = \infty,$$

hence

$$\mathcal{H}_X = \infty.$$

On the other hand, the pattern is always $11\ldots1$, hence

$$\mathcal{H}_\Psi = 0. \qquad \square$$

To prove that $\mathcal{H}_\Psi = \mathcal{H}_X$ for all discrete stationary processes with finite entropy, we use the innovation results of Section 9.1. We show that the probability that X_n is new, hence more "informative" than Ψ_n, is low for likely X^{n-1}, and that when X_n is not new, the conditional entropy of the pattern is roughly $\mathcal{H}_X$.

Theorem 54. For all finite-entropy discrete stationary processes,

$$\mathcal{H}_\Psi = \mathcal{H}_X.$$

Proof As before, we lower bound the conditional pattern entropy with a term that approaches $\mathcal{H}_X$. We show that

$$H(\Psi_n|\Psi^{n-1}) \geq H(X_n|X^{n-1}) - o(1),$$

and the theorem will follow from Lemma 51 as for all finite-entropy stationary processes,

$$\lim_{n\to\infty} H(X_n|X^{n-1}) = \mathcal{H}_X.$$

Recall that for $n \geq 1$, I_n indicates whether X_n is new, hence

$$\begin{aligned}
H(X_n|X^{n-1}) &= H(X_n, I_n|X^{n-1}) \\
&= H(I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n) \\
&= H(I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 0)p(I_n = 0) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(I_n|X^{n-1}) + H(\Psi_n|X^{n-1}, I_n = 0)p(I_n = 0) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(I_n|X^{n-1}) + H(\Psi_n|X^{n-1}, I_n = 0)p(I_n = 0) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&\qquad + H(\Psi_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(I_n|X^{n-1}) + H(\Psi_n|X^{n-1}, I_n) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(\Psi_n, I_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&= H(\Psi_n|X^{n-1}) + H(X_n|X^{n-1}, I_n = 1)p(I_n = 1) \\
&\leq H(\Psi_n|\Psi^{n-1}) + H(X_n|I_n = 1)p(I_n = 1)
\end{aligned}$$

We now use Corollary 48 to show that

$$H(X_n|I_n = 1)p(I_n = 1) = o(1).$$

Recall that $p_j = p(X_n = j)$, that

$$A_{\nu_n} = \{j : p_j > \nu_n\}$$

is the set of all elements whose probability exceeds ν_n, that $\nu_n = p(I_n = 1)$, and that $I_{n,j}$ indicates whether X_n is new and equals j. Define

$$\nu_{n,j} \overset{\text{def}}{=} p(I_{n,j} = 1) = \sum_{x^{n-1}:j\notin x^{n-1}} p(x^{n-1}, x_n = j),$$

so that $\nu_n = \sum_{j=1}^{\infty} \nu_{n,j}$. Then

$$\begin{aligned}
H(X_n|I_n = 1)p(I_n = 1) &= \sum_{j=1}^{\infty} \nu_{n,j} \log \frac{\nu_n}{\nu_{n,j}} \\
&= \nu_n \log \nu_n + \sum_{j\in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} + \sum_{j\notin A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} \\
&\overset{(a)}{\le} \nu_n \log (|A_{\nu_n}| + 1) + \sum_{j\notin A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} \\
&\overset{(b)}{\le} \nu_n \log \left(\frac{1}{\nu_n} + 1\right) + \sum_{j\notin A_{\nu_n}} p_j \log \frac{1}{p_j} \\
&\overset{(c)}{=} o(1),
\end{aligned}$$

where (a) follows because

$$\begin{aligned}
\sum_{j\in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} &\le \sum_{j\in A_{\nu_n}} \nu_{n,j} \log \frac{1}{\nu_{n,j}} + (\nu_n - \sum_{j\in A_{\nu_n}} \nu_{n,j}) \log \frac{1}{\nu_n - \sum_{j\in A_{\nu_n}} \nu_{n,j}} \\
&\le \nu_n \log (|A_{\nu_n}| + 1) + \nu_n \log \frac{1}{\nu_n},
\end{aligned}$$

(b) follows as $|A_{\nu_n}| < \frac{1}{\nu_n}$ and $\nu_{n,j} \le p_j \le \nu_n$, which for sufficiently large n is smaller than $\frac{1}{e}$, and (c) follows as Corollary 48 implies that $\nu_n \to 0$, and $\mathcal{H}_X < \infty$ implies that the marginal entropy is finite, hence

$$\lim_{n\to\infty} \sum_{j\notin A_{\nu_n}} p_j \log \frac{1}{p_j} = 0. \qquad \square$$

9.3 The rate of convergence

In the previous section we determined the pattern entropy rate—the limit of the per-symbol pattern entropy—of *i.i.d.* and certain related distributions. We now address the *convergence rate*

$$\rho_{X,n} \overset{\text{def}}{=} \left|\frac{1}{n}H(\overline{\Psi}) - \mathcal{H}_{\Psi}\right|$$

at which this limit is attained. In this section, we consider only discrete *i.i.d.* distributions. Then

$$\rho_{X,n} = \frac{1}{n}(H(\overline{X}) - H(\overline{\Psi})) = \frac{1}{n}H(\overline{X}|\overline{\Psi}),$$

where the first equality follows from Theorem 52, and the second from (9.9).

We first show that $\rho_{X,n}$ does not diminish uniformly for all distributions, or even for all distributions with a given entropy. We then bound $\rho_{X,n}$ in terms of the second moment of the self information.

To show that $\rho_{X,n}$ does not diminish uniformly, the next example shows that it can be made arbitrarily high for all n.

Example 5. The *i.i.d.* process $X_1, X_2, \ldots$, where each X_i is distributed uniformly over $\{1, \ldots, k\}$, has

$$\rho_{X,n} = \frac{1}{n}H(\overline{X}|\overline{\Psi}) \geq \frac{1}{n}H(X_1|\overline{\Psi}) = \frac{\log k}{n},$$

which can be made arbitrarily high by choosing a sufficiently large k. □

While the example shows that $\rho_{X,n}$ does not diminish uniformly for all *i.i.d.* distributions (and in fact is unbounded), the processes it uses have unbounded entropy themselves. It is natural to ask whether $\rho_{X,n}$ diminishes uniformly for all *i.i.d.* processes with a given entropy. The next example answers this question in the negative, showing that for all n, $\rho_{X,n}$ can be made arbitrarily close to the process entropy.

Example 6. For all n and $\epsilon > 0$, we construct an *i.i.d.* distribution that satisfies $\rho_{X,n} > H - \epsilon$.

Given $H > 0$, for all $k \geq 2^H$, there exists q_k such that

$$h(q_k) + q_k \log k = H.$$

Let $p^k = (1 - q_k, q_k/k, \ldots, q_k/k)$ be the distribution on $[k+1]$, where the element "1" has probability $1 - q_k$ and all the remaining k elements have probability q_k/k. The entropy of p_k is therefore H by construction.

For all $M \geq 2^H$, it follows that if $k > M$,

$$q_k < \frac{H}{\log M}.$$

As k increases to infinity, q_k diminishes, implying that $\exists N_1$ such that for $k \geq N_1$,

$$h(q_k) \leq \frac{\epsilon}{2}.$$

Consider an *i.i.d.* process on $[k+1]^n$ with the marginal distribution being p^k. For all n, by making k sufficiently large, the probability that any element with probability q_k/k appears more than once can be made arbitrary small. In particular, $\exists N_2$ such that for $k \geq N_2$,

$$\begin{aligned}\Pr(\text{Any of } \{2,\ldots,k+1\} \text{ appear} \geq 2 \text{ times }) &= k\left[1-\Big(1-\frac{q_k}{k}\Big)^n - n\frac{q_k}{k}\Big(1-\frac{q_k}{k}\Big)^{n-1}\right]\\ &\leq nq_k\\ &< \frac{\epsilon}{2n\log n}.\end{aligned}$$

Hence with probability $\geq 1-\frac{\epsilon}{2n\log n}$, there is a 1-1 correspondence between the pattern and the set of locations where the element 1 appears. Consequently,

$$H(\overline{\Psi}) < h(q_k) + \frac{\epsilon}{2n\log n} n \log n = h(q_k) + \frac{\epsilon}{2}.$$

It follows that for any fixed n, for $k \geq \max\{N_1, N_2\}$,

$$\rho_{X,n} = \frac{1}{n}H(\overline{X}|\overline{\Psi}) = \frac{1}{n}\big(H(\overline{X}) - H(\overline{\Psi})\big) \geq q\log k - \frac{\epsilon}{2} \geq H - \epsilon. \qquad \square$$

In the preceding examples we increased $\rho_{X,n}$ by constructing successively flatter distributions, raising the possibility that $\rho_{X,n}$ will diminish when the distribution $p_1, p_2, \ldots$ diminishes to 0 sufficiently quickly. In Theorem 56 and Corollary 57 we bound $\rho_{X,n}$ in terms of the second moment of the self information

$$\sigma^2 \stackrel{\text{def}}{=} \sum_{i\geq 1} p_i \log^2 \frac{1}{p_i}.$$

To do so, we first prove the following technical lemma.

Lemma 55. For any discrete distribution p, and all $I \geq 1$,

$$\sum_{i\geq I} p_i \log \frac{1}{p_i} \leq \sigma \sqrt{\sum_{i\geq I} p_i}.$$

Proof Using the Cauchy-Schwartz Inequality,

$$\sum_{i\geq I} p_i \log \frac{1}{p_i} = \sum_{i\geq I} \sqrt{p_i \cdot p_i \log^2 \frac{1}{p_i}} \leq \left(\sum_{i\geq I} p_i\right)^{1/2} \left(\sum_{i\geq I} p_i \log^2 \frac{1}{p_i}\right)^{1/2}. \qquad \square$$

Theorem 56. For all discrete *i.i.d.* distributions with entropy H,

$$H\left(1-\Theta\left(\frac{\sigma^2}{H\log n}\right)^{1/3}\right) \;\leq\; \frac{1}{n}H(\overline{\Psi}) \;\leq\; H.$$

Proof Let

$$\epsilon_n \stackrel{\text{def}}{=} \left(\frac{2H^2}{\sigma \log n}\right)^{2/3}$$

and

$$T_n \stackrel{\text{def}}{=} \left\{ x^{n-1} : p(I_n = 1|x^{n-1}) = \sum_{x\notin\mathcal{X}(\overline{x})} p(x) \leq \epsilon_n \right\},$$

be the set of strings whose missing mass is at most ϵ_n. From Lemma 55,

$$\min_{x^{n-1}\in T_n} H(\Psi_n|x^{n-1}) \geq H - \sigma\sqrt{\epsilon_n} = H\left(1-2\left(\frac{\sigma^2}{4H\log n}\right)^{1/3}\right) \stackrel{\text{def}}{=} H(1-2\delta_n).$$

Note that $E_{X^{n-1}} p(I_n = 1|X^{n-1}) = p(I_n = 1) = \nu_n$, hence from Markov's inequality and Lemma 49,

$$p(T_n) \geq 1 - \frac{H}{\epsilon_n \log n} = 1 - \left(\frac{\sigma^2}{4H\log n}\right)^{1/3} = 1-\delta_n.$$

It follows that for $n \geq 2$,

$$H(\Psi_n|\Psi^{n-1}) \geq H(\Psi_n|X^{n-1}) \geq \sum_{x^{n-1}\in T_n} p(x^{n-1}) H(\Psi_n|x^{n-1}) \geq H(1-3\delta_n).$$

Hence

$$\frac{1}{n}H(\Psi^n) \geq \frac{n-1}{n}H - \frac{3}{n}H\sum_{i=2}^{n}\delta_i = H - \Theta(H\delta_n) = H - \Theta\left(\frac{\sigma^2 H^2}{\log n}\right)^{1/3}. \qquad \square$$

Lemma 55 implies that

$$H \leq \sigma,$$

hence the rate of convergence of pattern entropy can be bounded as follows.

Corollary 57. For all discrete *i.i.d.* distributions,

$$\rho_{X,n} \leq \mathcal{O}\left(\frac{\sigma^4}{\log n}\right)^{1/3}. \qquad \square$$

9.4 Asymptotic equipartition of patterns

Shannon [73] showed that strings generated by *i.i.d.* distributions over finite alphabets satisfy an asymptotic equipartition property. Chung [74] generalized this result to infinite alphabets. We prove an equivalent property for patterns of such strings. Specifically, we show that

$$\frac{1}{n}\log\frac{1}{p(\overline{\Psi})} \xrightarrow{\text{p}} \frac{1}{n}E\log\frac{1}{p(\overline{\Psi})}, \tag{9.11}$$

where the convergence is in probability, uniformly over all *i.i.d.* distributions, see Theorem 61. Since by definition,

$$\frac{1}{n}E\log\frac{1}{p(\overline{\Psi})} \ \rightarrow\ \mathcal{H}_\Psi,$$

we obtain

$$\frac{1}{n}\log\frac{1}{p(\overline{\Psi})} \xrightarrow{\text{p}} \mathcal{H}_\Psi,$$

though here, the results of the last section show that we cannot have uniform convergence over all *i.i.d.* distributions. To prove (9.11) we use *profiles* of patterns, defined next.

The *multiplicity* of $\psi \in \mathbb{Z}^+$ in a pattern $\overline{\psi}$ is

$$\mu_\psi \stackrel{\text{def}}{=} |\{1 \leq i \leq |\overline{\psi}| : \psi_i = \psi\}|,$$

the number of times ψ appears in $\overline{\psi}$. The *prevalence* of a multiplicity $\mu \in \mathbb{N}$ in $\overline{\psi}$ is

$$\varphi_\mu \stackrel{\text{def}}{=} |\{\psi : \mu_\psi = \mu\}|,$$

the number of symbols appearing μ times in $\overline{\psi}$. The *profile* of $\overline{\psi}$ is

$$\overline{\varphi} \stackrel{\text{def}}{=} (\varphi_1, \ldots, \varphi_{|\overline{\psi}|})$$

the vector of prevalences of all possible multiplicities for $1 \le \mu \le |\overline{\psi}|$. For example, the pattern $\psi = 12131$ has multiplicities $\mu_1 = 3$, $\mu_2 = \mu_3 = 1$, and $\mu_\psi = 0$ for all other $\psi \in \mathbb{Z}^+$. Hence its prevalences are $\varphi_1 = 2$, $\varphi_2 = 0$, $\varphi_3 = 1$, $\varphi_4 = \varphi_5 = 0$, and its profile is $\varphi(\psi) = (2, 0, 1, 0, 0)$.

If p is an *i.i.d.* distribution, then all length-n patterns $\overline{\psi}$ with profile φ, have the same probability,

$$p(\overline{\psi}) = \frac{p(\varphi)}{N(\overline{\varphi})},$$

where

$$N(\overline{\varphi}) = \frac{n!}{\prod_\mu \mu!^{\varphi_\mu} \varphi_\mu!}$$

is the number of patterns with profile φ. Therefore

$$\log \frac{1}{p(\overline{\psi})} = \log \frac{1}{p(\varphi)} + \log N(\overline{\varphi}).$$

Let $\overline{\Phi}$ denote the profile of a random sequence $\mathcal{X}$. The following bound by McDiarmid can be used to show that $\log N(\overline{\Phi})$ concentrates around its mean.

Lemma 58. [McDiarmid [75]] Let $\overline{X} = X_1, \ldots, X_n$ be independent random variables and let the function $f(x_1, \ldots, x_n)$ be such that any change in a single x_i changes $f(x_1, \ldots, x_n)$ by at most η. Then,

$$p\left\{ \left| f(\overline{X}) - Ef(\overline{X}) \right| > \eta \sqrt{\frac{n \ln \frac{2}{\delta}}{2}} \right\} < \delta. \qquad \square$$

Corollary 59. For all $\alpha > 0$,

$$p\Big\{ \left| \log N(\overline{\Phi}) - E \log N(\overline{\Phi}) \right| > 3n^{\frac{1+\alpha}{2}} \log n \Big\} < \frac{2}{e^{2n^\alpha}}.$$

Proof Let $f(x_1, \ldots, x_n) = \log N(\overline{\varphi})$. A change in x_i can change $\log \prod \varphi_\mu!$ by at most $2 \log n$, and $\log \prod \mu!^{\varphi_\mu}$ by at most $\log n$. The corollary follows by setting $\delta = \frac{2}{e^{2n^\alpha}}$ in Lemma 58. □

We now show that with high probability, the profile self-information deviates from its expectation by at most roughly $n^{\frac{1+\alpha}{2}} \log n$.

Lemma 60. For all $\alpha > 0$,

$$p\left\{\left|\log\frac{1}{p(\overline{\Phi})} - H(\overline{\Phi})\right| \ge \left(\pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{1+\alpha}{2}}\log n\right\} \le \frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}}\log n\right)}.$$

Proof Let $\rho(n)$ be the number of profiles of length-n patterns. Then the entropy of $\overline{\Phi}$ can be bounded by

$$E\log\frac{1}{p(\overline{\Phi})} = H(\overline{\Phi}) \le \log\rho(n) \le \left(\pi\sqrt{\frac{2}{3}}\log e\right)\sqrt{n}.$$

where the second inequality follows as $\rho(n)$ is, see *e.g.* [2], the number of integer partitions of n, which has been computed by Hardy and Ramanujan [50].

Let $\ell = \left(\pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{1+\alpha}{2}}\log n$. Since $\ell \ge H(\overline{\Phi})$,

$$\left|\log\frac{1}{p(\overline{\Phi})} - H(\overline{\Phi})\right| \ge \ell \Rightarrow \log\frac{1}{p(\overline{\Phi})} \ge \ell,$$

hence

$$p\left\{\left|\log\frac{1}{p(\overline{\Phi})} - H(\overline{\Phi})\right| \ge \ell\right\} \le p\left\{\log\frac{1}{p(\overline{\Phi})} \ge \ell\right\} \le \frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}}\log n\right)},$$

where the last inequality follows as the probability of any profile with self-information $\ge \ell$ is at most $2^{-\ell}$ and there can be at most $\rho(n) \le \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$ such profiles. □

Corollary 59 and Lemma 60 imply the asymptotic equipartition property. Note that the convergence bound is uniform for all *i.i.d.* distributions.

Theorem 61. For all $\delta > 0$,

$$p\left\{\frac{1}{n}\left|\log\frac{1}{p(\overline{\Psi})} - H(\overline{\Psi})\right| \ge \delta\right\} = \exp\left(-\Omega\left(\frac{n\delta^2}{\log^2 n}\right)\right).$$

Proof Observe that $H(\overline{\Psi}) = E \log \frac{1}{p(\overline{\Psi})}$, and that

$$\begin{aligned}
p\left\{\frac{1}{n}\left|\log \frac{1}{p(\overline{\Psi})} - E\log\frac{1}{p(\overline{\Psi})}\right| \le \delta\right\} \\
&\ge p\left\{\frac{1}{n}\left|\log N(\overline{\Phi}) - E\log N(\overline{\Phi})\right| + \frac{1}{n}\left|\log\frac{1}{p(\overline{\Phi})} - E\log\frac{1}{p(\overline{\Phi})}\right| \le \delta\right\} \\
&\ge p\left\{\left\{\frac{1}{n}\left|\log N(\overline{\Phi}) - E\log N(\overline{\Phi})\right| \le 3n^{\frac{\alpha-1}{2}}\log n\right\}\right. \\
&\qquad \left.\bigcap\left\{\frac{1}{n}\left|\log\frac{1}{p(\overline{\Phi})} - E\log\frac{1}{p(\overline{\Phi})}\right| \le \left(\pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{\alpha-1}{2}}\log n\right\}\right\} \\
&\ge 1 - \frac{2}{e^{2n^\alpha}} - \frac{\exp\left(\pi\sqrt{\frac{2n}{3}}\right)}{\exp\left(\pi\sqrt{\frac{2n}{3}}\, n^{\frac{\alpha}{2}}\log n\right)},
\end{aligned}$$

where for sufficiently large n, $0 < \alpha \le 1$, is the solution of

$$\left(3 + \pi\sqrt{\frac{2}{3}}\log e\right) n^{\frac{\alpha-1}{2}}\log n = \delta.$$

The last inequality follows from Lemmas 59 and 60. Clearly,

$$n^\alpha = \frac{n\,\delta^2}{\left(3+\pi\sqrt{\frac{2}{3}}\log e\right)^2 \log^2 n},$$

and the theorem follows by observing that the $2e^{-2n^\alpha}$ term dominates. □

Acknowlegdements

This chapter is adapted from A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, July 2006. The author was a primary researcher of the sections reproduced here. We thank Ian Abramson and Tsachy Weissman for helpful discussions, and the anonymous reviewers of the paper mentioned for helpful comments.

Chapter 10

Relative Redundancy

We consider a relative redundancy measure for universal compression proposed in [76]. We state the problem formally below, summarize the underlying motivation, and cite some similar measures that have been previously considered.

As before $\mathcal{A}$ is the alphabet, and $\mathcal{A}^*$ be the collection of all finite sequences of symbols from $\mathcal{A}$. It will be convenient to adopt a different definition of an estimator for this chapter. An *estimator* p over $\mathcal{A}^*$ is a mapping $p : \mathcal{A}^* \to [0, 1]$ such that for every $n \in \mathbb{N} \overset{\text{def}}{=} \{0, 1, \ldots\}$, the restriction of p to $\mathcal{A}^n$ is a distribution.

Let $\mathcal{P}$ be a collection of estimators on $\mathcal{A}^*$. The *minimum encoding length* of $\overline{x} \in \mathcal{A}^*$ with respect to $\mathcal{P}$ is

$$\hat{\ell}(\overline{x}) = -\log \max_{p \in \mathcal{P}} p(\overline{x}),$$

the *codelength* of $\overline{x}$ using the estimator in $\mathcal{P}$ assigning $\overline{x}$ the highest probability.

The *redundancy* of a *universal* estimator q in describing $\overline{x} \in \mathcal{A}^*$ is

$$r(\overline{x}) = \log \frac{1}{q(\overline{x})} - \hat{\ell}(\overline{x}),$$

the excess number of bits over $\overline{x}$'s minimum encoding length. The *standard redundancy* is usually considered as a function of the sequence length n. In particular for every n, the standard (worst-case) redundancy of length-n sequences is the highest redundancy of q among such sequences.

A simple, but important collection of estimators is $\mathcal{I}_m^*$, the collection of *i.i.d.* distributions over strings from an alphabet of size m. In this case, several researchers have shown [24, 25, 27, 28, 29, 30] that the standard redundancy of length n sequences for the best universal estimator grows as $(m-1)/2\log n + \Theta(1)$. This implies that as the sequence length increases, the redundancy, incurred since the distribution is unknown, is negligible compared to the sequence length.

Note however that sequences of length n may have vastly different minimum encoding lengths. Hence, in comparison with their encoding lengths, such a bound on redundancy may be good for some sequences while lax for others.

For example, it is easy to see that most n-bit *i.i.d.* sequences have a minimum encoding length of $n - O(\sqrt{n})$ [77], which is significantly higher than their standard $\frac{1}{2}\log n + \Theta(1)$ redundancy.

However, for some sequences the standard redundancy is not small compared to their encoding lengths. For example, the minimum encoding length of the sequence of n zeros is 0. A code with $\frac{1}{2}\log n$ redundancy may be inefficient in describing this particular sequence. Similarly, the MDL of the sequence $0\ldots001$, consisting of $n-1$ zeroes and a single 1 is $\log n + O(1)$. A code with $\frac{1}{2}\log n$ redundancy describes this sequence using 50% more bits than the minimum necessary. It follows that the $\frac{1}{2}\log n$ redundancy bound for all n-bit sequences, while tight for most, is lax for many.

The slack of uniform standard redundancy bounds for low-MDL sequences becomes more pronounced as the size and complexity of the collection of distributions grows. Large and complex collections may have larger redundancy which may therefore be significant compared to the encoding length of a larger number of sequences. For example, a low-order Markov chain, or one with constrained transitions, may have a modest encoding length compared to the standard redundancy of the class of all Markov chains.

We therefore consider *relative redundancy* [76], defined as the maximum redundancy over all sequences whose minimum encoding length at most ℓ. This measure is in the same flavor as, but different from [78] and also [26, 15].

For the collection $\mathcal{P}$ of estimators over $\mathcal{A}^*$, the *(worst case) relative redundancy* of an estimator $\overline{q}$ for sequences whose minimum encoding length at most ℓ is

$$\hat{R}_r(\mathcal{P}, q, \ell) \stackrel{\text{def}}{=} \max_{\overline{x}\in\mathcal{A}^*:\hat{\ell}(\overline{x})\leq\ell} r(\overline{x}),$$

the highest redundancy over all such sequences.

Relative redundancy therefore measures the number of extra bits as a function of the smallest number of bits necessary if the distribution is known in advance. We say an estimator has *diminishing* relative redundancy if its relative redundancy is $o(\ell)$.

10.1 Results

We first note that the relative redundancy increases with the alphabet size.

Theorem 62. For all alphabet sizes m and estimators q, as ℓ grows,

$$\hat{R}_r(\mathcal{I}_m^*, q, \ell) \geq \Omega(m \log \ell).$$

Furthermore, for all $\ell > 0$, $\hat{R}_r(\mathcal{I}_m^*, q, \ell)$ increases with m and as $m \to \infty$,

$$\hat{R}_r(\mathcal{I}_m^*, q, \ell) \to \infty \qquad \square$$

As before, let $\mathcal{I}_\Psi^n$ be the set of distributions induced on patterns by *i.i.d.* distributions on length-n strings, and let

$$\mathcal{I}^\Psi = \cup_{n \geq 1} \mathcal{I}_\Psi^n.$$

We show that the relative redundancy of compressing patterns of *i.i.d.* strings,

$$\hat{R}_r(\mathcal{I}^\Psi, q, \ell) \leq \mathcal{O}\left(\frac{\ell}{\sqrt{\log \ell - \log \log \ell}} \right).$$

Hence, redundancy of patterns is only a small fraction of their encoding length, and patterns can be encoded with *diminishing relative redundancy.* The rest of the paper proves the result above, and a slight modification of the proof yields a bound on all $\ell > 0$ (not just asymptotics). The proof is constructive, in Section 10.3, we describe the estimator that achieves diminishing relative redundancy.

These result is analogous to large alphabet and pattern compression results known *e.g.* [26, 23, 2, 1], for standard redundancy formulation.

10.2 Preliminaries

As before, let Φ^n be the set of all possible profiles of length-n patterns, let Φ_m^n be the set of all possible profiles of length-n, m-symbol patterns, and for all $\varphi \in \Phi^n$, let $\Psi_{\overline{\varphi}}$ the set of all length-n patterns with profile φ.

We require the following bounds repeatedly in the proofs to follow.

Lemma 63. For $n \geq 2$ and for all $1 \leq m \leq \frac{n}{2}$,

$$\frac{\sqrt{m}}{\log \binom{n}{m}} \leq \frac{1}{\sqrt{\log n}} = o(1),$$

and hence

$$\frac{\log m}{\log \binom{n}{m}} \leq \frac{2}{\sqrt{\log n}} = o(1).$$

Proof Both inequalities trivially hold for $m = 1$. Since

$$\log \binom{n}{m} \geq m \log \frac{n}{m},$$

it follows that

$$\frac{\sqrt{m}}{\log \binom{n}{m}} \leq \frac{\sqrt{m}}{m \log \frac{n}{m}}.$$

If $1 < m \leq \log n$, $\sqrt{m} > 1$ and $\log \frac{n}{m} \geq \log \frac{n}{\log n}$, hence

$$\frac{\sqrt{m}}{m \log \frac{n}{m}} = \frac{1}{\sqrt{m} \log \frac{n}{m}} < \frac{1}{\log \frac{n}{\log n}}$$

When $\log n < m \leq \frac{n}{2}$, $\sqrt{m} > \sqrt{\log n}$ and $\log \frac{n}{m} \geq 1$, hence

$$\frac{1}{\sqrt{m} \log \frac{n}{m}} < \frac{1}{\sqrt{\log n}}.$$

For all $n \geq 2$, $\log n - \log \log n \geq \sqrt{\log n}$, and hence the first inequality follows.

Since for all $m \geq 1$, $\log m \leq 2\sqrt{m}$,

$$\log m \leq \frac{2 \log \binom{n}{m}}{\sqrt{\log n}}. \qquad \square$$

Lemma 64. For $n > 4$, and for non-negative φ_μ (prevalences), positive $\mu \in \mathbb{Z}$ (multiplicities) and $1 \leq \mu \leq n$, such that

$$n = \sum_{\mu \geq 1} \mu \varphi_\mu \qquad \text{and} \qquad m = \sum_{\mu \geq 1} \varphi_\mu < n,$$

the following inequality holds,

$$\frac{n!}{\prod_\mu \mu!^{\varphi_\mu} \varphi_\mu!} \geq \binom{n}{m-1}.$$

Proof If $m = 1$, $\varphi_n = 1$ and $\varphi_\mu = 0$ for all $1 \le \mu < n$, therefore the lemma follows trivially.

We consider the case when $m > 1$. Observe that if more than one $\varphi_\mu > 0$, $\prod_\mu \varphi_\mu! \le (m-1)!$. To see this, note that there is at least one j such that $1 \le \varphi_j \le m-1$, and that

$$(m-1)! = (m-1)!1! \ge \varphi_j!(m-\varphi_j)!$$

while

$$(m-\varphi_j)! \ge \prod_{\substack{\mu \\ \mu \ne j}} \varphi_\mu!.$$

Next, we show that if $m < n$

$$\frac{(n-m+1)!}{\prod_\mu \mu!^{\varphi_\mu}} \ge 1.$$

Both the numerator and denominator comprise of $n-m$ numbers > 1. The result is evident by writing the equation above as

$$\frac{2 \cdot 3 \cdot 4 \cdot \ldots (n-m+1)}{2 \cdot \ldots \cdot \mu_1 \cdot 2 \cdot \ldots \cdot \mu_2 \ldots 2 \cdot \ldots \cdot \mu_{m-\varphi_1}},$$

where $\mu_1, \ldots, \mu_{m-\varphi_1}$ are the multiplicites that are > 1.

If $\varphi_{n/m} = m$, note that if $\frac{n}{m} > 3$ then $n - m > 2m > 2$ and

$$\frac{(n-m+1)!}{\prod_\mu \mu!^{\varphi_\mu}} = \frac{2 \cdot 3 \cdot 4 \cdot \ldots (n-m+1)}{2 \cdot \ldots \cdot \frac{n}{m} \cdot 2 \cdot \ldots \cdot \frac{n}{m} 2 \cdot \ldots \cdot \frac{n}{m}} \ge \frac{n-m}{n/m-1} = m,$$

and if $\frac{n}{m} = 2$, namely $m = n/2$, when $n > 4$,

$$\frac{2 \cdot 3 \cdot 4 \cdot \ldots (n/2+1)}{2^{n/2}} \ge n/2,$$

with equality when $n = 6$. □

Lemma 65. Positive integers $n \ge 1$ can be described in a prefix free manner using $2\lfloor \log n \rfloor + 1$ bits. □

10.3 Description of the encoding

We provide a constructive proof, namely, we describe an estimator q for patterns with relative redundancy

$$\hat{R}_r(\mathcal{I}^\Psi, q, \ell) = o(\ell).$$

The estimator q describes a pattern $\psi \in \Psi^n$ by describing

- φ: first describe the number m of distinct symbols in Ψ^n, followed by the profile $\overline{\varphi} \in \Phi_m^n$ with m symbols,
- ψ: the description of ψ using a uniform encoding over $\Psi_{\overline{\varphi}}$.

Let q use $\ell_q(\overline{\varphi})$ bits to describe the profile φ, and and $\ell_q(\psi|\overline{\varphi})$ bits to describe $\psi \in \Psi_{\overline{\varphi}}$.

We now specify how the profile is described. Observe that the number of distinct symbols,

$$m = \sum_{i=1}^{n} \varphi_\mu,$$

satisfies $m \leq n$. To describe m, q uses one bit to describe if $m > \frac{n}{2}$, or if $m \leq \frac{n}{2}$, followed by a prefix free description of m using from Lemma 65 $2 \log m + 1$ bits.

If $m \leq n/2$, we concatenate the prefix free descriptions of

$$\varphi_\mu, \mu : \varphi_\mu \in \overline{\varphi} \text{ and } 0 < \mu\varphi_\mu \leq \frac{n}{2},$$

ascending order of $\mu\varphi_\mu$, with a random ordering in case of ties, with the description of

$$\varphi_\mu : \varphi_\mu \in \overline{\varphi} \text{ and } \mu\varphi_\mu > \frac{n}{2},$$

An extra bit is used in both cases to specify if both φ_μ and μ will be described.

If $m > \frac{n}{2}$, we use the one-one correspondance between profiles of length-n m-symbol patterns and the profiles of length-$(n-m)$ patterns to describe the profile using a uniform encoding of profiles in Φ^{n-m}.

10.4 Proof outline

10.4.1 Preliminaries

We bound $\ell_q(\overline{\varphi})$, the number of bits the estimator q uses to describe the profile in Lemma 66, and $\ell_q(\psi|\overline{\varphi})$, the number of bits needed to specify the pattern given the profile in Lemma 67. Using these results, in Lemma 68, we bound the relative redundancy of q, and in Theorem 69, we show that q has diminishing relative redundancy.

Lemma 66. For any profile $\overline{\varphi} \in \Phi^n$,

$$\ell_q(\overline{\varphi}) \leq \begin{cases} 2\log m + 2 + \\ \quad \sum_{\substack{\mu:\varphi_\mu \neq 0 \\ \mu\varphi_\mu \leq \frac{n}{2}}} (2\log \mu\varphi_\mu + 2) \\ \quad + \sum_{\substack{\mu:\varphi_\mu \neq 0 \\ \mu\varphi_\mu > \frac{n}{2}}} (2\log \varphi_\mu + 1) & m \leq \frac{n}{2} \\ 2\log(n-m) + 2 + \\ \quad \pi\sqrt{\frac{2}{3}} \log e\sqrt{n-m} & m > \frac{n}{2}. \end{cases}$$

Proof We can describe any positive integer m in a prefix-free fashion with $2\log m + 1$ bits. Therefore, including an additional bit to describe if $m > \frac{n}{2}$ or not

$$2\log\min\{m, n-m\} + 1 + 1$$

bits are used to describe m. When $m \leq \frac{n}{2}$, for all i, φ_μ such that $0 < \mu\varphi_\mu \leq \frac{n}{2}$, q uses

$$2\log\mu\varphi_\mu + 2 + 1$$

bits to describe (μ, φ_μ), the extra bit needed to specify that both μ and φ_μ are being specified, and for $\mu\varphi_\mu > \frac{n}{2}$, q uses

$$2\log\varphi_\mu + 1 + 1$$

bits.

Recall from [2] that the profile of a pattern is equivalent to the partition of a positive integer. If $m > \frac{n}{2}$, the number of partitions of n into m parts is the number of partitions of $n - m$. Therefore [61],

$$\pi\sqrt{\frac{2}{3}} \log e\sqrt{n-m}$$

bits are used describe a profile $\varphi \in \Phi^n_m$ of a pattern with more than $\frac{n}{2}$ symbols. □

Lemma 67. For all length-n patterns with profile $\varphi = (\varphi_1, \ldots, \varphi_\mu, \ldots \varphi_n)$,

$$\ell_q(\psi|\overline{\varphi}) = \log \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!}. \qquad \square$$

10.4.2 Estimation of relative redundancy

In this section, we bound the relative redundancy of the estimator q. In Lemma 68, we relate the relative redundancy to the blocklength, and use it in Theorem 69 to show that q has diminishing relative redundancy.

The asymptotic notation used below implies the minimum codelength, $\hat{\ell}(\psi)$ increases to infinity. Note that this also implies that the length of the pattern $n(\psi) \to \infty$. For example, $f = \mathcal{O}(g)$ reads as "$\exists L$ and $C > 0$, such that for all patterns with $\hat{\ell}(\psi) \geq L$, $f \leq Cg$".

For convenience, we write n and m for $n(\psi)$ and $m(\psi)$ respectively.

Lemma 68. For all patterns $\psi \in \Psi^n$, let $\hat{\ell}(\psi)$ be the minimum codelength of ψ from sources in $\mathcal{I}_\Psi^n$. Then, for the estimator q described in Section 10.3,

$$\ell_q(\psi) \leq \hat{\ell}(\psi) + \mathcal{O}_\ell\left(\frac{\hat{\ell}(\psi)}{\sqrt{\log n}}\right).$$

Proof Observe that the shortest codelength [2] of a pattern assigned by a source in $\mathcal{I}_\Psi^n$, $\hat{\ell}$, is

$$\hat{\ell}(\psi) \geq \log \frac{n!}{\prod_{i=1}^{n} (i!)^{\varphi_\mu} \varphi_\mu!},$$

where $\varphi = (\varphi_1, \ldots, \varphi_n)$ is the profile of the pattern. Hence

$$\ell_q(\psi|\overline{\varphi}) < \hat{\ell}(\psi).$$

We show that

$$\ell_q(\overline{\varphi}) \leq \mathcal{O}\left(\frac{\ell_q(\psi|\overline{\varphi})}{\sqrt{\log n}}\right) \tag{10.1}$$

from which the lemma follows.

We prove Equation (10.1) for profiles of patterns with $1 < m \leq \frac{n}{2}$ and $\frac{n}{2} < m < n$ symbols separately.

Case 1: $\frac{n}{2} < m < n$

Observe that

$$\ell_q(\psi|\overline{\varphi}) = \log \frac{n!}{\prod_\mu \mu!^{\varphi_\mu} \varphi_\mu!} \geq \log \binom{n}{m-1}. \tag{10.2}$$

From Lemma 66 the number of bits needed to describe the profile is

$$\begin{aligned}\ell_q(\overline{\varphi}) &= \Theta\big(\sqrt{n-m} + 2\log(n-m)\big)\\ &\le \mathcal{O}\left(\frac{\binom{n}{m}}{\sqrt{\log n}}\right),\end{aligned}$$

from Lemma 63. Therefore, from Equation (10.2)

$$\ell_q(\overline{\varphi}) \le \mathcal{O}\left(\frac{\ell_q(\psi|\overline{\varphi})}{\sqrt{\log n}}\right).$$

Case 2: $1 < m \le \frac{n}{2}$

From Lemma 66,

$$\frac{\ell_q(\overline{\varphi})}{\ell_q(\psi|\overline{\varphi})} = \frac{2\log m + 2 + \sum_{\mu:\mu\varphi_\mu>0} \ell_q^{iv_i}(\overline{\varphi})}{\sum_{\mu:\mu\varphi_\mu>0} \ell_q^{iv_i}(\psi|\overline{\varphi})}$$

where

$$\ell_q^{iv_i}(\overline{\varphi}) = \begin{cases} 2\log\mu\varphi_\mu + 2 & 0 < \mu\varphi_\mu \le \frac{n}{2},\\ 2\log\varphi_\mu + 1 & \mu\varphi_\mu > \frac{n}{2},\end{cases}$$

and

$$\ell_q^{iv_i}(\psi|\overline{\varphi}) = \log\frac{n!^{\frac{\mu\varphi_\mu}{n}}}{\mu\varphi_\mu!} + \log\frac{\mu\varphi_\mu!}{\mu!^{\varphi_\mu}\varphi_\mu!}.$$

Noting that for μ such that $\mu\varphi_\mu > \frac{n}{2}$ and $\varphi_\mu = 1$,

$$\ell_q^{iv_i}(\overline{\varphi}) = 1,$$

we write,

$$\frac{\ell_q(\overline{\varphi})}{\ell_q(\psi|\overline{\varphi})} \le \frac{\substack{2\log m+2+1+\sum_{0<\mu\varphi_\mu\le\frac{n}{2}} \ell_q^{iv_i}(\overline{\varphi})\\ +\sum_{\substack{\mu\varphi_\mu>\frac{n}{2}\\ \varphi_\mu\neq 1}} \ell_q^{iv_i}(\overline{\varphi})}}{\substack{\frac{1}{2}\ell_q(\psi|\overline{\varphi})+\frac{1}{2}\left(\sum_{0<\mu\varphi_\mu\le\frac{n}{2}} \ell_q^{iv_i}(\psi|\overline{\varphi})\right.\\ \left.+\sum_{\substack{\mu\varphi_\mu>\frac{n}{2}\\ \varphi_\mu\neq 1}} \ell_q^{iv_i}(\psi|\overline{\varphi})\right)}}.$$

It can be seen that

$$\frac{2\log m + 2 + \sum_{\mu:\mu\varphi_\mu>0} \ell_q^{iv_i}(\overline{\varphi})}{\sum_{\mu:\mu\varphi_\mu>0} \ell_q^{iv_i}(\psi|\overline{\varphi})}$$
$$\leq 2\max\left\{\frac{2\log m+3}{\ell_q(\psi|\overline{\varphi})},\ \max_{\substack{\mu\varphi_\mu>\frac{n}{2}\\ \varphi_\mu\neq 1}} \frac{\ell_q^{iv_i}(\overline{\varphi})}{\ell_q^{iv_i}(\psi|\overline{\varphi})},\right.$$
$$\left.\max_{0<\mu\varphi_\mu\leq\frac{n}{2}} \frac{\ell_q^{iv_i}(\overline{\varphi})}{\ell_q^{iv_i}(\psi|\overline{\varphi})}\right\}$$

We show that

$$T_1 \overset{\text{def}}{=} \frac{2\log m+3}{\ell_q(\psi|\overline{\varphi})} \leq \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right),$$
$$T_2 \overset{\text{def}}{=} \max_{\substack{\mu\varphi_\mu>\frac{n}{2}\\ \varphi_\mu\neq 1}} \frac{\ell_q^{iv_i}(\overline{\varphi})}{\ell_q^{iv_i}(\psi|\overline{\varphi})} \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$
$$T_3 \overset{\text{def}}{=} \max_{0<\mu\varphi_\mu\leq\frac{n}{2}} \frac{\ell_q^{iv_i}(\overline{\varphi})}{\ell_q^{iv_i}(\psi|\overline{\varphi})} \leq \mathcal{O}\left(\frac{1}{\log n}\right),$$

thus proving (10.1).

Bound on T_1

We consider the bound on T_1 first. For all $m > 1$, and $\ell_q(\psi|\overline{\varphi})$ large enough that $n > 4$

$$\ell_q(\psi|\overline{\varphi}) = \log\frac{n!}{\prod_\mu \mu!^{\varphi_\mu}\varphi_\mu!} \geq \log\binom{n}{m-1},$$

and from Lemma 63,

$$\log m \leq \frac{\log\binom{n}{m-1}}{\sqrt{\log n}}.$$

Therefore for $\ell_q(\psi|\overline{\varphi})$ large enough that $n > 4$

$$\frac{2\log m+3}{\ell_q(\psi|\overline{\varphi})} \leq \frac{7}{\sqrt{\log n}}.$$

Bound on T_2

To prove the bound on T_2, it can be shown that

$$\ell_q^{iv_i}(\psi|\overline{\varphi}) \geq \log\frac{\mu\varphi_\mu!}{\mu!^{\varphi_\mu}\varphi_\mu!} \geq \log\left(\varphi_\mu!^{\mu-1}\right),$$

and, if $\mu\varphi_\mu > \frac{n}{2}$, and $\varphi_\mu > 1$,

$$\ell_q^{iv_i}(\psi|\overline{\varphi}) \geq \log\left(\varphi_\mu!^{i-1}\right) \geq \frac{\sqrt{n}}{4}\log n$$

bits are needed for describing the pattern while at most $2\log n + 2$ bits are needed to describe for describing (μ, φ_μ) in this case.

Bound on T_3

We now prove the bound on T_3. If $\mu\varphi_\mu \leq \frac{n}{2}$, using Feller's bounds on Stirling's approximation we obtain

$$\begin{aligned}
\frac{1}{T_3} &= \frac{\log \frac{n!^{\frac{\mu\varphi_\mu}{n}}}{\mu\varphi_\mu!} + \log \frac{\mu\varphi_\mu!}{\mu!^{\varphi_\mu}\varphi_\mu!}}{2\log \mu\varphi_\mu + 3} \\
&\geq \frac{\mu\varphi_\mu \log \frac{n}{\mu\varphi_\mu} - \frac{1}{2}\log \mu\varphi_\mu - \frac{1}{12\mu\varphi_\mu}\log e - \frac{1}{2}\log 2\pi}{2\log \mu\varphi_\mu + 3} \\
&= \frac{\mu\varphi_\mu \log \frac{n}{\mu\varphi_\mu}}{2\log \mu\varphi_\mu + 3} - \frac{\frac{1}{2}\log \mu\varphi_\mu}{2\log \mu\varphi_\mu + 3} - \frac{\frac{1}{12\mu\varphi_\mu}\log e}{2\log \mu\varphi_\mu + 3} \\
&\qquad - \frac{\frac{1}{2}\log 2\pi}{2\log \mu\varphi_\mu + 3} \\
&\geq \min_{1\leq x\leq \frac{n}{2}} g(x) - \frac{1}{4} - \frac{1}{36}\log e - \frac{1}{8}\log 2\pi
\end{aligned}$$

where

$$g(x) \overset{\text{def}}{=} \frac{x\log\frac{n}{x}}{2\log x + 3}$$

To see that

$$g(x) = \Omega(\log n)$$

for all $1 \leq x \leq \frac{n}{2}$, note that the derivative,

$$g' = -\frac{2y^2 - (2\log n - 3)y + 3 - \log n}{(2y+3)^2}.$$

where $y = \log x$ is a ratio of quadratic polynomials in $\log x$. The denominator is positive in the entire range $1 \leq x \leq \frac{n}{2}$ and the numerator has one root between

$$\log\frac{n}{2} - \frac{4}{2\log n - 1} \leq \log x \leq \log\frac{n}{2}$$

and another between

$$-\frac{1}{2} \leq \log x \leq -\frac{1}{2} + \frac{4}{2\log n - 1}.$$

The latter root is less than 1 for $n \geq 4$. Since the numerator g' is quadratic and the denominator is positive, we conclude that for $n \geq 4$, that is one *maximum* of x between 1 and $\frac{n}{2}$, namely g is unimodal.

Since at $x = 1$ and $x = \frac{n}{2}$,

$$g(x) = \begin{cases} \frac{1}{3}\log n & x = 1, \\ \frac{n/2}{2\log n + 3} & x = \frac{n}{2} \end{cases}$$

it follows that for all $1 \leq x \leq \frac{n}{2}$, and n large enough,

$$g(x) \geq \frac{1}{3}\log n.$$

Therefore,

$$T_3 \leq \mathcal{O}\left(\frac{1}{\log n}\right).$$

This concludes the proof of (10.1), thus proving the Theorem. □

We now show that the above result implies that the relative redundancy is asymptotically a negligible fraction of the MDL.

Theorem 69. As ℓ grows,

$$\hat{R}_r(\mathcal{I}^\Psi, q, \ell) \leq \mathcal{O}\left(\frac{\ell}{\sqrt{\log \ell - \log\log \ell}}\right) = o(\ell).$$

Proof Note that the right side increases to infinity, so we ignore the relative redundancy of patterns with short minimum encoding length. Observe that the longest minimum encoding length of any length n pattern is $\log n! \leq n \log n$. Hence, for a given length $\hat{\ell}$, the smallest blocklength $n(\hat{\ell})$ of any pattern with maximum likelihood codelength $\hat{\ell}$ should satisfy

$$n(\hat{\ell})\log n(\hat{\ell}) \geq \hat{\ell},$$

hence,

$$n(\hat{\ell}) \geq \frac{\hat{\ell}}{\log \hat{\ell}}.$$

Therefore, for sufficiently large ℓ,

$$\hat{R}_r(\mathcal{I}^\Psi, q, \ell) \leq \mathcal{O}\left(\frac{\ell}{\sqrt{\log \ell - \log\log \ell}}\right).$$

which grows $o(\ell)$. □

Note that the scheme q always uses 2 bits to code patterns with zero maximum likelihood codelength. It follows from the Lemma 68 that for long enough patterns, the relative redundancy is small. Furthermore, for all long enough, non-zero minimum encoding length patterns the redundancy of q is a negligible fraction of the encoding length.

It is possible to bound the relative redundancy for all values of $\ell \geq 0$, rather than an asymptotic result as described above. This proof has been omitted since it involves little more than keeping the constants in the proofs instead of covering them with the asymptotic notation.

Acknowlegdements

This chapter is adapted from A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy of large alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2006. The author was a primary researcher of the sections reproduced here.

Bibliography

[1] A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004.

[2] 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.

[3] N. Jevtić, A. Orlitsky, and N.P. Santhanam. A lower bound on compression of unknown alphabets. Theoretical Computer Science, Feb 2005.

[4] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427—431, October 17 2003. See also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003.

[5] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J.Zhang. On modeling profiles instead of values. In *Uncertainty in Artificial Intelligence*, 2004.

[6] A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy of large alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2006.

[7] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *Proceedings of the 44th Annual Symposium on Foundations of Computer Sciece*, October 2003.

[8] B. Fittingoff. Universal methods of coding for the case of unknown statistics. In *Proceedings of the 5th Symposium on Information Theory*, pages 129—135. Moscow-Gorky, 1972.

[9] L.D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783—795, November 1973.

[10] J. Shtarkov. Coding of discrete sources with unknown statistics. In I. Csiszár and P. Elias, editors, *Topics in Information Theory (Coll. Math. Soc. J. Bolyai, no. 16)*, pages 559—574. Amsterdam, The Netherlands: North Holland, 1977.

[11] J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674—682, November 1978.

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

[13] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124—2147, October 1998.

[14] L. Györfi, I. Pali, and E.C. Van der Meulen. On universal noiseless source coding for infinite source alphabets. *European Transactions on Telecommunications and Related Technologies*, 4:125—132, 1993.

[15] D.P. Foster, R.A. Stine, and A.J. Wyner. Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Transactions on Information Theory*, 48(6):1713—1720, June 2002.

[16] T. Uyematsu and F. Kanaya. Asymptotic optimality of two variations of Lempel-Ziv codes for sources with countably infinite alphabet. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

[17] P. Elias. Universal codeword sets and representations of integers. *IEEE Transactions on Information Theory*, 21(2):194—203, March 1975.

[18] D. He and E Yang. On the universality of grammar-based codes for sources with countably infinite alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2003.

[19] J. Åberg, Y.M. Shtarkov, and B.J.M. Smeets. Multialphabet coding with separate alphabet description. In *Proceedings of Compression and Complexity of Sequences*, 1997.

[20] N. Jevtić, A. Orlitsky, and N.P. Santhanam. Universal compression of unknown alphabets. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

[21] A. Orlitsky and N.P. Santhanam. Performance of universal codes over infinite alphabets. In *Proceedings of the Data Compression Conference*, March 2003.

[22] G. Shamir and L. Song. On the entropy of patterns of *i.i.d.* sequences. In *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, pages 160—170, October 2003.

[23] G. Shamir. Universal lossless compression with unknown alphabets—the average case. Submitted for publication, IEEE Transactions on Information Theory, 2003.

[24] R.E. Krichevsky and V.K. Trofimov. The preformance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199—207, March 1981.

[25] T.M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1—29, January 1991.

[26] Y.M. Shtarkov, T.J. Tjalkens, and F.M.J. Willems. Multialphabet universal coding of memoryless sources. *Problems of Information Transmission*, 31(2):114—127, 1995.

[27] T.M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348—363, March 1996.

[28] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40—47, January 1996.

[29] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34(2):142—146, 1998.

[30] Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431—445, March 2000.

[31] M. Drmota and W. Szpankowski. The precise minimax redundancy. In *Proceedings of IEEE Symposium on Information Theory*, 2002.

[32] N. Littlestone and M.K. Warmuth. The weighted majority algorithm. In *IEEE Symposium on Foundations of Computer Science*, 1992.

[33] V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153—173, 1998.

[34] N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 12—18, 1999.

[35] P.S. Laplace. *Philosphical essays on probabilities.* Springer Verlag, New York, Translated by A. Dale from the 5th (1825) edition, 1995.

[36] I.H. Witten and T.C. Bell. The zero frequency problem. *IEEE Transactions on Information Theory*, 37(4):1085—1094, 1991.

[37] B.S. Clarke and A.R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37—60, 1994.

[38] W. Gale and K. Church. What is wrong with adding one? In N. Oostdijk and P. de Haan, editors, *Corpus based research into language*, pages 189—198. Rodopi, Amsterdam, 1994.

[39] F.H. Hinsley and A. Stripp. *Codebreakers: The inside story of Bletchley Park.* Oxford University Press, 1993.

[40] I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237—264, December 1953.

[41] F. Song and W.B. Croft. A general language model for information retrieval (poster abstract). In *Research and Development in Information Retrieval*, pages 279—280, 1999.

[42] K.W. Church and W.A. Gale. Probability scoring for spelling correction. *Statistics and Computing*, 1:93—103, 1991.

[43] W.A. Gale, K.W. Church, and D. Yarowsky. A method for disambiguating word senses. *Computers and Humanities*, 26:415—419, 1993.

[44] S.F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310—318, San Francisco, 1996. Morgan Kaufmann Publishers.

[45] A. Nadas. On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33(6):1414—1416, December 1985.

[46] A. Nadas. Good, Jelinek, Mercer, and Robins on Turing's estimate of probabilities. *American Journal of Mathematical and Management Sciences*, 11:229—308, 1991.

[47] I.J. Good. Turing's anticipation of Empirical Bayes in connection with the cryptanalysis of the Naval Enigma. *Journal of Statistics Computation and Simulation*, 66:101—111, 2000.

[48] D. McAllester and R. Schapire. On the convergence rate of Good Turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[49] E. Drukh and Y. Mansour. Concentration bounds on unigrams language model. In *17th Annual Conference on Learning Theory*, pages 170—185, 2004.

[50] G.H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75—115, 1918.

[51] G. Shamir. Sequence patterns, entropy, and infinite alphabets. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, October 2004.

[52] G. Gemelos and T. Weissman. On the entropy rate of pattern processes. Technical Report HPL-2004-159, HP Labs, September 2004.

[53] G. Gemelos and T. Weissman. Submitted for publication, Data Compression Conference, November 2004.

[54] G. Shamir. Bounds on the entropy of patterns of iid sequences. In *IEEE Information Theory Workshop*, 2005.

[55] G. Gemelos and T. Weissman. On the relationship between process and pattern entropy rate. In *Proceedings of the IEEE Symposium on Information Theory*, 2005.

[56] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Innovation and pattern entropy of stationary processes. In *Proceedings of the IEEE Symposium on Information Theory*, 2005.

[57] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Information theoretic approach to modeling low probabilities. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2004.

[58] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. In *Information Theory Workshop*, 2004.

[59] Y.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3—17, 1987.

[60] W. Feller. *An introduction to probability theory.* Wiley, 1968.

[61] J.H. van Lint and R.M. Wilson. *A course in combinatorics.* Cambridge University Press, 1996.

[62] N.J.A. Sloane. Online encyclopedia of integer sequences. `http://www.research.att.com/~njas/sequences/`.

[63] G.H. Hardy and E.M. Wright. *An introduction to the theory of numbers.* Oxford University Press, 1985.

[64] I. Csiszár and P.C. Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information Theory*, 42(6):2065—2072, November 1996.

[65] A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50(10):2215—2230, October 2004.

[66] W.K. Hayman. A generalization of Stirling's formula. *Journal für die reine und angewandte Mathematik*, 196:67—95, 1956.

[67] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences.* Wiley, 2001.

[68] P. Flajolet and R. Sedgewick. Average case analysis of algorithms: Saddle point asymptotics. Technical Report 2376, INRIA, 1994.

[69] E.W. Weisstein. "Ratio Test." From Mathworld—A Wolfram Web Resource. `http://mathworld.wolfram.com/RatioTest.html`.

[70] E.W. Weisstein. "Root Test." From Mathworld—A Wolfram Web Resource. `http://mathworld.wolfram.com/RootTest.html`.

[71] R.P. Stanley. *Enumerative Combinatorics*. Cambridge University Press, 1 edition, 1997.

[72] A. Orlitsky and K. Viswanathan. One-way communication and error-correcting codes. *IEEE Transactions on Information Theory*, 49(7):1781—1788, July 2003.

[73] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379—423, 623—656, 1948.

[74] K.L. Chung. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612—614, 1961.

[75] C. McDiarmid. *Surveys in Combinatorics 1989*, chapter On the method of bounded differences, pages 148—188. Cambridge University Press, 1989.

[76] A. Orlitsky, N.P. Santhanam, and J. Zhang. Relative redundancy: A more stringent performance guarantee for universal coding. In *Proceedings of IEEE Symposium on Information Theory*, 2004.

[77] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and sons., 1991.

[78] N. Merhav and M. Feder. Hierarchical universal coding. *it*, 42(5):1354—1364, September 1996.