Always Good Turing: Asymptotically Optimal Probability Estimation

Alon Orlitsky[1,2] Narayana P. Santhanam[1] Junan Zhang[1]

[1]ECE [2]CSE Departments, UCSD

{alon,nsanthan,j6zhang}@ucsd.edu

Abstract

While deciphering the Enigma Code during World War II, I.J. Good and A.M. Turing considered the problem of estimating a probability distribution from a sample of data. They derived a surprising and unintuitive formula that has since been used in a variety of applications and studied by a number of researchers. Borrowing an information-theoretic and machine-learning framework, we define the attenuation of a probability estimator as the largest possible ratio between the per-symbol probability assigned to an arbitrarily-long sequence by any distribution, and the corresponding probability assigned by the estimator. We show that some common estimators have infinite attenuation and that the attenuation of the Good-Turing estimator is low, yet larger than one. We then derive an estimator whose attenuation is one, namely, as the length of any sequence increases, the per-symbol probability assigned by the estimator is as high as possible. Interestingly, some of the proofs use celebrated results by Hardy and Ramanujan on the number of partitions of an integer. To better understand the behavior of the estimator, we study the probability it assigns to several simple sequences. We show that for some sequences this probability agrees with our intuition, while for others it is rather unexpected.

1. Introduction

In preparation for your next safari, you observe a random sample of African animals. You find 3 giraffes, 1 zebra, and 2 elephants. How would you estimate the probability distribution 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 this estimate is clearly amiss as the poor estimator will be completely unprepared for an encounter with an offended lion.

To address this unseen-elements problem, Laplace [19] proposed an estimator that would add one to the count of each species, including to the collection of unseen ones. In the previous sample for example, the *Laplace*, or *add-one*, estimator assigns 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* [18], estimator was shown to possess certain optimality properties when the number of possible elements is fixed and the sample size increases to infinity [33, 5].

However, when the number of possible elements is large compared to the sample size, add-constant estimators are lacking too. To see that, suppose that during your safari trip you evaluate the distribution of animals' DNA sequences. You observe the DNA sequences of a large number n of animals and, predictably, find that each sequence is unique. 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 $(n + nc)/(n + nc + c)$ that the estimator assigns to all observed sequences is close to one, while the probability it assigns to all unseen sequences is close to zero. Clearly, the opposite better represents the truth. Additional shortcomings of add-constant estimators can be found in [10].

Supported by NSF Grant CCR-0313367.

I.J. Good and A.M. Turing encountered this problem while trying to break the Enigma Cipher [16] during World War II. British intelligence was in possession of the German cipher book which 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 knowledge to estimate the distributions of pages that each U-boat commander picked secret keys from.

Good and Turing came up with a surprising estimator, described in Section 3, that bears little resemblance to either the empirical-frequency or the add-constant estimators above. After the war, Good published the estimator [12] mentioning that Turing had an "intuitive demonstration" for it, but not describing what this intuition was.

Since its publication, the *Good-Turing estimator* has been incorporated into a variety of applications such as information retrieval [30], spelling correction [4], and word-sense disambiguation [11]. Perhaps its most common use is in language modeling for speech recognition, *e.g.*, [3] where it is applied to estimate the probability distribution of words.

While the Good-Turing estimator performs well in general, it is known to be suboptimal for elements that appear frequently. Consequently, several modifications have been proposed, including the Jelinek-Mercer, Katz, Witten-Bell and Kneser-Ney estimators [3]. In language modeling for example, Good Turing is usually used to estimate the probability of infrequent words, *e.g.*, hapax legomena, whereas the probability of frequent words is estimated via their empirical frequency.

On the theoretical side, interpretations of the Good-Turing estimator have been proposed [23, 24, 13], and its convergence rate was analyzed [21]. 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.

Borrowing an information-theoretic and machine-learning framework, we derive a natural measure for the performance of an estimator. Instead of using the estimator once, we apply it repeatedly to a sequence of elements all drawn according to the same distribution. Before each element is revealed, we use the estimator to evaluate its conditional probability given the previous elements. Doing so in turn for each element in the sequence, and multiplying the conditional probability estimates together, we obtain the probability that the estimator assigns to the whole sequence. We then compare that probability to the probability assigned to the sequence by any distribution, including the actual underlying one.

This measure is similar to one used to evaluate estimators of distributions over known and small alphabets [8, 28] and applied in a variety of fields including universal compression, *e.g.*, [29, 6, 7, 22], on-line algorithms, and learning, *e.g.*, [20, 31, 2].

To extend it to unknown, potentially large, infinite, and even continuous, alphabets, we abstract the actual symbols that appear in the sequence and consider only their *pattern*, the order in which they appear. This allows us to enumerate sequences over infinite alphabets, and to calculate their best probability assignment.

To describe the results obtained, we need a more detailed account of the quantities involved. We provide their informal description here, and formalize them in Section 2. The *sequence attenuation* of an estimator q for a sequence $\overline{x}$ is the ratio between the highest probability assigned to $\overline{x}$ by any distribution (including the one underlying the data) and the probability assigned to $\overline{x}$ by q. The *symbol attenuation* of q for $\overline{x}$ is the nth root of the sequence attenuation, namely, the ratio between the highest *per-symbol* probability assigned to the sequence by any distribution, and that assigned by q. Finally, the *(asymptotic, symbol) attenuation* of q is the highest symbol attenuation maximized over all sequences of increasing length.

Every estimator corresponds to a probability distribution over sequences of any given length. Hence the attenuation of any estimator is always at least one. Since the number of distinct symbols in a sequence can be as large as its length, the sequence attenuation of an estimator on a length-n sequence can $n!$ or even higher, hence its attenuation can be infinite.

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*, implies that the estimator assigns to each sequence a probability that is at most sub-exponentially smaller than the best possible, and hence the per-symbol probability assigned by the estimator is asymptotically the best possible.

The main objectives of this paper are to evaluate the attenuation of some existing estimators, to derive diminishing-attenuation estimators, and to establish bounds on the performance of any estimator.

In Section 3, we consider add-constant and Good-

Turing estimators. We show that add-constant estimators have infinite attenuation. We then analyze three versions of the Good-Turing estimator. We show that they perform well in the sense that their attenuation is low. However for some sequences they assign a probability that is exponentially smaller than the best possible, hence their attenuation is strictly above one.

In Section 4, we use the attenuation measure to 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 diminishes to one 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 diminishes to one at the faster rate of $2^{\mathcal{O}(n^{-1/2})}$.

All constants involved in the asymptotic terms are small. The techniques for evaluating 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 [14] on the number of partitions of an integer.

In Section 5 we evaluate the $2^{\mathcal{O}(n^{-1/3})}$ and $2^{\mathcal{O}(n^{-1/2})}$ rates at which the attenuations of the estimators approach one by upper bounding the rate at which the attenuation of any estimator can approach one. Converting the problem to a universal coding problem, and using Hayman's Theorem [15] as in [17, 25], and similar to [1], we show that the sequence attenuation of any estimator is at least $2^{\Omega(n^{1/3})}$, hence the rate in which the symbol attenuation decreases to one must be slower than $2^{\Omega(n^{-2/3})}$.

To better understand the behavior of diminishing-attenuation estimators we study the probability that the computationally-efficient estimator assigns to some simple sequences in Section 6. We 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.

2. Definitions and a preliminary result

We formally define the terms outlined in the introduction.

Estimators A *sample* is a sequence of elements. An estimator associates with every sample a probability distribution over the set of elements in the sample, and "new". For example, after observing the sample

giraffe, zebra, giraffe, elephant, elephant, giraffe,

an estimator postulates a distribution over the set {giraffe, zebra, 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. Since the sample space is not known in advance, the estimator cannot know which elements it hasn't yet seen, hence classifies all of them as "new". Observe also that if the sample space is known to the estimator in advance, then since all unseen elements are equivalent, an estimator that lumps them together can be easily converted to one that does not by assigning each unseen element the probability of "new" divided by the number of unseen elements.

Patterns Since 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, zebras 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 and call the *pattern* of the original sequence. The pattern of a sequence $\overline{x} = x_1, x_2, \ldots, x_n$ is denoted by $\Psi(\overline{x})$. For example, $\Psi(\text{g,z,g,e,e,g}) = 121331$.

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. Additionally we no longer need to refer to an element as "new" the first time it appears, and by its name thereafter. We always know the name of a new element in advance. It is one more than the number of elements hitherto seen.

A string of positive integers is the pattern of some sequence iff the first appearance of any $i \geq 2$ occurs after that of $i-1$. For example, the empty string Λ and the strings 1, 12, and 121 are patterns (of the empty string, and, say, "a", "ad", and "ada", respectively), while 2, 21, and 132 are not.

We let Ψ^n denote the set of length-n patterns, and let Ψ^* denote the set of all finite-length patterns. For example, $\Psi^0 = \{\Lambda\}$, $\Psi^1 = \{1\}$, $\Psi^2 = \{11, 12\}$, $\Psi^3 = \{111, 112, 121, 122, 123\}$, and so on, and $\Psi^* = \{\Lambda, 1, 11, 12, 111, 112, 121, 122, 123, \ldots\}$. It can be shown that every length-n pattern corresponds to a partition of a set of cardinality n, hence $|\Psi^n|$ is the n'th Bell number.

Probability of patterns If $\mathcal{A}$ is an alphabet, we let $\mathcal{A}^n$ denote the set of length-n sequences of elements in $\mathcal{A}$ and let $\mathcal{A}^*$ denote the set of finite strings of elements in $\mathcal{A}$. For example, $\{a,b\}^2 = \{aa, ab, ba, bb\}$. Let p be a probability distribution over an alphabet $\mathcal{A}$. For every $n \in \mathbb{Z}^+$, p induces a probability distribution p^Ψ over Ψ^n where

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

denotes the probability that a sequence of elements, each selected according to p will form the pattern $\overline{\psi} \in \Psi^n$. For example, for any probability p over an alphabet $\mathcal{A}$, $p^\Psi(1) = p(\mathcal{A}) = 1$, indicating that the first element of any pattern is 1 ("new"). If p is a distribution over $\{a, b\}$ where $p(a) = p$ and $p(b) = 1 - p \stackrel{\text{def}}{=} \overline{p}$, then $p^\Psi(11) = p\{aa, bb\} = p^2 + \overline{p}^2$, the probability that two elements will be identical, and $p^\Psi(12) = p\{ab, ba\} = 2p\overline{p}$, the probability that the two elements will be distinct.

Continuous distributions induce probabilities over patterns as well. For example, if p is any continuous distribution, then for all n, $p^\Psi(12\ldots n) = p\{x_1 \ldots x_n : x_i \neq x_j\} = 1$, indicating that if we pick any number of elements according to a continuous distribution, with probability 1, they will all be distinct. It follows that for continuous distributions, every $\overline{\psi} \in \Psi^n - \{12\ldots n\}$, namely every pattern with repetitions, has $p^\Psi(\overline{\psi}) = 0$, corresponding to the fact that elements repeat with probability 0.

Maximum pattern probability Our goal is to derive an estimator that, though unaware of the underlying probability p, assigns to every pattern $\overline{\psi}$ a probability that is not much smaller than the induced probability $p^\Psi(\overline{\psi})$. Since we do not know the underlying distribution, we must consider the one that assigns to $\overline{\psi}$ the highest probability. The *maximum probability* of a pattern $\overline{\psi}$ is

$$\hat{p}^\Psi(\overline{\psi}) \stackrel{\text{def}}{=} \max_p p^\Psi(\overline{\psi}),$$

the highest probability assigned to the pattern by any distribution. For example, since any distribution p has $p^\Psi(1) = 1$, we have $\hat{p}^\Psi(1) = 1$. Since any constant distribution p has $p^\Psi(1\ldots1) = 1$ for any number of 1's, we obtain $\hat{p}^\Psi(1\ldots1) = 1$, and, since any continuous distribution p has $p^\Psi(12\ldots n) = 1$, we derive $\hat{p}^\Psi(12\ldots n) = 1$. In general however, it is difficult to determine the maximum probability of a pattern. For example, some work [27] is needed to show that $\hat{p}^\Psi(112) = 1/4$.

Sequential estimators Let $m \stackrel{\text{def}}{=} m(\psi_1^n) \stackrel{\text{def}}{=} |\{\psi_1, \ldots, \psi_n\}|$ be the number of distinct symbols appearing in a pattern $\psi_1^n = \psi_1, \ldots, \psi_n \in \Psi^n$. A *sequential estimator* is a mapping q that associates with every pattern ψ_1^n a probability distribution $q(\psi_{n+1}|\psi_1^n)$ over $[m+1] = \{1, \ldots, m+1\}$, representing the probability that the estimator assigns to the possible values of ψ_{n+1}, after seeing ψ_1^n. For example, $q(\psi_1) \stackrel{\text{def}}{=} q(\psi_1|\Lambda)$ is a distribution over $\{1\}$, namely, $q(1|\Lambda) = 1$, while $q(\psi_3|12)$ and $q(\psi_4|121)$ are distributions over $\{1, 2, 3\}$.

For a simple example, consider the add-one estimator mentioned in the introduction, henceforth denoted q_{+1}. After observing the pattern ψ_1^n, it assigns to any $\psi_{n+1} \in [m+1]$ a probability proportional to one more than the number of times ψ_{n+1} appeared in ψ_1^n. For instance, 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

$$q(\psi_1^n) = \prod_{i=0}^{n-1} q(\psi_{i+1}|\psi_1^i).$$

For example, the probability that the add-one estimator ascribes to the pattern 1213 is

$$\begin{aligned} q_{+1}(1213) &= q_{+1}(1|\Lambda) \cdot q_{+1}(2|1) \cdot q_{+1}(1|12) \cdot q_{+1}(3|121) \\ &= \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6} = \frac{1}{45}. \end{aligned}$$

It is important to note that while for mathematical convenience estimators are defined in terms of patterns, they associate probabilities with the original sequences themselves. For example, for the sequence giraffe, zebra, giraffe, elephant, abbreviated g,z,g,e, the add-one estimator will associate the probability

$$q_{+1}(\text{new}) \cdot q_{+1}(\text{new}|\text{g}) \cdot q_{+1}(\text{g}|\text{g,z}) \cdot q_{+1}(\text{new}|\text{g,z,g}) = \frac{1}{1}\cdot\frac{1}{3}\cdot\frac{2}{5}\cdot\frac{1}{6} = \frac{1}{45},$$

the same probability it associates with its pattern 1213. Similarly, all our results are phrased in terms of patterns, but apply to the actual underlying sequences.

Attenuation We would like to derive sequential estimators that assign to every pattern 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 pattern ψ_1^n to be

$$R(q, \psi_1^n) \overset{\text{def}}{=} \frac{\hat{p}^{\Psi}(\psi_1^n)}{q(\psi_1^n)},$$

the ratio between the highest probability assigned to ψ_1^n by any distribution and the probability assigned to it by q. The *worst-case sequence attenuation* of q for length-n patterns is

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

the largest sequence attenuation of q for any length-n pattern. Note that $(R^n(q))^{1/n}$ is the *worst-case symbol attenuation* of q for length-n patterns, 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

$$R^*(q) \overset{\text{def}}{=} \limsup_{n\to\infty} (R^n(q))^{1/n},$$

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

As mentioned in the introduction, $R^*(q) \geq 1$ for every estimator q. If $R^*(q) > 1$ then q assigns to some length-n pattern a probability that is $(R^*(q))^n$ times smaller than its highest possible probability, and if $R^*(q) = 1$ then the probability that q assigns to every length-n pattern is at most subexponentially smaller than the highest possible.

A preliminary result In this paper we evaluate the attenuation of several existing and new estimators. Some of the proofs use the following upper bound on maximum pattern probabilities.

Let $\psi_1^n = \psi_1, \ldots, \psi_n$ be a pattern. The *multiplicity* of $\psi \in \mathbb{Z}^+$ in ψ_1^n is

$$\mu_\psi \overset{\text{def}}{=} \mu_\psi(\psi_1^n) \overset{\text{def}}{=} |\{1 \leq i \leq n : \psi_i = \psi\}|,$$

the number of times ψ appears in ψ_1^n. The *prevalence* of the multiplicity $\mu \in \mathbb{N}$ in $\overline{\psi}$ is

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

the number of symbols appearing μ times in ψ_1^n. For example, in the pattern 1213, the number 1 appears twice, hence it has multiplicity $\mu_1 = 2$, and each of the numbers 2 and 3 appears once, hence $\mu_2 = \mu_3 = 1$. It follows that two symbols appear once and one symbol appears twice, hence $\varphi_1 = 2$, $\varphi_2 = 1$, and $\varphi_\mu = 0$ for all other μ's.

The number of patterns with prevalences $\varphi_1, \varphi_2, \ldots, \varphi_n$ can be shown to be

$$\frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!} \overset{\text{def}}{=} N(\varphi_1, \ldots, \varphi_n).$$

Since any distribution assigns to each of them the same probability the maximum probability of each of these patterns is upper bounded by

$$\hat{p}^{\Psi}(\overline{\psi}) \leq \frac{1}{N(\varphi_1, \ldots, \varphi_n)}. \tag{1}$$

3. Unbounded- and constant-attenuation estimators

We show that add-constant estimators have unbounded attenuation and that a modified version of the add-one estimator and the Good-Turing estimator have constant, albeit non-diminishing, attenuations.

3.1. 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 q_{+1}. To the pattern $123\ldots n$ it assigns probability

$$q_{+1}(123\ldots n) = \frac{1}{1}\cdot\frac{1}{3}\cdot\ldots\cdot\frac{1}{2n+1} = \frac{2^n \cdot n!}{(2n+1)!}.$$

Since, as we saw in the introduction,

$$\hat{p}^{\Psi}(12\ldots n) = 1,$$

we obtain that q_{+1} has symbol attenuation

$$\left(R^n(q_{+1})\right)^{1/n} \geq \frac{(2n+1)!}{2^n \cdot n!} \geq \frac{2n}{e},$$

hence

Theorem 1

$$R^*(q_{+1}) = \infty. \qquad \square$$

By applying the add-one estimator in two steps, we obtain a *modified add-one* estimator $q_{+1'}$ whose attenuation is between 1.69 and 2.85. The estimator uses the add-one rule to estimate the probability of the next symbol being new or repeated, and for repeated symbols it assigns a probability proportional to the number of occurrences of the symbol. Recall that m is the number of distinct symbols appearing in a pattern ψ_1^n, and that for $1 \le \psi \le m$, μ_ψ is the multiplicity of ψ in ψ_1^n. Then $q_{+1'}$ assigns to each $1 \le \psi_{n+1} \le m+1$ the probability

$$q_{+1'}(\psi_{n+1}|\psi_1^n) \stackrel{\text{def}}{=} \begin{cases} \frac{m+1}{n+2}, & \psi_{n+1} = m+1 \\ \frac{n-m+1}{n+2} \cdot \frac{\mu_{\psi_{n+1}}}{n}, & 1 \le \psi_{n+1} \le m. \end{cases}$$

It can be shown that for sequences where the number of distinct symbols is a vanishing fraction of the sequence length, namely, $m = o(n)$, the modified add-one estimator has subexponential sequence attenuation, hence diminishing symbol attenuation. However, as illustrated below, sequences with more symbols may have an exponential sequence attenuation.

Theorem 2

$$1.69 < R^*(q_{+1'}) \le 2.85.$$

Proof To lower bound $R^*(q_{+1'})$, consider the pattern

$$\overline{\psi} \stackrel{\text{def}}{=} 12\ldots\frac{n}{2}12\ldots\frac{n}{2},$$

to which the modified add-one estimator assigns probability

$$q_{+1'}(\overline{\psi}) = \frac{\left(\frac{n}{2}!\right)^2\left(\frac{n}{2}-1\right)!}{(n+1)!(n-1)!} \approx 0.58^n n^{-n/2},$$

while the uniform distribution over an alphabet of size $0.628n$ assigns to $\overline{\psi}$ the probability $0.98^n n^{-n/2}$. Therefore,

$$R^*(q_{+1'}) \ge \frac{0.98}{0.58} > 1.69.$$

To upper bound $R^*(q_{+1'})$, we compare the probability that $q_{+1'}$ assigns to any pattern $\overline{\psi}$ with the upper bound on $\hat{p}^\Psi(\overline{\psi})$ given in Equation (1). We show that the sequence attenuation of any length-n pattern $\overline{\psi}$ with m distinct symbols is bounded by

$$R(q_{+1'}, \overline{\psi}) \le 2^{nh\left(\frac{m}{n}\right) + m\log\frac{n}{m} + o(1)}.$$

The theorem follows by maximizing this expression over m. $\square$

3.2. The Good-Turing estimator

We show that the attenuation of the Good-Turing estimator is a constant between 1.39 and 2.

Recall that μ_ψ is the multiplicity of $\psi \in \mathbb{Z}^+$ in $\psi_1^n \in \Psi^n$, *i.e.*, the number of times ψ appears in ψ_1^n, and that φ_μ is the prevalence of μ, $\mu \in \mathbb{N}$, *i.e.*, 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 [12] is then defined by

$$q(\psi_{n+1}|\psi_1^n) = \begin{cases} \frac{\varphi'_1}{n}, & r = 0 \\ \frac{r+1}{n}\frac{\varphi'_{r+1}}{\varphi'_r}, & r \ge 1, \end{cases}$$

where φ'_μ is a smoothed value of φ_μ. As observed already by Good [12], smoothing is needed for a variety of reasons. One of them is that if $\varphi_{r+1}(\psi_1^n) = 0$, 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. Here we consider only one of the simplest smoothing technique,

$$\varphi'_\mu = \max(\varphi_\mu, 1),$$

which ensures nonzero probabilities for all symbols in $[1, m(\psi_1^n)+1]$. This smoothing method results in the estimator

$$q_{\text{GT1}}(\psi_{n+1}|\psi_1^n) \stackrel{\text{def}}{=} \begin{cases} \frac{\max(\varphi_1,1)}{S_{\text{GT1}}(\psi_1^n)}, & r = 0 \\ \frac{r+1}{S_{\text{GT1}}(\psi_1^n)}\frac{\max(\varphi_{r+1},1)}{\varphi_r}, & r \ge 1, \end{cases}$$

where

$$S_{\text{GT1}}(\psi_1^n) \stackrel{\text{def}}{=} \max(\varphi_1, 1) + \sum_{\mu:\varphi_\mu>0} \varphi_\mu \cdot (\mu+1)\frac{\max(\varphi_{\mu+1}, 1)}{\varphi_\mu}$$

is a normalization factor.

The attenuation of q_{GT1} is a constant greater than one, as outlined below.

Theorem 3

$$1.39 < R^*(q_{\text{GT1}}) \le 2.$$

Proof outline To prove the lower bound, consider the pattern

$$\overline{\psi} \stackrel{\text{def}}{=} 12(132)^{n/3} \stackrel{\text{def}}{=} 12132132\ldots132,$$

to which the estimator q_{GT1} assigns probability

$$q_{\text{GT1}}(\overline{\psi}) = \Theta(72^{-n/3}),$$

while the maximum probability of $\overline{\psi}$ can be shown to be

$$\hat{p}^{\Psi}(\overline{\psi}) = \Theta(3^{-n}).$$

Hence,

$$R^*(q_{\text{GT1}}) \ge \frac{72^{\frac{1}{3}}}{3} > 1.39.$$

To upper bound $R^n(q_{\text{GT1}})$, let

$$r(i) \overset{\text{def}}{=} \mu_{\psi_{i+1}}(\psi_1^i),$$

be the multiplicity of ψ_{i+1} in ψ_1^i, and for $1 \le \mu \le i$, let

$$\varphi_\mu^i \overset{\text{def}}{=} \varphi_\mu(\psi_1^i)$$

be the prevalence of the multiplicity μ in ψ_1^i. It can be shown by induction on n that

$$q_{\text{GT1}}(\psi_1^n) = \frac{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu}}{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)} \cdot \prod_{i=1}^{n-1} \frac{\max(\varphi^i_{r(i)+1}, 1)}{\varphi^i_{r(i)}}.$$

Comparing the probability that q_{GT1} assigns to any pattern $\overline{\psi}$ with the upper bound on $\hat{p}^{\Psi}(\overline{\psi})$ given in Equation (1), we derive

$$\begin{aligned} R^n(q_{\text{GT1}}) \le & \max_{\psi_1^n} \frac{\prod_{\mu=1}^{n} \varphi_\mu^n!}{\prod_{i=1}^{n-1} \max(\varphi^i_{r(i)+1}, 1)/\varphi^i_{r(i)}} \times \\ & \max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{\text{GT1}}(\psi_1^i)}{n!} \\ \overset{\text{def}}{=} & R_G^n \cdot R_S^n. \end{aligned}$$

Observing that

$$\prod_{i=1}^{n-1} \frac{\varphi^i_{r(i)+1} + 1}{\varphi^i_{r(i)}} = \prod_{\mu=1}^{n} \varphi_\mu^n!,$$

we obtain

$$R_G^n \le 2^{n-1}. \qquad (2)$$

It can be shown that for all ψ_1^n,

$$S_{\text{GT1}}(\psi_1^n) \le n + \sqrt{8n},$$

hence,

$$R_S^n \le \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)^{n-1} \cdot \frac{1}{n}. \qquad (3)$$

The Theorem follows by multiplying Equations (2) and (3). □

We also considered two other variants of Good-Turing estimators. The first is the "simple Good-Turing" estimator developed by Gale [9]. The second, motivated by the fact that Good-Turing estimates perform well for infrequent symbols but not for frequent ones, uses the Good-Turing estimator to predict the probability of infrequent symbols and uses empirical frequency for frequent ones.

These two estimators seem too complex to analyze mathematically, but we showed empirically that while they perform well in general, for for some sequences, their attenuation does not approach one as the sequence length increases. Therefore none of the Good-Turing estimators we considered had diminishing attenuation.

4. Diminishing-attenuation estimators

We describe two diminishing-attenuation estimators. The first is computationally more efficient, while the second's attenuation diminishes to one faster.

4.1. A low complexity estimator

We construct a diminishing-attenuation estimator whose sequence attenuation is at most $2^{\mathcal{O}(n^{2/3})}$, hence its symbol attenuation diminishes to one at a rate of at least $2^{\mathcal{O}(n^{-1/3})}$. The estimator uses just a constant number of operations per symbol, hence has linear complexity for the whole sequence.

Recall that μ_ψ is the multiplicity of ψ, that φ_μ is the prevalence of μ, and that $r \overset{\text{def}}{=} \mu_{\psi_{n+1}}(\psi_1^n)$. For $c \in \mathbb{Z}^+$, let

$$f_c(\varphi) \overset{\text{def}}{=} \max(\varphi, c),$$

and for $\varphi \in \mathbb{N}$, let

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

Define also the sequence

$$c_n = \lceil n^{1/3} \rceil.$$

The estimator assigns

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

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

$$q_{1/3}(\psi_{n+1}|\psi_1^n) = \frac{1}{S_{c_{n+1}}(\psi_1^n)} \times \begin{cases} f_{c_{n+1}}(\varphi_1 + 1), & r = 0 \\ (r+1)\frac{f_{c_{n+1}}(\varphi_{r+1}+1)}{f_{c_{n+1}}(\varphi_r)}, & r > 0, \end{cases}$$

where

$$S_{c_{n+1}}(\psi_1^n) \overset{\text{def}}{=} f_{c_{n+1}}(\varphi_1 + 1) + \sum_{\mu=1}^{n} \varphi_\mu \cdot (\mu + 1) \frac{f_{c_{n+1}}(\varphi_{\mu+1} + 1)}{f_{c_{n+1}}(\varphi_\mu)}$$

is a normalization factor.

The estimator has an attenuation that is asymptotically unity.

Theorem 4 For all n,

$$R^n(q_{1/3}) \le 2^{\mathcal{O}(n^{2/3})},$$

where the implied constant is at most 10.

Proof outline It can be shown that for all $n \ge 2$ and $\psi_1^n \in \Psi^n$,

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

where as before, for $1 \le \mu \le i$, we let $\varphi_\mu^i = \varphi_\mu(\psi_1^i)$. Again, the upper bound is obtained by comparing the probability that $q_{1/3}$ assigns to a pattern with the upper bound on the maximum probability of the pattern given in Equation (1), yielding

$$\begin{aligned} R^n(q_{1/3}) \le & \max_{\psi_1^n} \prod_{\mu=1}^{n} \frac{\varphi_\mu^n!}{g_{c_n}(\varphi_\mu^n)} \times \\ & \max_{\psi_1^n} \frac{\prod_{i=1}^{n-1} S_{c_i}(\psi_1^i)}{n!} \times \\ & \max_{\psi_1^n} \prod_{i=1}^{n-1} \prod_{\mu=1}^{i} \frac{g_{c_{i+1}}(\varphi_\mu^i)}{g_{c_i}(\varphi_\mu^i)} \\ & \overset{\text{def}}{=} R_G^n \cdot R_S^n \cdot R_L^n. \end{aligned} \tag{4}$$

We bound each of R_G^n, R_S^n, and R_L^n individually. Observing that for all $c \in \mathbb{Z}^+$ and $\varphi \in \mathbb{N}$, $g_c(\varphi) \ge \varphi!$, we obtain

$$R_G^n \le 1. \tag{5}$$

It can be shown [26] that for all $\psi_1^n \in \Psi^n$ and all $\gamma \in \mathbb{Z}^+$,

$$S_\gamma(\psi_1^n) \le (1 + \frac{1}{\gamma})n + \sqrt{\frac{2n(2\gamma+1)^2}{\gamma}},$$

implying,

$$R_S^n \le \frac{1}{n} \left(\frac{\sum_{i=1}^{n-1} \left(1 + \frac{1}{c_{i+1}} + \sqrt{\frac{2(2c_{i+1}+1)^2}{ic_{i+1}}} \right)}{n-1} \right)^{n-1}. \tag{6}$$

It can also be shown that

$$R_L^n \le \prod_{i=1}^{n-1} \left(\frac{c_{i+1}}{c_i} \right)^{\sqrt{2ic_{i+1}}}. \tag{7}$$

The Theorem follows by multiplying Equations (5,6,7), and setting $c_n = \lceil n^{1/3} \rceil$. □

It can also be shown that the estimator requires only a constant number of calculations per symbol, hence has linear complexity for the whole sequence.

4.2. A low attenuation estimator

Building on an equivalence between set partitions and patterns, we obtain an estimator $q_{1/2}$ achieving a sequence attenuation of $2^{\mathcal{O}(n^{1/2})}$, hence a symbol attenuation that diminishes to one at a rate of at least $2^{\mathcal{O}(n^{-1/2})}$. However, the estimator has super polynomial, albeit subexponential, complexity.

To describe $q_{1/2}$, we need some additional definitions. For a pattern ψ_1^n, let

$$z(\psi_1^n) \overset{\text{def}}{=} \frac{\prod_{\mu=1}^n \mu!^{\varphi_\mu} \varphi_\mu!}{n!},$$

and define the distribution $\tilde{p}$ over Ψ^n by

$$\tilde{p}(\psi_1^n) = \frac{z(\psi_1^n)}{\sum_{\overline{y} \in \Psi^n} z(\overline{y})},$$

let

$$t_n \overset{\text{def}}{=} 2^{\lceil \log n + 1 \rceil - 1}$$

be the largest power of 2 that is $\le n$, and finally let

$$\Psi^{2t_n}(\psi_1^n) \overset{\text{def}}{=} \{ y_1^{2t_n} \in \Psi^{2t_n} : y_1^n = \psi_1^n \}$$

be the set of patterns of length $2t_n$ with prefix ψ_1^n. Then the estimator $q_{1/2}$ is defined by

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

and for all $n \ge 1$ and $\psi_1^n \in \Psi^n$,

$$q_{1/2}(\psi_{n+1} | \psi_1^n) = \frac{\sum_{\overline{y} \in \Psi^{2t_n}(\psi_1^n . \psi_{n+1})} \tilde{p}(\overline{y})}{\sum_{\overline{y} \in \Psi^{2t_n}(\psi_1^n)} \tilde{p}(\overline{y})}.$$

While $q_{1/2}$ is computationally complex, it achieves a lower attenuation. To upper bound its attenuation, we related it to the number of partitions of an integer. A *partition* of an integer n is a sum $a_1 + \ldots + a_k$, where $a_1 \ge a_2 \ge \ldots \ge a_k$ are positive integers. For example, 4, 3+1, 2+2, 2+1+1, and 1+1+1+1 are the five possible partitions of 4.

In what is considered by some to be “one of the jewels of 20th century mathematics” [32], Hardy and Ramanujan [14] gave an expression for the exact number of partitions of any positive integer n, and showed that it grows as $\exp(\pi\sqrt{2/3}\sqrt{n}(1 + o(1)))$. We use this result to show that

Theorem 5 For all n,

$$R^n(q_{1/2}) \le \exp\left(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})} \sqrt{n} \right).$$

Proof outline The theorem holds trivially for $n = 1$. For all $n \ge 2$ and ψ_1^n, we show by induction that

$$q_{1/2}(\psi_1^n) \ge \frac{\sum_{\overline{y} \in \Psi^{2t_n}(\psi_1^n)} \tilde{p}(\overline{y})}{\exp\left(\pi \sqrt{\frac{2}{3}} \sqrt{2} \frac{\sqrt{t_n} - 1}{\sqrt{2} - 1} \right)}. \tag{8}$$

To relate the numerator to $\hat{p}^{\Psi}(\psi_1^n)$ we use results on integer partitions by Hardy and Ramanujan [14]. We thus obtain

$$\frac{\sum_{\overline{y}\in\Psi^{2t_n}(\psi_1^n)} \tilde{p}(\overline{y})}{\hat{p}^{\psi_1^n}} \geq \frac{1}{\exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{2t_n}\right)},$$

which, together with Equation (8) implies that for all $n \geq 2$ and ψ_1^n,

$$\frac{\hat{p}^{\psi_1^n}(\psi_1^n)}{q_{1/2}(\psi_1^n)} \leq \exp\left(\frac{4\pi}{\sqrt{3}(2-\sqrt{2})}\sqrt{n}\right). \qquad \square$$

5. A lower bound on attenuation

In the previous section we derived estimators with sequence attenuations of $2^{\mathcal{O}(n^{2/3})}$ and $2^{\mathcal{O}(n^{1/2})}$. We now show that attenuation cannot be made arbitrarily small. The sequence attenuation of any estimator grows at least exponentially in the cube root of the sequence length.

Theorem 6 For every estimator q and all n,

$$R^n(q) \geq \exp\left(\frac{3}{2}n^{1/3}(1-o(1))\right).$$

Proof outline To prove this lower bound on the attenuation of any estimator, we combine an equivalence between estimation and sequential universal compression with Shtarkov's argument [29]. We then incorporate a lower bound on the maximum probabilities of patterns [17], and apply an asymptotic analysis technique developed by Hayman [15] to obtain the result. □

6. Examples

To better understand the behavior of the diminishing-attenuation estimators, we consider the conditional probabilities that the low-complexity estimator $q_{1/3}$ assigns to some simple sequences and compare it to what one would intuitively expect.

Consider first the sequence *aaa*.... 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 $1 - \Theta(1/n)$ for the next symbol being '*a*' and probability $\Theta(1/n)$ to a new symbol.

For the alternating sequence *abab*..., one would predict probability half for the next symbol being each of '*a*' and '*b*'. Correspondingly, the estimator assigns probability $\Theta(1/n)$ 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 *abcde*..., we would expect the next symbol to be new. Indeed the estimator assigns probability $1 - \Theta(1/n^{2/3})$ that the next symbol will be new.

But for large-alphabet sequences where the probability of new does not approach 1, intuition may not serve well. Consider perhaps the simplest such case, the sequence *aabbcc*.... 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 [27] 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 symbol is about $(0.12n)/(0.62n) \approx 0.2$.

Acknowledgments

We thank N. Alon, Y. Freund, I.J. Good, N. Jevtić, Sajama, W. Szpankowski, and K. Viswanathan for helpful discussions.

References

[1] J. Åberg, Y. Shtarkov, and B. Smeets. Multialphabet coding with separate alphabet description. In *Proceedings of compression and complexity of sequences*, 1997.

[2] 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.

[3] S. 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.

[4] K. Church and W. Gale. Probability scoring for spelling correction. *Statistics and Computing*, 1:93–103, 1991.

[5] B. Clarke and A. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[6] T. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, Jan 1991.

[7] I. Csisźar and P. Shields. Redundancy rates for renewal and other processes. *IEEE Transactions on Information Theory*, 42(6):2065–2072, Nov 1996.

[8] L. Davison. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, Nov 1973.

[9] W. Gale. Good Turing smoothing without tears. Technical report, AT&T Bell Laboratories, Murray Hill, NJ, 1994.

[10] W. Gale and K. Church. What is wrong with adding one? In N. Oostdijk and P. de Haan, editors, *Corpus based research into language*. Rodopi, Amsterdam, 1994.

[11] W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses. *Computers and Humanities*, 26:415–419, 1993.

[12] I. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, Dec 1953.

[13] I. 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.

[14] G. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75–115, 1918.

[15] W. Hayman. A generalization of Stirling's formula. *Journal für die reine und angewandte Mathematik*, 196:67–95, 1956.

[16] A. Hodges. *Alan Turing: The Enigma*. Walker & Co., 2000.

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

[18] R. Krichevsky and V. Trofimov. The preformance of universal coding. *IEEE Transactions on Information Theory*, 27:199–207, 1981.

[19] P. Laplace. *Philosphical essays on probabilities*. Springer Verlag, New York, Translated by A. Dale from the 5th (1825) edition, 1995.

[20] N. Littlestone and M. Warmuth. The weighted majority algorithm. In *IEEE Symposium on Foundations of Computer Science*, 1992.

[21] 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.

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

[23] A. Nadas. On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-33(6):1414–1416, Dec 1985.

[24] A. Nadas. Good, Jelinek, Mercer, and Robins on Turing's estimate of probabilities. *American Journal of Mathematical and Management Sciences*, 11:229–308, 1991.

[25] A. Orlitsky and N. Santhanam. Performance of universal codes over infinite alphabets. In *Proceedings of the Data Compression Conference*, Mar 2003.

[26] A. Orlitsky, N. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. Submitted to *IEEE Transactions on Information Theory*.

[27] A. Orlitsky and K. Viswanathan. The most likely alphabet. In preparation.

[28] J. Shtarkov. Coding of discrete sources with unknown statistics. In I. Csiszar and P. Elias, editors, *Topics in Information Theory (Coll. Math. Soc. J. Bolyai, no. 16)*, pages 559–574. Amsterdam, The Netherlands: North Holland, 1977.

[29] Y. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, Jul-Sep. 1987.

[30] F. Song and W. Croft. A general language model for information retrieval (poster abstract). In *Research and Development in Information Retrieval*, pages 279–280, 1999.

[31] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[32] H. Wilf. *Generatingfunctionology*, page 91. Academic Press, 1990.

[33] I. Witten and T. Bell. The zero frequency problem. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.