Entropy **2014**, *16*, 4168-4184; doi:10.3390/e16074168

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Characterizing the Asymptotic Per-Symbol Redundancy of Memoryless Sources over Countable Alphabets in Terms of Single-Letter Marginals

Maryam Hosseini and Narayana Santhanam *

Department of Electrical Engineering, University of Hawaii at Manoa, Honolulu, HI 96822, USA;
E-Mail: hosseini@hawaii.edu

* Author to whom correspondence should be addressed; E-Mail: nsanthan@hawaii.edu.

Received: 27 May 2014; in revised form: 24 June 2014 / Accepted: 7 July 2014 /
Published: 23 July 2014

Abstract: The minimum expected number of bits needed to describe a random variable is its entropy, assuming knowledge of the distribution of the random variable. On the other hand, universal compression describes data supposing that the underlying distribution is unknown, but that it belongs to a known set $\mathcal{P}$ of distributions. However, since universal descriptions are not matched exactly to the underlying distribution, the number of bits they use on average is higher, and the excess over the entropy used is the redundancy. In this paper, we study the redundancy incurred by the universal description of strings of positive integers ($\mathbb{Z}_+$), the strings being generated independently and identically distributed (*i.i.d.*) according an unknown distribution over $\mathbb{Z}_+$ in a known collection $\mathcal{P}$. We first show that if describing a single symbol incurs finite redundancy, then $\mathcal{P}$ is tight, but that the converse does not always hold. If a single symbol can be described with finite worst-case regret (a more stringent formulation than redundancy above), then it is known that describing length n *i.i.d.* strings only incurs vanishing (to zero) redundancy per symbol as n increases. On the contrary, we show it is possible that the description of a single symbol from an unknown distribution of $\mathcal{P}$ incurs finite redundancy, yet the description of length n *i.i.d.* strings incurs a constant (> 0) redundancy per symbol encoded. We then show a sufficient condition on single-letter marginals, such that length n *i.i.d.* samples will incur vanishing redundancy per symbol encoded.

Keywords: universal compression; redundancy; large alphabets; tightness; redundancy-capacity theorem

1. Introduction

A number of statistical inference problems of significant contemporary interest, such as text classification, language modeling and DNA microarray analysis, are what are called large alphabet problems. They require inference on sequences of symbols where the symbols come from a set (alphabet) with a size comparable or even larger than the sequence length. For instance, language models for speech recognition estimate distributions over English words using text samples much smaller than the English vocabulary.

An abstraction behind several of these problems is universal compression over large alphabets. The general idea here is to model the problem at hand with a collection of models $\mathcal{P}$ instead of a single distribution. The model underlying the data is assumed or known to belong to the collection $\mathcal{P}$, but the exact identity of the model remains unknown. Instead, we aim to use a universal description of data.

The universal description uses more bits on average (averaged over the random sample) than if the underlying model were known, and the additional number of bits used by the universal description is called the redundancy against the true model. The average excess bits over the entropy of the true model will be referred to as the model redundancy for that model. Since one does not know the true model in general, a common approach is to consider collection redundancy or simply redundancy, which is the supremum of the model redundancy, the supremum being taken over all models of the collection.

Typically, we look at sequences of *i.i.d.* symbols, and therefore, we usually refer to the redundancy of distributions over length n sequences obtained by *i.i.d.* sampling from distributions from $\mathcal{P}$. The length n of sequences considered will typically be referred to as the sample size.

The nuances of prediction, the compression or estimation where the alphabet size and sample size are roughly equal are not well captured by studying a collection over a finite alphabet when the sample size is increased to infinity. Rather, they are better captured when we begin with a countably infinite support and let the sample size approach infinity or when we let the alphabet size scale as a function of the sample size. However, the collection of all *i.i.d.* distributions over countably infinite supports has infinite redundancy that renders most estimations or prediction problems impossible. Therefore, there are several alternative formulations to tackle language modeling, classification and estimation questions over large alphabets.

Patterns: One line of work is the patterns [1] approach that considers the compression of the pattern of a sequence rather than the sequence itself. Patterns abstract the identities of symbols and indicate only the relative order of appearance. For example, the pattern of TATTLE is 121134, while that of HONOLULU is 12324545. The point to note is that patterns of length n *i.i.d.* sequences can be compressed (no matter what the underlying countably infinite alphabet is) with redundancy that grows sublinearly in n [1]; therefore, the excess bits needed to describe patterns are asymptotically vanishing per symbol encoded. Indeed, insights learned in this line of work will be used to understand the compression of sequences, as well, in this paper.

Envelope on Model Classes: A second line of work considers restricted model classes for applications, particularly where the collection of models can be described in terms of an envelope [2]. This approach leads to an understanding of the worst-case formulations. In particular, we are interested in the result that if the worst-case regret (different from and a more stringent formulation than the redundancy described here) of describing a single sample is finite, then the per-symbol redundancy diminishes to zero. We will interpret this result towards the end of the Introduction. While envelope classes are usually chosen so that they are compressible in the worst case, a natural extension is the possibility of choosing classes that are only average-case, but not worst-case, compressible. For this, we need to understand how the single-letter average case redundancy of a class influences the redundancy of compressing strings sampled *i.i.d.* from distributions in the class—the focus of this paper.

Data-derived Consistency: A third line of work ignores the uniform convergence framework underlying redundancy or regret formulations. This is useful for large or infinite alphabet model collections that have poor or no redundancy guarantees, but ask a question that cannot be answered with the approaches above. In this line of work, one obtains results on the model redundancy described above instead of (the collection) redundancy. For example, a model collection is said to be weakly compressible if there is a universal measure that ensures that for all models, the model redundancy normalized by the sample size (per-symbol) diminishes to zero. The rate at which the per-symbol model redundancy diminishes to zero depends on the underlying model and for some models could be arbitrarily slower than others. Given a particular block length n, however large, there may be, hence, no non-trivial guarantee that holds over the entire model collection, unlike the redundancy formulation.

However, if we add on the additional constraint that we should estimate the rate of convergence from the data, we get the data-derived consistency formulations in [3]. Fundamental to further research in this direction is a better understanding of how single-letter redundancy (of $\mathcal{P}$) relates to the redundancy of length n strings (that of $\mathcal{P}^n$). The primary theme of this paper is to collect such results on the redundancy of classes over countably infinite support.

In the fixed alphabet setting, this connection is well understood. If the alphabet has size k, the redundancy of $\mathcal{P}$ is easily seen to be always finite (in fact, $\leq \log k$) and that of $\mathcal{P}^n$ scales as $\frac{k-1}{2}\log n$. However, when $\mathcal{P}$ does not have a finite support, the above bounds are meaningless.

Redundancy Capacity Theorem: On the other hand, the redundancy of a collection $\mathcal{P}$ over a countably infinite support may be infinite. In this paper we let $\mathbb{Z}_+ = \{1, 2, 3, ...\}$ be the set of positive integers and $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of non-negative integers. However, what about the case where the redundancy of a collection $\mathcal{P}$ over $\mathbb{Z}_+$ is finite? Now, a well-known redundancy-capacity [4] argument can be used to interpret the redundancy, which equates the redundancy to the amount of information we can get about the source from the data. In this case, finite (infinite, respectively) redundancy of $\mathcal{P}$ implies that a single symbol contains a finite (infinite, respectively) amount of information about the model.

The natural question then is the following. If a collection $\mathcal{P}$ over $\mathbb{Z}_+$ has finite redundancy, does it imply that the redundancy of length n *i.i.d.* strings from $\mathcal{P}$ grows sublinearly? Equivalently, do finite redundancy collections behave similar to their fixed alphabet counterparts? If true, roughly speaking, such a result would inform us that as the universal encoder sees more and more of the sequence, it learns less and less of the underlying model. This would be in line with our intuition, where seeing more data fixes the model. Therefore, the more data we have already seen, the less there is to learn. Yet, as we will show, that is not the case.

Results: To understand these nuances, we first show that if the redundancy of a collection $\mathcal{P}$ of distributions over $\mathbb{Z}_+$ is finite, then $\mathcal{P}$ is tight. This turns out to be a useful tool to check if the redundancy is finite in [3], for example.

However, in a departure from other worst-case regret formulations, as in [2], we demonstrate that it is possible for a class $\mathcal{P}$ to have finite redundancy, yet the asymptotic per-symbol redundancy of strings sampled *i.i.d.* from $\mathcal{P}$ is bounded away from zero. Therefore, roughly speaking, no matter how much of the sequence the universal encoder has seen, it learns at least a constant number of bits about the underlying model each time it sees an additional symbol. No matter how much data we see, there is more to learn about the underlying model! We finally obtain a sufficient condition on a class $\mathcal{P}$, such that the asymptotic per-symbol redundancy of length n *i.i.d.* strings diminishes to zero.

2. Notation and Background

We introduce the notation used in the paper, as well as some prior results that will be used. Following information theoretic conventions, log indicates logarithms to base two and ln to base e. In this paper we let $\mathbb{Z}_+ = \{1, 2, 3, ...\}$ be the set of positive integers and $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of non-negative integers.

2.1. Redundancy

The notation used here is mostly standard, but we include it for completeness. Let $\mathcal{P}$ be a collection of distributions over $\mathbb{Z}_+$. Let $\mathcal{P}^n$ be the set of distributions over length-n sequences obtained by *i.i.d.* sampling from distributions in $\mathcal{P}$.

$\mathcal{P}^\infty$ is the collection of measures over infinite length sequences of $\mathbb{Z}_+$ obtained by *i.i.d.* sampling as follows. Observe that $\mathbb{Z}_+^n$ is countable for every n. For simplicity of exposition, we will think of each length n string $\mathbf{x}$ as a subset of $\mathbb{Z}_+^\infty$—the set of all semi-infinite strings of positive integers that begin with $\mathbf{x}$. Each subset of $\mathbb{Z}_+^n$ is therefore a subset of $\mathbb{Z}_+^\infty$. Now the collection $\mathcal{J}$ of all subsets of $\mathbb{Z}_+^n$ and all $n \in \mathbb{Z}_+$, is a semi-algebra [5]. The probabilities *i.i.d.* sampling assigns to finite unions of disjoint sets in $\mathcal{J}$ is the sum of that assigned to the components of the union. Therefore, there is a sigma-algebra over the uncountable set $\mathbb{Z}_+^\infty$ that extends $\mathcal{J}$ and matches the probabilities assigned to sets in $\mathcal{J}$ by *i.i.d.* sampling. The reader can assume that $\mathcal{P}^\infty$ is the measure on the minimal sigma-algebra that extends $\mathcal{J}$ and matches what the probabilities *i.i.d.* sampling gives to sets in $\mathcal{J}$. See, e.g., [5], for a development of elementary measure theory that lays out the above steps.

Let q be a measure over infinite sequences that we call:

$$R_n(\mathcal{P}^\infty) = \inf_q \sup_{p \in \mathcal{P}^\infty} E_p \log \frac{p(X^n)}{q(X^n)} \tag{1}$$

the redundancy of length n sequences, or length n *i.i.d.* redundancy, or simply length n redundancy. The single-letter redundancy refers to the special case when $n = 1$. We often normalize $R_n(\mathcal{P}^\infty)$ in (1) by the block length n. We will call $R_n(\mathcal{P}^\infty)/n$ the per-symbol length n redundancy.

In particular, note the distinction between single letter and per-symbol length n redundancy. In the definition (1), we do not require q to be *i.i.d.*. The single-letter redundancy would correspond to obtaining the infimum in (1) only over the restricted class of *i.i.d.* measures, while the per-symbol length n redundancy allows for the infimum over all possible measures q. Thus, the per-symbol length n redundancy is upper bounded by the single letter redundancy. Any difference between the two can be thought of as the advantage accrued, because the universal measure learns the underlying measure p.

In this paper, our primary goal is to understand the connections between the single-letter redundancy, on the one hand, and the behavior of length n *i.i.d.* redundancy, on the other. As mentioned in the Introduction, length n redundancy is the capacity of a channel from $\mathcal{P}$ to $\mathbb{Z}_+^n$, where the conditional probability distribution over $\mathbb{Z}_+^n$ given $p \in \mathcal{P}$ is simply the distribution p over length n sequences. Roughly speaking, it quantifies how much information about the source we can extract from the sequence.

We will often speak of the per-symbol length n redundancy, which is simply length n redundancy normalized by n, *i.e.*,, $R_n(\mathcal{P}^\infty)/n$. Furthermore, the limit $\limsup_{n\to\infty} R_n(\mathcal{P}^\infty)/n$ is the asymptotic per-symbol redundancy. Whether the asymptotic per-symbol redundancy is zero (we will equivalently say that the asymptotic per-symbol redundancy *diminishes* to zero to keep in line with prior literature) is in many ways a litmus test for compression, estimation and other related problems. Loosely speaking, if $R_n(\mathcal{P}^\infty)/n \to 0$, the redundancy-capacity interpretation [4] mentioned above implies that after a point, there is little further information to be learned when we see an additional symbol, no matter what the underlying source is. In this sense, this is the case where we can actually learn the underlying model at a uniform rate over the entire class.

We note that it is possible to define an even more stringent notion—a worst-case-regret. For length n sequences, this is:

$$\inf_q \sup_{p\in\mathcal{P}^\infty} \sup_{X^n} \log \frac{p(X^n)}{q(X^n)}.$$

Single-letter regret is the special case where $n = 1$, and asymptotic per-symbol regret is the limit as $n \to \infty$ of the length n regret normalized by n. We will not concern ourselves with the worst case formulation in this paper, but mention it in passing for comparison. In the worst-case setting, finite single letter redundancy is necessary and sufficient [2] for the asymptotic per-symbol worst-case regret to diminish to zero.

Yet, we show in this paper that it is not necessarily the case for redundancy. It is quite possible that collections with finite single-letter redundancy have asymptotic per-symbol redundancy bounded away from zero.

2.2. Patterns

Recent work [1] has formalized a similar framework for countably infinite alphabets. This framework is based on the notion of patterns of sequences that abstract the identities of symbols and indicates only the relative order of appearance. For example, the pattern of PATTERN is 1233456. The k-th distinct symbol of a string is given an index k when it first appears, and that index is used every time the symbol appears henceforth. The crux of the patterns approach is to consider the set of measures induced over patterns of the sequences instead of considering the set of measures $\mathcal{P}$ over infinite sequences,

Denote the pattern of a string $\mathbf{x}$ by $\Psi(\mathbf{x})$. There is only one possible pattern of strings of length one (no matter what the alphabet, the pattern of a length one string is one), two possible patterns of strings of length two (11 and 12), and so on. The number of possible patterns of length n is the n-th Bell number [1], and we denote the set of all possible length n patterns by Ψ^n. The measures induced on patterns by a corresponding measure p on infinite sequences of positive integers assigns to any pattern ψ a probability:

$$p(\psi) = p(\{\mathbf{x} : \Psi(\mathbf{x}) = \psi\}).$$

In [1], the length n pattern redundancy,

$$\inf_q \sup_{p \in \mathcal{P}^\infty} E_p \log \frac{p(\Psi(X^n))}{q(\Psi(X^n))},$$

was shown to be upper bounded by $\pi(\log e)\sqrt{\frac{2n}{3}}$. It was also shown in [6] that there is a measure q over infinite length sequences that satisfies for all n simultaneously:

$$\sup_{p \in \mathcal{P}^\infty} \sup_{X^n} \log \frac{p(\Psi(X^n))}{q(\Psi(X^n))} \leq \pi(\log e)\sqrt{\frac{2n}{3}} + \log(n(n+1)).$$

Let the measure induced on patterns by q be denoted as q_Ψ for convenience.

We can interpret the probability estimator q_Ψ as a sequential prediction procedure that estimates the probability that the symbol X_{n+1} will be "new" (has not appeared in X_1^n) and the probability that X_{n+1} takes a value that has been seen so far. This view of estimation also appears in the statistical literature on Bayesian nonparametrics that focuses on exchangeability. Kingman [7] advocated the use of exchangeable random partitions to accommodate the analysis of data from an alphabet that is not bounded or known in advance. A more detailed discussion of the history and philosophy of this problem can be found in the works of Zabell [8,9] collected in [10].

2.3. Cumulative Distributions and Tight Collections

For our purposes, the cumulative distribution function of any probability distribution p on $\mathbb{Z}_+$ ($\mathbb{N}$, respectively) is a function $F_p : \mathbb{R} \cup \{\infty\} \to [0,1]$ defined in the following (slightly unconventional) way. We let $F_p(0) = 0$ in case the support is $\mathbb{Z}_+$ ($F_p(-1) = 0$ if the support is $\mathbb{N}$, respectively). We then define F_p on points in the support of p in the way cumulative distribution functions are normally defined. Specifically for all y in the support of p,

$$F_p(y) = \sum_{j \geq 0}^{y} p(j).$$

We let $F_p(-\infty) := 0$ and $F_p(\infty) := 1$. Finally, we extend the definition of F_p to all real numbers by linearly interpolating between the values defined already.

Let $F_p^{-1} : [0,1] \mapsto \mathbb{R} \cup \{\infty\}$ denote the inverse function of F_p defined as follows. To begin with,

$$F_p^{-1}(0) = \sup\{y : F_p(y) = 0\},$$

If p has infinite support, then $F_p^{-1}(1) = \infty$, else $F_p^{-1}(1)$ is the smallest positive integer y, such that $F_p(y) = 1$. It follows [11] then that:

$$p\{x \in \mathbb{Z}_+ : x \geq F_q^{-1}(1-\gamma)\} > \gamma \text{ and } p\{x \in \mathbb{Z}_+ : x > 2F_q^{-1}(1-\frac{\gamma}{2})\} \leq \gamma$$

A collection $\mathcal{P}$ of distributions on $\mathbb{Z}_+$ is defined to be tight if for all $\gamma > 0$,

$$\sup_{p\in\mathcal{P}} F_p^{-1}(1-\gamma) < \infty.$$

3. Redundancy and Tightness

We focus on the single-letter redundancy in this section and explore the connections between the single-letter redundancy of a collection $\mathcal{P}$ and the tightness of $\mathcal{P}$.

Lemma 1. A collection $\mathcal{P}$ over $\mathbb{N}$ with bounded length n redundancy is tight. Namely, if the single-letter redundancy of $\mathcal{P}$ is finite, then for any $\gamma > 0$:

$$\sup_{p\in\mathcal{P}} F_p^{-1}(1-\gamma) < \infty.$$

Proof Since $\mathcal{P}$ has bounded single-letter redundancy, fix a distribution q over $\mathbb{N}$, such that:

$$\sup_{p\in\mathcal{P}} D(p||q) < \infty.$$

We define $R \overset{\text{def}}{=} \sup_{p\in\mathcal{P}} D(p||q)$ where $D(p||q)$ is the Kullback–Leibler distance between p and q. We will first show that for all $p \in \mathcal{P}$ and any $m > 0$,

$$p(\left|\log\frac{p(X)}{q(X)}\right| > m) \le (R + (2\log e)/e)/m. \tag{2}$$

To see Equation (2), let S be the set of all $x \in \mathbb{N}$, such that $p(x) < q(x)$. A well-known convexity argument shows that the partial contribution to KL divergence from S,

$$\sum_{x\in S} p(x)\log\frac{p(x)}{q(x)} \ge p(S)\log\frac{p(S)}{q(S)} \ge -\frac{\log e}{e},$$

and hence:

$$\sum_{x\in\mathbb{Z}_+} p(x)\left|\log\frac{p(x)}{q(x)}\right| = \sum_{x\in\mathbb{Z}_+} p(x)\log\frac{p(x)}{q(x)} - 2\sum_{x\in S} p(x)\log\frac{p(x)}{q(x)} \le R + 2\frac{\log e}{e}.$$

Then, Equation (2) follows by a simple application of Markov's inequality.

We will now use Equation (2) to complete the proof of the lemma. Specifically, we will show that for all $\gamma > 0$,

$$\sup_{p\in\mathcal{P}} F_p^{-1}(1-\gamma) < 2F_q^{-1}(1-\gamma/2^{m^*+2})$$

where m^* is the smallest integer, such that $(R + (2\log e)/e)/m^* < \gamma/2$. Equivalently, for all $\gamma > 0$ and $p \in \mathcal{P}$, we show that:

$$p\{x \in \mathbb{N} : x > 2F_q^{-1}(1-\gamma/2^{m^*+2})\} \le \gamma.$$

We prove the above by partitioning q's tail; numbers $x \ge 2F_q^{-1}(1-\gamma/2^{m^*+2})$ into two parts.

(i) the set $W_1 = \{x \in \mathbb{N} : x > 2F_q^{-1}(1 - \gamma/2^{m+2})$ and $\log \frac{p(x)}{q(x)} > m^*\}$. Clearly:

$$W_1 \subseteq \{y \in \mathbb{N} : \left|\log \frac{p(y)}{q(y)}\right| > m^*\},$$

and thus:

$$p(W_1) \leq p\{y \in \mathbb{N} : \left|\log \frac{p(y)}{q(y)}\right| > m^*\} \leq \frac{\gamma}{2}$$

where the right inequality follows from Equation (2).

(ii) the set $W_2 = \{x \in \mathbb{N} : x > 2F_q^{-1}(1 - \gamma/2^{m^*+2})$ and $\log \frac{p(x)}{q(x)} \leq m^*\}$. Clearly:

$$W_2 \subseteq \{y \in \mathbb{N} : y > 2F_q^{-1}(1 - \gamma/2^{m^*+2})\}$$

and therefore:

$$q(W_2) \leq q\{y \in \mathbb{N} : y > 2F_q^{-1}(1 - \gamma/2^{m^*+2})\} \leq \frac{\gamma}{2^{m^*+1}}.$$

By definition, all $x \in W_2$ satisfy $\log \frac{p(x)}{q(x)} \leq m^*$ or that $p(x) \leq q(x)2^{m^*}$. Hence, we have:

$$p(W_2) \leq q(W_2)2^{m^*} \leq \frac{\gamma 2^{m^*}}{2^{m^*+1}} = \frac{\gamma}{2}.$$

The lemma follows. □

The converse is not necessarily true. Tight collections need not have finite single-letter redundancy, as the following example demonstrates.

Construction: Consider the following collection $\mathcal{I}$ of distributions over $\mathbb{Z}_+$. First, partition the set of positive integers into the sets T_i, $i \in \mathbb{N}$, where:

$$T_i = \{2^i, \ldots, 2^{i+1} - 1\}.$$

Note that $|T_i| = 2^i$. Now, $\mathcal{I}$ is the collection of all possible distributions that can be formed as follows: for all $i \in \mathbb{Z}_+$, pick exactly one element of T_i and assign probability $1/((i+1)(i+2))$ to the element of T_i chosen choosing the support as above implicitly assumes the axiom of choice. Note that the set $\mathcal{I}$ is uncountably infinite. □

Corollary 2. The set $\mathcal{I}$ of distributions is tight.

Proof For all $p \in \mathcal{I}$,

$$\sum_{\substack{x \geq 2^k \\ x \in \mathbb{Z}_+}} p(x) = \frac{1}{k+1},$$

namely, all tails are uniformly bounded over the collection $\mathcal{I}$. Put another way, for all $\delta > 0$ and all distributions: $p \in \mathcal{I}$,

$$F_p^{-1}(1 - \delta) \leq 2^{\frac{1}{\delta}}.$$ □

On the other hand:

Proposition 1. The collection $\mathcal{I}$ does not have finite redundancy.

Proof Suppose q is any distribution over $\mathbb{Z}_+$. We will show that $\exists p \in \mathcal{I}$, such that:

$$\sum_{x \in \mathbb{Z}_+} p(x) \log \frac{p(x)}{q(x)}$$

is not finite. Since the entropy of every $p \in \mathcal{I}$ is finite, we just have to show that for any distribution q over $\mathbb{Z}_+$, there exists $p \in \mathcal{I}$, such that:

$$\sum_{x \in \mathbb{Z}_+} p(x) \log \frac{1}{q(x)}$$

is not finite.

Consider any distribution q over $\mathbb{Z}_+$. Observe that for all i, $|T_i| = 2^i$. It follows that for all i, there is $x_i \in T_i$, such that:

$$q(x_i) \leq \frac{1}{2^i}.$$

However, by construction, $\mathcal{I}$ contains a distribution p^* that has for its support $\{x_i : i \in \mathbb{Z}_+\}$ identified above. Furthermore p^* assigns:

$$p^*(x_i) = \frac{1}{(i+1)(i+2)} \qquad \forall\, i \in \mathbb{Z}_+.$$

The KL divergence from p^* to q is not finite, and the Lemma follows, since q is arbitrary. □

4. Length n Redundancy

We study how the single-letter properties of a collection $\mathcal{P}$ of distributions influence the compression of length n strings obtained by *i.i.d.* sampling from distributions in $\mathcal{P}$. Namely, we try to characterize when the length n redundancy of $\mathcal{P}^\infty$ grows sublinearly in the block length n.

Lemma 3. Let $\mathcal{P}$ be a collection of distributions over a countable support $\mathcal{X}$. For some $m \in \mathbb{Z}_+$, consider m pairwise disjoint subsets $S_i \subset \mathcal{X}$ ($1 \leq i \leq m$), and let $\delta > 1/2$. If there exist $p_1, \ldots, p_m \in \mathcal{P}$, such that:

$$p_i(S_i) \geq \delta,$$

then for all distributions q over $\mathcal{X}$,

$$\sup_{p \in \mathcal{P}} D(p||q) \geq \delta \log m.$$

In particular if there are an infinite number of sets S_i, $i \in \mathbb{Z}_+$ and distributions $p_i \in \mathcal{P}$, such that $p_i(S_i) \geq \delta$, then the redundancy is infinite.

Proof This is a simplified formulation of the distinguishability concept in [4]. For a proof, see e.g., [12]. □

4.1. Counterexample

We now show that it is possible for the single-letter redundancy of a collection $\mathcal{B}$ of distributions to be finite, yet the asymptotic per-symbol redundancy (the length n redundancy of $\mathcal{B}^\infty$ normalized by n) remains bounded away from zero; in the limit, the block length goes to infinity. To show this, we obtain such a collection $\mathcal{B}$.

Construction: As before, partition the set $\mathbb{Z}_+$ into $T_i = \{2^i, \dots, 2^{i+1} - 1\}$, $i \in \mathbb{N}$. Recall that T_i has 2^i elements. For all $0 < \epsilon \leq 1$, let $n_\epsilon = \lfloor \frac{1}{\epsilon} \rfloor$. Let $1 \leq j \leq 2^{n_\epsilon}$, and let $p_{\epsilon,j}$ be a distribution on $\mathbb{Z}_+$ that assigns probability $1 - \epsilon$ to the number one (or equivalently, to the set T_0) and ϵ to the j-th smallest element of T_{n_ϵ}, namely the number $2^{n_\epsilon} + j - 1$. $\mathcal{B}$ (mnemonic for binary, since every distribution has a support of size two) is the collection of distributions $p_{\epsilon,j}$ for all $\epsilon > 0$ and $1 \leq j \leq 2^{n_\epsilon}$. $\mathcal{B}^\infty$ is the set of measures over infinite sequences of numbers corresponding to *i.i.d.* sampling from $\mathcal{B}$. □

We first verify that the single-letter redundancy of $\mathcal{B}$ is finite.

Proposition 2. Let q be a distribution that assigns $q(T_i) = \frac{1}{(i+1)(i+2)}$ and for all $j \in T_i$,

$$q(j|T_i) = \frac{1}{|T_i|}.$$

Then:

$$\sup_{p \in \mathcal{B}} \sum_{x \in \mathbb{Z}_+} p(x) \log \frac{p(x)}{q(x)} \leq 2. \qquad \square$$

However, the redundancy of compressing length n sequences from $\mathcal{B}^\infty$ scales linearly with n.

Proposition 3. For all $n \in \mathbb{Z}_+$,

$$\inf_q \sup_{p \in B^\infty} E_p \log \frac{p(X^n)}{q(X^n)} \geq n \left(1 - \frac{1}{n}\right)^n.$$

Proof Let the set $\{1^n\}$ denote a set containing a length n sequence of only ones. For all n, define 2^n pairwise disjoint sets S_i of $\mathbb{Z}_+^n$, $1 \leq i \leq 2^n$, where:

$$S_i = \{1, 2^n + i - 1\}^n - \{1^n\}$$

is the set of all length n strings containing at most two numbers (one and $2^n + i - 1$) and at least one occurrence of $2^n + i - 1$. Clearly, for distinct i and j between one and 2^n, S_i and S_j are disjoint. Furthermore, the measure $p_{\frac{1}{n},i} \in \mathcal{B}^\infty$ assigns S_i the probability:

$$p_{\frac{1}{n},i}(S_i) = 1 - \left(1 - \frac{1}{n}\right)^n > 1 - \frac{1}{e}.$$

From Lemma 3, it follows that length n redundancy of $\mathcal{B}^\infty$ is lower bounded by:

$$\left(1 - \frac{1}{e}\right) \log 2^n = n\left(1 - \frac{1}{e}\right). \qquad \square$$

In a preview of what is to come, we notice that though the single-letter redundancy of the class $\mathcal{B}$ over $\mathbb{Z}_+$ is finite, the single-letter tail redundancy, as described in the equation below, does not diminish to zero; namely, for all M:

$$\sup_{p\in\mathcal{B}} \sum_{x\geq M} p(x)\log\frac{p(x)}{q(x)} \geq 1.$$

In fact, in the next section, we relate the single-letter tail redundancy above diminishing to zero to sublinear growth of the *i.i.d.* length n redundancy.

4.2. Sufficient Condition

In this section, we show a sufficient condition on single-letter marginals of $\mathcal{P}$ and its redundancy that allows for *i.i.d.* length n redundancy of $\mathcal{P}^\infty$ to grow sublinearly with n. This condition is, however, not necessary; and the characterization of a condition that is both necessary and sufficient is as yet open.

For all $\epsilon > 0$, let $A_{p,\epsilon}$ be the set of all elements in the support of p with probability $\geq \epsilon$, and let $T_{p,\epsilon} = \mathbb{Z}_+ - A_{p,\epsilon}$. Let $G_0 = \{\phi\}$, where ϕ denotes the empty string. For all i, the sets:

$$G_i = \{x^i : A_{p,\frac{2\ln(i+1)}{i}} \subseteq \{x_1, x_2, \ldots, x_i\}\}$$

where, in a minor abuse of notation, we use $\{x_1, \ldots, x_i\}$ to denote the set of distinct symbols in the string x_1^i. Let $B_0 = \{\}$, and let $B_i = \mathbb{Z}_+^i - G_i$. Observe from an argument similar to the coupon collector problem that:

Lemma 4. For all $i \geq 2$,

$$p(B_i) \leq \frac{1}{(i+1)\ln(i+1)}.$$

Proof The proof follows from an elementary union bound:

$$\begin{aligned}
p(B_i) &\leq |A_{p,\frac{2\ln(i+1)}{i}}| \left(1 - \frac{2\ln(i+1)}{i}\right)^i \\
&\leq \frac{i}{2\ln(i+1)}\left(1 - \frac{2\ln(i+1)}{i}\right)^i \\
&\leq \frac{i}{2\ln(i+1)} e^{\frac{-2i\ln(i+1)}{i}} \\
&\leq \frac{i}{2(i+1)^2\ln(i+1)} \\
&\leq \frac{1}{(i+1)\ln(i+1)}.
\end{aligned}$$

□

Theorem 5. Suppose $\mathcal{P}$ is a collection of distributions over $\mathbb{Z}_+$. Let the entropy be uniformly bounded over the entire collection, and in addition, let the redundancy of the collection be finite. Namely,

$$\sup_{p\in\mathcal{P}} \sum_{x\in\mathbb{Z}_+} p(x)\log\frac{1}{p(x)} \overset{\text{def}}{=} H < \infty \quad \text{and } \exists q_1 \text{ over } \mathbb{Z}_+ \text{ s.t.} \quad \sup_{p\in\mathcal{P}} \sum_{x\in\mathbb{Z}_+} p(x)\log\frac{p(x)}{q_1(x)} < \infty.$$

We will denote:

$$R = \sup_{p\in\mathcal{P}} \sum_{x\in\mathbb{Z}_+} p(x)\log\frac{p(x)}{q_1(x)}.$$

Recall that for any distribution p, the set $T_{p,\delta}$ denotes the support of p, all of whose probabilities are $< \delta$. Let:

$$\lim_{\delta\to 0}\sup_{p\in\mathcal{P}}\sum_{x\in T_{p,\delta}} p(x)\log\frac{1}{p(x)} = 0 \quad \text{and } \exists q_1 \text{ over } \mathbb{Z}_+ \text{ s.t.} \quad \lim_{\delta\to 0}\sup_{p\in\mathcal{P}}\sum_{x\in T_{p,\delta}} p(x)\log\frac{p(x)}{q_1(x)} = 0. \tag{3}$$

Then, the redundancy of length n distributions obtained by *i.i.d.* sampling from distributions in $\mathcal{P}$, denoted by $R_n(\mathcal{P}^\infty)$, grows sublinearly:

$$\limsup_{n\to\infty}\frac{1}{n}R_n(\mathcal{P}^\infty) = 0.$$

Remark If the conditions of the theorem are met, we can always assume without loss of generality that there is a distribution q_1 that satisfies (3) and simultaneously has finite redundancy. To see this, suppose $q_1^{'}$ satisfies the finite-redundancy condition, namely:

$$\sup_{p\in\mathcal{P}}\sum_{x\in T_{p,\delta}} p(x)\log\frac{p(x)}{q_1^{'}(x)} = 0;$$

while a different distribution $q_1^{''}$ satisfies the second tail-redundancy condition,

$$\lim_{\delta\to 0}\sup_{p\in\mathcal{P}}\sum_{x\in T_{p,\delta}} p(x)\log\frac{p(x)}{q_1^{''}(x)} = 0.$$

It is easy to verify that the distribution q that assigns to any $x \in \mathbb{Z}_+$, $q_1(x) = \frac{q_1^{'}(x)+q_1^{''}(x)}{2}$ satisfies both conditions simultaneously. □

Proof In what follows, x^i represents a string $x_1, \ldots, x_i$ and x^0 denotes the empty string. For all n, we denote $\Psi(x^n) = \psi_1, \ldots, \psi_n$ and $\Psi(X^n) = \Psi_1, \ldots, \Psi_n$.

We will construct q, such that $\limsup_{n\to\infty}\frac{1}{n}E_p\log\frac{p(X^n)}{q(X^n)} = 0$. Recall that q_Ψ is the optimal universal pattern encoder over patterns of *i.i.d.* sequences defined in Section 2.2. Furthermore, recall that the redundancy of $\mathcal{P}$ is finite and that q_1 is the universal distribution over $\mathbb{Z}_+$ that attains redundancy R for $\mathcal{P}$.

The universal encoder q is now defined as follows:

$$\begin{aligned} q(x^n) &= q(x^n, \Psi(x^n)) \\ &= q(\psi_1, x_1, \psi_2, x_2, \ldots, \psi_n, x_n) \\ &= \prod_{i\in\mathbb{Z}_+} q(\psi_i|\psi_1^{i-1}, x_1^{i-1}) \prod_{j\in\mathbb{Z}_+} q(x_j|\psi_1^j, x_1^{j-1}) \\ &\overset{\text{def}}{=} \prod_{i\in\mathbb{Z}_+} q_\Psi(\psi_i|\psi_1^{i-1}) \prod_{j\in\mathbb{Z}_+} q(x_j|\psi_1^j, x_1^{j-1}). \end{aligned}$$

Furthermore, we define for all $x_1^{i-1} \in \mathbb{Z}_+^{i-1}$ and all $\psi^i \in \Psi^i$, such that $\psi^{i-1} = \Psi(x^{i-1})$,

$$q(x_i|\psi_1^i, x_1^{i-1}) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } x_i \in \{x_1, \ldots, x_{i-1}\} \text{ and } \Psi(x^i) = \psi^i \\ q_1(x_i) & \text{if } x_i \notin \{x_1, \ldots, x_{i-1}\} \text{ and } \Psi(x^i) = \psi^i. \end{cases}$$

Namely, we use an optimal universal pattern encoder over patterns of *i.i.d.* sequences and encode any new symbol using a universal distribution over $\mathcal{P}$. We now bound the redundancy of q as defined above. We have for all $p \in \mathcal{P}^\infty$,

$$
\begin{aligned}
E_p \log \frac{p(X^n)}{q(X^n)} &= \sum_{x^n} p(x^n) \log \prod_{i \in \mathbb{Z}_+} \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})} \prod_{j \in \mathbb{Z}_+} \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})} \\
&= \sum_{x^n} p(x^n) \sum_{i=1}^n \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})} + \sum_{x^n} p(x^n) \sum_{j=1}^n \log \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})}.
\end{aligned}
$$

Since ψ_1 is always one, $p(\psi_1) = q_\Psi(\psi_1) = 1$. Therefore, we have:

$$
\sum_{x^n} p(x^n) \sum_{i=1}^n \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})} = \sum_{x^n} p(x^n) \sum_{i=2}^n \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})}.
$$

The first term, normalized by n, can be upper bounded as follows:

$$
\begin{aligned}
&\frac{1}{n} \sum_{x^n} p(x^n) \sum_{i=2}^n \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})} \\
&\qquad \le \frac{1}{n} \sum_{i=2}^n \sum_{x_1^i} p(x_1^i) \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{p(\psi_i|\psi_1^{i-1})} + \frac{1}{n}\left(\pi \log e \sqrt{\frac{2n}{3}} + \log n(n+1)\right) \\
&\qquad = \frac{1}{n} \sum_{i=2}^n (H(\Psi_i|\Psi_1^{i-1}) - H(\Psi_i|X_1^{i-1})) + \frac{1}{n}\left(\pi \log e \sqrt{\frac{2n}{3}} + \log n(n+1)\right) \\
&\qquad \le \frac{1}{n}(nH_p) - \frac{1}{n} \sum_{i=2}^n H(\Psi_i|X_1^{i-1})) + \frac{1}{n}\left(\pi \log e \sqrt{\frac{2n}{3}} + \log n(n+1)\right)
\end{aligned}
$$

where we define H_p as:

$$
H_p \overset{\text{def}}{=} \sum_{x \in \mathbb{Z}_+} p(x) \log \frac{1}{p(x)}
$$

and the last inequality follows, since:

$$
H(\Psi^n) \le H(X^n) = nH_p.
$$

Now for $i \ge 2$,

$$
\begin{aligned}
H(\Psi_i|X_1^{i-1}) &= \sum p(\psi_i|x^{i-1}) p(x^{i-1}) \log \frac{1}{p(\psi_i|x_1^{i-1})} \\
&= \sum p(x^{i-1}) \left(\sum_{x \in \{x_1, \ldots, x_{i-1}\}} p(x) \log \frac{1}{p(x)} + \sum_{y \notin \{x_1, \ldots, x_{i-1}\}} p(y) \log \frac{1}{\sum_{y \notin \{x_1, \ldots, x_{i-1}\}} p(y)} \right)
\end{aligned}
$$

Then,

$$\begin{aligned}
H_p - H(\Psi_i|X_1^{i-1}) &= \sum_{x^{i-1}} p(x^{i-1}) H_p - H(\Psi_i|X_1^{i-1}) \\
&= \sum_{x^{i-1}} p(x^{i-1}) \sum_{x_i \notin \{x_1,\dots,x_{i-1}\}} p(x_i) \log \frac{1}{p(x_i)} \\
&\qquad - \sum_{x^{i-1}} p(x^{i-1}) \sum_{x_i \notin \{x_1,\dots,x_{i-1}\}} p(x_i) \log \frac{1}{\sum_{x_i \notin \{x_1,\dots,x_{i-1}\}} p(x_i)} \\
&\le \sum_{x^{i-1}} p(x_1^{i-1}) \sum_{x_i \notin \{x_1,\dots,x_{i-1}\}} p(x_i) \log \frac{1}{p(x_i)} \\
&\le p(G_{i-1}) \sum_{x_i \in T_{p,2\frac{\ln i}{i-1}}} p(x_i) \log \frac{1}{p(x_i)} + p(B_{i-1}) H \\
&\le \sum_{x_i \in T_{p,2\frac{\ln i}{i-1}}} p(x_i) \log \frac{1}{p(x_i)} + \frac{H}{i \ln i}.
\end{aligned}$$

We have split the length $i-1$ sequences into the sets G_{i-1} and B_{i-1} and use separate bounds on each set that hold uniformly over the entire model collection. The last inequality above follows from Lemma 4. From Condition (3) of the Theorem, we have that:

$$\lim_{i\to\infty} \sup_{p\in\mathcal{P}} \sum_{x \in T_{p,2\frac{\ln i}{i-1}}} p(x) \log \frac{1}{p(x)} = 0.$$

Therefore, we have:

$$\lim_{n\to\infty} \sup_{p\in\mathcal{P}} \frac{1}{n} \sum_{i=2}^{n} \left(\sum_{x \in T_{p,2\frac{\ln i}{i-1}}} p(x) \log \frac{1}{p(x)} + \frac{H}{i \ln i} \right) \le \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} \left(\sup_{p\in\mathcal{P}} \sum_{x \in T_{p,2\frac{\ln i}{i-1}}} p(x) \log \frac{1}{p(x)} + \frac{H}{i \ln i} \right) \overset{(a)}{=} 0.$$

The first term on the left in the first equation above is non-negative, hence the limit above has to equal zero. The equality (a) follows from Cesaro's lemma asserting that for any sequence $\{a_i, i \in \mathbb{Z}_+\}$ with $a_i < \infty$ for all i, if $\lim_{i\to\infty} a_i$ exists, then:

$$\lim_{i\to\infty} a_i = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} a_j.$$

Therefore,

$$\lim_{n\to\infty} \sup_{p\in\mathcal{P}} \frac{1}{n} \sum_{x^n} p(x^n) \sum_{i=2}^{n} \log \frac{p(\psi_i|\psi_1^{i-1}, x_1^{i-1})}{q_\Psi(\psi_i|\psi_1^{i-1})} = 0.$$

For the second term, observe that:

$$\sum_{x^n} p(x^n) \sum_{j=1}^{n} \log \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})} \le R + \sum_{x^n} p(x^n) \sum_{j=2}^{n} \log \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})}.$$

Furthermore,

$$\sum_{x^n} p(x^n) \sum_{j=2}^{n} \log \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})} = \sum_{j=2}^{n} \sum_{x^j} p(x^j) \log \frac{p(x_j|\psi_1^j, x_1^{j-1})}{q(x_j|\psi_1^j, x_1^{j-1})}$$
$$\leq \sum_{j=2}^{n} \sum_{x^{j-1}} p(x^{j-1}) \sum_{x_j \notin \{x_1,\ldots,x_{j-1}\}} p(x_j) \log \frac{p(x_j)}{q_1(x_j)}$$
$$\leq \sum_{j=2}^{n} \left(p(G_{j-1}) \sum_{x_j \in T_{p,\frac{2\ln j}{j-1}}} p(x_j) \log \frac{p(x_j)}{q_1(x_j)} + Rp(B_{j-1}) \right)$$
$$\leq \sum_{j=2}^{n} \sum_{x_j \in T_{p,\frac{2\ln j}{j-1}}} p(x_j) \log \frac{p(x_j)}{q_1(x_j)} + \frac{R}{j \ln j}.$$

As before, the last inequality is from Lemma 4. Again, from Condition (3), we have:

$$\lim_{j\to\infty} \left(\sup_{p\in\mathcal{P}} \sum_{x_j \in T_{p,\frac{2\ln j}{j-1}}} p(x_j) \log \frac{p(x_j)}{q_1(x_j)} + \frac{R}{j \ln j} \right) = 0.$$

Therefore, as before:

$$\lim_{n\to\infty} \sup_{p\in\mathcal{P}} \frac{1}{n} \left(\sum_{j=1}^{n} \sum_{x_j \in T_{p,\frac{2\ln j}{j-1}}} p(x_j) \log \frac{p(x_j)}{q_1(x_j)} + \sum_{j=2}^{n} \frac{R}{j \ln j} \right) = 0$$

as well. The theorem follows. □

A few comments about (3) in Theorem 5 are in order. Neither condition automatically implies the other. The set $\mathcal{B}$ of distributions in Section 4.1 is an example where every distribution has finite entropy, the redundancy of $\mathcal{B}$ is finite,

$$\lim_{\delta\to 0} \sup_{p\in\mathcal{B}} \sum_{x\in T_{p,\delta}} p(x) \log \frac{1}{p(x)} = 0 \quad \text{but } \forall q \text{ over } \mathbb{Z}_+ \text{ s.t.} \quad \lim_{\delta\to 0} \sup_{p\in\mathcal{B}} \sum_{x\in T_{p,\delta}} p(x) \log \frac{p(x)}{q(x)} > 0.$$

We will now construct another set $\mathcal{U}$ of distributions over $\mathbb{Z}_+$, such that every distribution in $\mathcal{U}$ has finite entropy, the redundancy of $\mathcal{U}$ is finite,

$$\lim_{\delta\to 0} \sup_{p\in\mathcal{U}} \sum_{x\in T_{p,\delta}} p(x) \log \frac{1}{p(x)} > 0 \quad \text{but } \exists q \text{ over } \mathbb{Z}_+ \text{ s.t.} \quad \lim_{\delta\to 0} \sup_{p\in\mathcal{U}} \sum_{x\in T_{p,\delta}} p(x) \log \frac{p(x)}{q(x)} = 0. \tag{4}$$

At the same time, the length n redundancy of $\mathcal{U}^\infty$ diminishes sublinearly. This is therefore also an example to show that the conditions in Theorem 5 are only sufficient, but, in fact, not necessary. It is yet open to find a condition on single-letter marginals that is both necessary and sufficient for the asymptotic per-symbol redundancy to diminish to zero.

Construction: $\mathcal{U}$ is a countable collection of distributions p_k, $k \in \mathbb{Z}_+$, on $\mathbb{N}$ where:

$$p_k(x) = \begin{cases} 1 - \frac{1}{k^2} & x = 0, \\ \frac{1}{k^2 2^{k^2}} & 1 \leq x \leq 2^{k^2}, \\ 0 & x > 2^{k^2}. \end{cases}$$ □

The entropy of $p_k \in \mathcal{U}$ is therefore $1 + h\big(\frac{1}{k^2}\big)$. Note that the redundancy of $\mathcal{U}$ is finite, too. To see this, first note that:

$$\sum_{x\in\mathbb{Z}_+} \sup_{k\in\mathbb{Z}_+} p_k(x) \le \sum_{x\in\mathbb{Z}_+} \sum_{p_k:k\in\mathbb{Z}_+} p_k(x) = \sum_{p_k:k\in\mathbb{Z}_+} \sum_{x\in\mathbb{Z}_+} p_k(x) = \sum_{p_k:k\in\mathbb{Z}_+} \frac{1}{k^2} = \frac{\pi^2}{6}. \tag{5}$$

Now, letting: $R^+ \overset{\text{def}}{=} \log\left(\sum_{x\in\mathbb{Z}_+} \sup_{k\in\mathbb{N}} p_k(x)\right)$, observe that the distribution:

$$q(x) = \begin{cases} 1/2 & x = 0, \\ \frac{\sup_{k\in\mathbb{Z}_+} p_k(x)}{2^{R^+ +1}} & x \in \mathbb{Z}_+. \end{cases}$$

satisfies for all $p_k \in \mathcal{U}$:

$$\sum_{x\in\mathbb{N}} p_k(x) \log \frac{p_k(x)}{q(x)} \le 1 + \frac{R^+ + 1}{k^2} \le R^+ + 2,$$

implying that the redundancy of $\mathcal{U}$ is $\le R^+ + 2$. Furthermore, Equation (5) implies that worst-case regret is finite, and from [2] the length n redundancy of $\mathcal{U}^\infty$ diminishes sublinearly. Now, pick an integer $m \in \mathbb{Z}_+$. We have for all $p \in \mathcal{U}$,

$$\sum_{x\in T_{p,\frac{1}{m^2 2^{m^2}}}} p(x) \log \frac{p(x)}{q(x)} \le \frac{R^+ + 1}{m^2},$$

yet, for all $k \ge m$, we have:

$$\sum_{x\in T_{p,\frac{1}{m^2 2^{m^2}}}} p_k(x) \log \frac{1}{p_k(x)} = 1.$$

Thus, the length n redundancy of $\mathcal{U}^\infty$ diminishes to zero, while not satisfying all of the requirements of Theorem 5. Therefore, the conditions of Theorem 5 are only sufficient, not necessary.

5. Open Problems

We have demonstrated that finite single-letter redundancy of a collection $\mathcal{P}$ of distributions over a countably infinite support does not imply that the asymptotic per-symbol redundancy of *i.i.d.* samples from $\mathcal{P}$ diminishes to zero. This is in contrast to the scenario for worst-case regret, where single-letter worst-case regret, being finite, is both necessary and sufficient for asymptotic per-symbol regret to diminish to zero. We have also demonstrated sufficient conditions on the collection $\mathcal{P}$, so that asymptotic per-symbol redundancy of *i.i.d.* samples diminish to zero in this paper. However, as we show, the sufficient conditions we provide are not necessary. It is yet open to find a condition on single-letter marginals that is both necessary and sufficient for the asymptotic per-symbol redundancy to diminish to zero.

Acknowledgments

This research was supported by National Science Foundation Grants CCF-1065632 and CCF-1018984.

Conflicts of Interest

The authors do not have any conflicts of interest.

References

1. Orlitsky, A.; Santhanam, N.P.; Zhang, J. Universal compression of memoryless sources over unknown alphabets. *IEEE Tran. Inf. Theory* **2004**, *50*, 1469–1481.
2. Boucheron, S.; Garivier, A.; Gassiat, E. Coding on countably infinite alphabets. **2008**, arXiv.org: 0801.2456.
3. Santhanam, N.; Anantharam, V.; Kavcic, A.; Szpankowski, W. Data driven weak universal redundancy. In Proceedings of 2014 IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014.
4. Merhav, N.; Feder, M. Universal prediction. *IEEE Tran. Inf. Theory* **1998**, *44*, 2124–2147.
5. Rosenthal, J.S. *A First Look at Rigorous Probability Theory*, 2nd ed.; World Scientific: Singapore, Singapore, 2008.
6. Santhanam, N. Probability Estimation and Compression Involving Large Alphabets. Ph.D. Thesis, University of California, San Diego, CA, USA, 2006.
7. Kingman, J.F.C. *The Mathematics of Genetic Diversity*; SIAM: Philadelphia, PA, USA, 1980.
8. Zabell, S.L. Predicting the unpredictable. *Synthese* **1992**, *90*, 205–232.
9. Zabell, S.L. The continuum of inductive methods revisited. In *The Cosmos of Science: Essays of Exploration*; Earman, J., Norton, J.D., Eds.; The University of Pittsburgh Press: Pittsburgh, PA, USA, 1997; Chapter 12.
10. Zabell, S.L. *Symmetry and Its Discontents: Essays on the History of Inductive Probability*; Cambridge Studies in Probability, Induction, and Decision Theory; Cambridge University Press: Cambridge, UK, 2005.
11. Santhanam, N.; Anantharam, V. Agnostic insurance of model classes. **2012**, arXiv.org: 1212:3866.
12. Orlitsky, A.; Santhanam, N. Lecture notes on universal compression. Available online: http://www-ee.eng.hawaii.edu/~prasadsn/ (accessed on 9 July 2014).