Tail Redundancy and Its Characterization of Compression of Memoryless Sources
Maryam Hosseini and Narayana Santhanam, Member, IEEE

Abstract—We formalize the tail redundancy of a collection \( P \) of distributions over a countably infinite alphabet, and show that this fundamental quantity characterizes the asymptotic per-symbol minimax redundancy of universally compressing sequences generated i.i.d. from \( P \). Contrary to the worst case formulations of universal compression, finite single letter minimax (average case) redundancy of \( P \) does not automatically imply that the expected minimax redundancy of describing length-\( n \) strings sampled i.i.d. from \( P \) grows sublinearly with \( n \). Instead, we prove that universal compression of length-\( n \) i.i.d. sequences from \( P \) is characterized by how well the tails of distributions in \( P \) can be universally described, showing that the asymptotic per-symbol redundancy of i.i.d. strings is equal to the tail redundancy.

Index Terms—Source coding, universal compression, minimax redundancy, tail redundancy.

I. INTRODUCTION

UNIVERSAL compression [2] captures the observation that it is often unreasonable to posit knowledge of the underlying probability law \( p \) generating data. Rather, one formalizes a setup where the generating probability law \( p \) is unknown, instead compressing data with just the knowledge that \( p \) belongs to a known collection \( P \) of probability laws, e.g., i.i.d. or Markov distributions. Since the generating law is unknown, we use a single universal probability law \( q \) for the collection \( P \) that, hopefully, simultaneously encodes as well as the underlying unknown \( p \) as closely as possible.

In the process, the idea is that the universal \( q \) and how well it encodes the data against the true probability laws, captured by metrics such as minimax redundancy should provide insights on how much information about the generating model we can glean from the data. We show that the asymptotic minimax redundancy of universally compressing memoryless sequences from a large, potentially countably infinite alphabet is captured by the complexity in the tails of the distributions, which this paper formalizes as the tail redundancy of a collection of distributions.

A countably infinite alphabet setup coupled with a finite number of observations often leads to the question of describing novelty, something that the observations have not yet revealed. The tail redundancy can be seen as another handle on this puzzle from an average case universal compression perspective, complementing the work done on the Good Turing estimators [3], which captures the worst case universal compression formulations [4].

a) Universal compression: Universal compression schemes are applied in several commercial compression algorithms, but as implied above, their theoretical underpinnings have implications beyond compression. Metrics used to quantify universal compression algorithms, in particular, redundancy, have interpretations that lend themselves to regularization [5], [6], quantifying the information content in observations about the generating source via the redundancy-capacity theorem [7], [8], [9], [10], [11], while their Bayesian formulations lead naturally to non-informative priors [12], [13], [14]. In addition, these metrics have been shown to capture the average reduction of the log compounded wealth in finance and gambling theory [15].

Different applications require different formalizations of redundancy, but this paper focuses on the average-case formulation. For a probability distribution \( q \), the redundancy incurred by a collection \( P \) of sources over an alphabet \( \mathcal{X} \) is the supremum over all sources \( p \in P \), of the expected excess code length of the universal scheme:

\[
\sup_{p \in P} \mathbb{E} \left[ \log \frac{1}{q(X)} - \log \frac{1}{p(X)} \right] = \sup_{p \in P} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},
\]

where the \( \mathcal{X} \)-valued random variable \( X \) above is distributed according to \( p \). Alternate worst case formulations measure the excess code length of a universal probability law \( q \) for the most inconvenient choice of source and data \( x \in \mathcal{X} \).

We focus now on the optimal value (1) can attain—the minimax redundancy which is the infimum of (1) over all possible distributions \( q \) over \( \mathcal{X} \). The minimax redundancy has an elegant statistical interpretation: the minimax redundancy is the capacity of the channel from \( P \) to \( \mathcal{X} \)—therefore, the amount of information the observation can provide about the model. This equivalence has been well known since [9], [11] when \( \mathcal{X} \) is finite, and was extended to arbitrary alphabets in [16].

Leaving formal definitions to Section III, the redundancy in (1) when \( \mathcal{X} \) is the set of all length-\( n \) sequences is termed as the length-\( n \) redundancy incurred by a distribution \( q \). We see this distribution \( q \) on length-\( n \) strings as the marginal
induced on length-\(n\) strings by a probability measure on infinite sequences. For such a probability measure (which we also denote by \(q\) below) on infinite sequences, the asymptotic per-symbol redundancy,
\[
\limsup_{n \to \infty} \frac{1}{n} \sum_{x \in \mathcal{X}^n} p(x) \log \frac{p(x)}{q(x)},
\]
captures the growth of the (normalized) length-\(n\) redundancy. This paper specifically addresses this question, characterizing the asymptotics of the redundancy of compressing sequences of symbols from a countable alphabet, generated i.i.d.

b) Countably infinite alphabets: The asymptotics when we have countably infinite supports help provide insights about scenarios when we have alphabets that are comparable or even exceed the sample length from which we are learning. This situation is not uncommon in language modeling, for example, and different scalings of the sample length and alphabet sizes have been considered in [17], [18], [19] to shed light on different nuances in this setup. One fundamental aspect of many of these problems, as mentioned before, is the aspect of describing novelty, i.e., something the finite observation may not have revealed about the source.

The tail redundancy is another hook to think about describing novelty, and indeed, it captures the rate at which the minimax redundancy scales asymptotically. That it should be so is satisfying intuitively, since it means that we conceptually divide up the description of sequences into two parts: a part that involves describing novel symbols that have never been seen prior (which is asymptotically the dominant contributor to the minimax redundancy) and the description of statistics of those symbols that we have already seen (which can be done efficiently). This is indeed reflected in the proofs as well.

A different issue arises from the fact that the asymptotic minimax redundancy of compression is not as well characterized in the countably infinite case. Starting from [2], the case where \(\mathcal{X}\) is a set of length-\(n\) sequences from a set of size \(k\), and \(\mathcal{P}\) is a collection of i.i.d. or Markov probability laws have been studied extensively. A cursory set of these papers include [14], [20], [21] for compression of i.i.d. sequences of sequences drawn from \(k\)-sized alphabets, [22], [23] for context tree sources, as well as extensive work involving renewal processes [24], [25], finite state sources [26], etc. For the infinite or large alphabet cases, the results that exist are not as comprehensive, though this has been approached from multiple angles. We describe these in the context of this paper in detail in Section II. This paper provides the leading term for the minimax redundancy in general for memoryless sequences.

The specific case, where the asymptotic per-symbol redundancy is zero is interesting as well. It is well understood that in this scenario, one can learn from the samples the (unknown) underlying probabilities of sequences. Asymmetric zero per-symbol redundancy is also sufficient to be able to learn the marginal distributions as well, though it is not necessary—see [27, Th. 9] for a full characterization of the learnability of the marginals in the i.i.d. case.

c) Summary: In this paper, we obtain a complete single letter characterization for the asymptotic per-symbol redundancy of length-\(n\) strings generated by i.i.d. sampling, and in the process also settle the open problem of when a class of i.i.d. sources can be compressed with asymptotically zero per-symbol redundancy. Section III defines the notion of redundancies formally, the connection between redundancy and tightness, and considers results connecting single letter and asymptotic per-symbol redundancy that will motivate the development of the paper. Section IV introduces the notion of tail redundancy that is central to the paper, and captures how much complexity lurks in tails of distributions:
\[
T(\mathcal{P}) \overset{\text{def}}{=} \liminf_{n \to \infty} \inf_{\mathcal{P}' \in \mathcal{P}} \frac{1}{n} \sum_{x \in \mathcal{X}^n} p(x) \log \frac{p(x)}{q(x)}.
\]
Section IV-A then sets about simplifying the above definition, showing in several steps that there is a single distribution \(q\) satisfying
\[
\limsup_{n \to \infty} \sup_{\mathcal{P} \in \mathcal{P} \cup \mathcal{P}' \in \mathcal{P}'} \frac{1}{n} \sum_{x \in \mathcal{X}^n} p(x) \log \frac{p(x)}{q(x)} = T(\mathcal{P}).
\]
We do so by first showing that the limsup in (2) can be replaced by a limit, that the limsup and inf can be interchanged (so it is unnecessary to consider a separate distribution for each \(m\)), and that the inf can be replaced by a minimization (there is a distribution that achieves the tail redundancy). Section IV-B develops some properties of the tail redundancy: (i) non-negativity, and (ii) the tail redundancy of finite unions equals the max of tail redundancies of the components of the union (unlike the single-letter redundancy of a union of distributions, but like the asymptotic per-symbol redundancy of unions of classes).

Section V uses the material developed to show our main result, that the asymptotic per-symbol redundancy of i.i.d. classes equals the tail redundancy of their marginals. That is, for all sets \(\mathcal{P}\) of distributions over \(\mathbb{N}\),
\[
\limsup_{n \to \infty} \inf_{\mathcal{P}' \in \mathcal{P}} \frac{1}{n} \sum_{x \in \mathcal{X}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = T(\mathcal{P}),
\]
where \(\mathcal{P}'\) is the collection of distributions on \(\mathbb{N}^\infty\) obtained as i.i.d. assignments from (i.e., products of) the distributions \(p \in \mathcal{P}\) and the infimum is over all distributions over \(\mathbb{N}^\infty\). Put another way, using the redundancy capacity theorem [9], [11], [16], [28], the tail redundancy tells us the rate at which we will keep learning new information about the underlying source from the data.

II. Prior Work

Consider a collection of all probability measures over a Borel sigma-algebra over infinite sequences of natural numbers obtained by i.i.d. sampling from a distribution in \(\mathcal{P}\) over \(\mathbb{N}\). A fundamental and natural question that has remained open on this topic is characterizing the per-symbol asymptotic minimax redundancy of compressing i.i.d. sequences. It is not even known what would be necessary and sufficient conditions on \(\mathcal{P}\) so that i.i.d. sequences from sources in \(\mathcal{P}\) could be compressed with asymptotically zero per-symbol minimax redundancy.

On the other hand, the question of what source classes can be compressed with asymptotically zero worst case per-symbol redundancy...
redundancy is settled in work by Boucheron et al. [29]. These are classes \( \mathcal{P} \) satisfying

\[
\inf_{q_n} \sup_{x^n \in [n]} \sup_{p \in \mathcal{P}} \frac{1}{n} \log \frac{p(x^n)}{q_n(x^n)} \to 0,
\]

where \( q_n, n \geq 1 \) are distributions over \( \mathbb{N}^n \) respectively, and \( p(X^n) \) is the \textit{i.i.d.} probability assignment on \( x^n \), i.e., \( p(x^n) = \prod_n p(x) \). The necessary and sufficient characterization for such classes is that \( \mathcal{P} \) has finite single letter worst case redundancy, i.e.,

\[
\inf_{q_n} \sup_{x^n \in [n]} \sup_{p \in \mathcal{P}} \log \frac{p(x)}{q(x)} < \infty.
\]

Of course, since the average case redundancy is always upper bounded by the worst case redundancy, we infer that finiteness of the worst case redundancy of single letter marginals is sufficient to guarantee asymptotically zero per-symbol average redundancy. On the other hand, finiteness of single worst case redundancy is not necessary, it is possible to construct an \textit{i.i.d.} class whose worst case redundancy of single letter marginals is infinite, yet the asymptotic per-symbol average redundancy is 0.

It is also helpful to compare with the single letter characterization of weak universality. A class of stationary ergodic sources \( \mathcal{P} \) over sequences of natural numbers is weakly compressible if there is a distribution \( q \) on naturals that incurs only a finite excess code length over any single letter marginal of \( \mathcal{P} \). Note that the excess code length of \( q \) is finite for each source, but need not be universally bounded over all sources. It is therefore tempting to consider whether finiteness of single letter average redundancy guarantees asymptotic zero per-symbol redundancy. Unfortunately, this need not be true either (Corollary 2 in Section III-B). While finite single letter average redundancy is necessary for zero asymptotic per-symbol redundancy, it is not sufficient.

From another direction, a series of results on grammar based codes [30], and in particular [31], have shown how to obtain encoding schemes for general stationary ergodic model classes that incur zero redundancy for all sources. Here the convergence is not necessarily uniform over the class as in the formulations we have considered, namely, there is no sup over \( p \) in (1)—rather convergence is considered pointwise for each \( p \in \mathcal{P} \). It is curious therefore, to see if results on grammar based codes may be extended to shed light on our problem. But as explained in the Appendix, approaches from [31] do not lend themselves to a sufficiency condition even for \textit{i.i.d.} classes.

In yet another direction, partial results were obtained by Haussler and Opper [16], who proved that if a class \( \mathcal{P} \) is not totally bounded in the Hellinger metric, then the class cannot incur asymptotically zero per-symbol redundancy. But as we will see, this result is also incomplete. We show in Proposition 6 in Section III-C that the partial result of [16] is only sufficient, not necessary. Specifically, we provide an example of a class bounded in the Hellinger metric but incurring asymptotically non-zero per-symbol redundancy.

III. Definitions and Background

The following development of universal compression is essentially standard. However, for formal simplicity in definitions we define redundancy and its asymptotics by means of measures over infinite sequences rather than sequences of distributions over various lengths.

Let \( N = 1, 2, \ldots \) be the set of naturals, \( \mathbb{N}^* \) be the collection of all finite strings of naturals, and let \( \mathbb{P}(\mathbb{N}) \ (\mathbb{P}(\mathbb{N}^*)) \) be the collection of all probability distributions over \( \mathbb{N} \ (\mathbb{N}^*) \) respectively. Let \( \mathcal{P} \subset \mathbb{P}(\mathbb{N}) \) be a collection of distributions over \( \mathbb{N} \) and \( \mathcal{P}^\infty \) be the set of distributions over length-\( n \) sequences, \( \mathbb{N}^n \), obtained via \textit{i.i.d.} assignments from marginals \( p \in \mathcal{P} \) (i.e., products distributions). For all \( p, \text{i.i.d.} \) assignments of probability measures for finite length strings can be naturally extended to a probability measure on the Borel sigma-algebra on the natural product topology in \( \mathbb{R}^\infty \), see, e.g., [32, Ch. 2].

Let \( \mathcal{P}^\infty \) be the collection of all such probability measures over infinite length sequences of \( \mathbb{N} \) obtained through the above construction. We use the same symbol \( p \) to indicate the probability measure in \( \mathcal{P}^\infty \), or its marginals—the distributions in \( \mathcal{P} \) or \( \mathcal{P}^n \).

We will use \( \mathbb{P}(\mathbb{N}^\infty) \) for the set of all probability measures over the Borel sigma-algebra on the product topology in \( \mathbb{R}^\infty \) (all sequential estimators as sometimes used in compression literature), and these probability measures can be specified uniquely by simply specifying the probabilities they assign on every finite string of naturals. The term sequential estimator in compression literature recalls the fact that the induced distributions on \( \mathbb{P}(\mathbb{N}^n) \) are consistent, and one can assign probabilities on any finite string via a sequence of conditional probabilities. Let \( q \) be an arbitrary (not necessarily \textit{i.i.d.}) probability measure in \( \mathbb{P}(\mathbb{N}^\infty) \). For all \( n, \) the redundancy of \( q \) against any \( p \in \mathcal{P}^\infty \) is

\[
R_n(p, q) = \sum_{x^n \in \mathbb{N}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = D_n(p||q),
\]

where \( D_n() \) above denotes the KL divergence between the distributions on \( \mathbb{N}^n \) induced by the measures \( p \) and \( q \) respectively. A collection \( \mathcal{P}^\infty \) is \textit{weakly compressible} if there exists a probability measure \( q \in \mathbb{P}(\mathbb{N}^\infty) \) such that for all \( p \in \mathcal{P}^\infty \)

\[
\lim_{n \to \infty} \frac{1}{n} R_n(p, q) = 0.
\]

A collection \( \mathcal{P}^\infty \) is \textit{strongly compressible} if there exists a probability measure \( q \in \mathbb{P}(\mathbb{N}^\infty) \) such that

\[
\lim_{n \to \infty} \sup_{p \in \mathcal{P}^\infty} \frac{1}{n} R_n(p, q) = 0.
\]

Define the length-\( n \) per-symbol redundancy of any \( q \in \mathbb{P}(\mathbb{N}^n) \) (or completely equivalently, \( q \in \mathbb{P}(\mathbb{N}^\infty) \)) against \( \mathcal{P}^\infty \) to be

\[
R_n(\mathcal{P}^\infty, q) = \sup_{p \in \mathcal{P}^\infty} \frac{1}{n} R_n(p, q).
\]

Since \( \mathbb{N}^\infty \) is not countable, for completeness sake we define an appropriate sigma-algebra that allows us to focus on probability assignments on finite strings in \( \mathbb{N}^* \) alone.

\[1\]All the logarithms are in base 2, unless otherwise specified.
and the length-$n$ per-symbol redundancy of $P^\infty$ to be

$$R_n(P^\infty) \overset{\text{def}}{=} \inf_q R_n(P^\infty, q) = \inf_{p \in P^n} \frac{1}{n} R_n(p, q)$$

(4)

where the infimum is taken over all $q \in P(N^\infty)$, or completely equivalently, $q \in P[N^n]$ (as is commonly done). For $n = 1$, the length-$1$ redundancy, $R_1(P^\infty)$ will be called the single-letter redundancy of $P$, and to emphasize the point, we will use $R_1(P)$ to denote the case $n = 1$. Finally, the asymptotic, per-symbol redundancy of $q \in P(N^\infty)$ against $P^\infty$, $R(P^\infty, q)$, as well as the asymptotic, per-symbol redundancy of $P^\infty$, $R(P^\infty)$, (see discussion below and in Appendix A) to be

$$R(P^\infty, q) \overset{\text{def}}{=} \limsup_{n \to \infty} \frac{1}{n} R_n(p, q)$$

and

$$R(P^\infty) = \inf_{q \in P[N^\infty]} R(P^\infty, q)$$

(5)

which, as shown in Appendix A to also satisfy

$$R(P^\infty) = \limsup_{n \to \infty} \frac{1}{n} R_n(p, q) = \limsup_{n \to \infty} R_n(p, q)$$

(6)

The definition of asymptotic-per symbol redundancy is a minor technical departure from some of prior literature, but is completely equivalent while clarifying the following. In certain prior expositions, as in the equation (6) above, the infimum over $q$ does not enforce that the length $n$ distributions for all $n$ be consistent, i.e., marginals of the same probability measure. Instead, there is seemingly extra freedom allowed, where different (potentially inconsistent) probability distributions $\{q_n \in P[N^n] : n \geq 1\}$ could be chosen for different $n$. The seeming additional "freedom" in allowing potentially inconsistent probability distributions is a red herring, and does not yield any actual advantage, something automatically clarified by the definition in (5). See Appendix A.

We will need the following elementary results on the redundancy.

**Proposition 1:** For all $P$ and all numbers $n \geq 1$, $\frac{1}{n} R_n(P^\infty) \leq R_1(P)$, and therefore $R(P^\infty) \leq R_1(P)$.

**Proof:** See [27, Proposition 38] for a proof that for all $n \geq 1$, $\frac{1}{n} R_n(P^\infty) \leq R_1(P)$. The proposition then follows from Appendix A that proves that $R(P^\infty) = \limsup_{n \to \infty} \frac{1}{n} R_n(P^\infty)$.

**Proposition 2:** For all $P$, $R(P^\infty) < \infty$ if $R_1(P) < \infty$.

**Proof:** See [27, Corollary 39] for a proof.

**A. Tightness**

A collection $P \subset P(N)$ of distributions on $N$ is defined to be **tight** if for all $\gamma > 0$, there is a number $N_\gamma$ such that

$$\sup_{p \in P} P(X_p > N_\gamma) < \gamma$$

where $X_p$ above is a random variable distributed according to $p$.

**Lemma 1:** Let $P \subset P(N)$ be a class of distributions on $N$ with finite single letter redundancy, namely $R_1 < \infty$. Then $P$ is tight.

**Proof:** This is a well known result, see [1], [16], [27] for three separate proofs.

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

**Example 1:** Consider the following collection $I$ of distributions over $N$. First partition the set of naturals into the sets $T_i$, $i \in N$, where

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

Note that $|T_i| = 2^i$. Now, $I$ is the collection of all possible distributions that can be formed as follows—for all $i \in Z_+$, pick exactly one element of $T_i$ and assign probability $1/(i+1)(i+2)$ to the element of $T_i$ chosen. Note that the set $I$ is uncountably infinite. We will show in the following Corollary and Proposition that $I$ is tight, but does not have finite redundancy.

**Corollary 1:** The set $I$ of distributions is tight.

**Proof:** For all $p \in I$,

$$\sum_{x \geq 2^k} p(x) \log \frac{p(x)}{p(x_0)}$$

namely, all tails are uniformly bounded over the collection $I$.

Put another way, for all $\delta > 0$ and all distributions $p \in I$,

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

On the other hand, **Proposition 3:** The collection $I$ does not have finite redundancy.

**Proof:** Suppose $q$ is any distribution over $Z_+$. We will show that $\exists p \in I$ such that $\sum_{x \in Z_+} p(x) \log \frac{p(x)}{q(x)}$ is not finite. Since the entropy of every $p \in I$ is finite, we just have to show that for any distribution $q$ over $Z_+$, there $\exists p \in I$ such that $\sum_{x \in Z_+} p(x) \log \frac{p(x)}{q(x)}$ is not finite.

Consider any distribution $q$ over $Z_+$. Observe that for all $i$, $|T_i| = 2^i$. It follows that for all $i$ there is $x_i \in T_i$ such that

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

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

$$p^*(x_i) = \frac{1}{(i+1)(i+2)} \quad \forall i \in Z_+.$$

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

**B. Single Letter and Asymptotic Per-Symbol Redundancy**

Although arbitrary collections of stationary ergodic distributions over finite alphabets are weakly compressible, Kieffer [33] showed the collection of all i.i.d. distributions over $N$ is not even weakly compressible. Indeed, here the finiteness of single letter redundancy is sufficient for weak compressibility. Any collection of stationary ergodic measures over infinite sequences is weakly compressible if $R_1 < \infty$.

$R_1$ being finite, however, is not sufficient for strong compression guarantees to hold even when while dealing with i.i.d.
We reproduce the following Example 2 from [34] to illustrate the pitfalls with strong compression, and to motivate the notion of tail redundancy that will be central to our main result. Proposition 4 shows that the collection in the Example below has finite single letter redundancy, but Proposition 5 shows that its length \( n \) redundancy does not diminish to zero as \( n \to \infty \).

**Example 2:** Partition the set \( \mathbb{N} \) into \( T_i = \{2^i, \ldots, 2^{i+1} - 1\} \), \( i \in \mathbb{N} \). Recall that \( T_i \) has 2\( i \) elements. For all \( n \geq 1 \), let \( 1 \leq j \leq 2^n \) and let \( p_{n,j} \) be a distribution on \( \mathbb{N} \) that assigns probability \( 1 - \frac{1}{n} \) to the number 1 (or equivalently, to the set \( T_0 \)), and \( \frac{1}{n} \) to the \( j \)th smallest element of \( T_n \), namely the number \( 2^n + j - 1 \). \( B \) (mnemonic for binary, since every distribution has at support of size 2) is the collection of distributions \( p_{n,j} \) for all \( n > 0 \) and \( 1 \leq j \leq 2^n \). \( B^\infty \) is the set of measures over infinite sequences of numbers corresponding to i.i.d. sampling from \( B \).

We first verify that the single letter redundancy of \( B \) is finite.

**Proposition 4:** Let \( q \in \mathcal{P}(\mathbb{N}) \) 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

\[
\sup_{p \in B} \sum_{x \in \mathbb{N}} p(x) \log \frac{p(x)}{q(x)} \leq 2.
\]

However, the redundancy of compressing length-\( n \) sequences from \( B^\infty \) scales linearly with \( n \).

**Proposition 5:** For all \( n \in \mathbb{N} \),

\[
\inf \sup_{q \in \mathcal{P} B^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} \geq 1 - \frac{1}{e} - \frac{1}{n} \ln \left(\frac{1}{e}\right),
\]

where (\( x \) is a fixed value for which \( \ln x \) is defined) is the binary entropy function and the infimum is over all distributions over \( \mathbb{N}^n \).

**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 \( \mathbb{N}^n \), 1 \( \leq i \leq 2^n \), where

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

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

\[
p_{1,j}(S_i) = 1 - \left(1 - \frac{1}{n}\right)^n > 1 - \frac{1}{e}.
\]

Since there are 2\( n \) pairwise disjoint sets \( S_i \), no matter what the universal distribution \( q \) over length-\( n \) sequences, there must be a set \( S_j \) such that

\[
q(S_j) \leq 2^{-j}.
\]

Therefore, the redundancy incurred by \( q \) against \( p_{1,j}^{1/n} \) is

\[
E_{p_{1,j}} \log \frac{p_{1,j}(X^n)}{q(X^n)} \geq p_{1,j}(S_i) \log \frac{p_{1,j}(S_i)}{q(S_i)} + p_{1,j}(S_j) \log \frac{p_{1,j}(S_j)}{q(S_j)}
\]

which is in turn lower bounded by

\[
\left(1 - \frac{1}{e}\right) \log 2^n - h\left(\frac{1}{e}\right) = n\left(1 - \frac{1}{e}\right) - h\left(\frac{1}{e}\right).
\]

We will see later that the asymptotic per-symbol redundancy of \( B^\infty \) equals 1.

**Corollary 2:** There exists a collection of distributions \( \mathcal{P} \in \mathcal{P}(\mathbb{N}) \) with finite single letter redundancy, yet \( \mathcal{P}^\infty \), the set of i.i.d. processes with single letter marginals from \( \mathcal{P} \), has asymptotic per-symbol redundancy bounded away from 0.

**Proof:** \( B \) from Example 2 is one such class.

It is instructive to compare what happens when we try to describe a single digit output from an unknown distribution in \( B \). An observation we make is that there is no number \( m \) such that some universal distribution over \( \mathbb{N} \) describes numbers \( \geq m \) as well as the best distribution in \( B \), in the sense the redundancy incurred is always bounded below by 1, no matter how large \( m \) is. This is reflected when compressing strings of length \( n \)—we incur a heavy penalty against those distributions that contain an element with probability \( O(\frac{1}{n}) \).

It is a different set of distributions that hit us for different sequence lengths, and this does not stop no matter how large the sequence length becomes. This is the essence of the problem in Proposition 5, and what motivates our definition of tail redundancy in Section IV.

### C. Asymptotic Zero Per-Symbol Redundancy and Boundedness in Hellinger Metric

To connect single letter redundancy to length-\( n \) redundancy, the authors in [16] obtain partial lower and upper bounds on the asymptotic per-symbol redundancy using the total-boundedness of the probability set under the Hellinger metric. This work perhaps comes closest to our results. For this reason, we present both the result as well as why the result is not a full characterization in this Section.

**Definition 1 (Hellinger Distance):** Let \( p_1 \) and \( p_2 \) be two distributions in \( \mathcal{P}(\mathbb{N}) \). The Hellinger distance \( h \) is defined as

\[
h^2(p_1, p_2) = \frac{1}{2} \sum_{x \in \mathbb{N}} \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2.
\]

**Definition 2 (Totally Bounded Set [16]):** Let \( (S, \rho) \) be any complete separable metric space. A partition \( \Pi \) of set \( S \) is a collection of disjoint Borel subsets of \( S \) such that their union is \( S \). Then diameter of a subset \( A \subset S \) is \( d(A) = \sup_{x,y \in A} \rho(x,y) \) and diameter of partition \( \Pi \) is supremum of diameters of the sets in the partition. For \( \delta > 0 \), let \( D_\delta(S, \rho) \) be the cardinality of the smallest finite partition of \( S \) of diameter \( \leq \delta \). We say \( S \) is totally bounded if \( D_\delta(S, \rho) < \infty \) for all \( \delta > 0 \).

**Lemma 2 [16]:** If length-\( n \) redundancy is finite it can grow at most linearly in \( n \). If \( \mathcal{P}, h \) is not totally bounded in the Hellinger metric and single letter redundancy is finite then \( \limsup_{n \to \infty} \frac{1}{n} R_n(\mathcal{P}^n) \) is bounded away from zero and \( \liminf_{n \to \infty} \frac{1}{n} R_n(\mathcal{P}^n) < \infty \).

**Proof:** See [16, part 5, Th. 4].

The above is not a complete characterization of when the asymptotic per-symbol redundancy is bounded away from 0, and the converse of the Lemma above does not hold. For example, recall the collection \( B \) from Example 1. We show below that \( B \) is a counter-example that proves that the converse of the above lemma cannot hold.
Proposition 6: The collection \((B, h)\) is totally bounded in the Hellinger metric, but
\[
\liminf_{n \to \infty} \frac{1}{n} R_n(B^n) > 0.
\]

Proof: To see \(B\) is totally bounded in the Hellinger metric, consider the following partition for \(\delta > 0\). Let \(N = \lceil \frac{1}{2\delta} \rceil\) and we partition \(B\) into \(\leq N^2 + 1\) parts, each with diameter \(\leq \delta\). All but the last part contains exactly one distribution each among \(p_{n,j}\) where \(n \leq N\) and \(1 \leq j \leq 2^n\) (Therefore, \(\leq N^2\) parts). The last part of the partition contains all the other distributions of \(B\). The diameter of all but the last part is 0 (since they are sets with only one distribution each). We bound the Hellinger distance between any two distributions in the last part by noticing that for numbers \(n_1\) and \(n_2\) both \(> N\),
\[
\left( \sqrt{1 - \frac{1}{n_1}} - \sqrt{1 - \frac{1}{n_2}} \right)^2 + \frac{1}{n_1} + \frac{1}{n_2} \leq 3 \frac{1}{N},
\]
which in turn implies that the Hellinger distance between any pair of distributions in the last part is bounded by \(\sqrt{\frac{1}{2N}} \leq \delta\) by our choice of \(N\). The redundancy result follows from Proposition 5.

In the development below, we give a complete characterization of the asymptotic per-symbol redundancy.

IV. TAIL REDUNDANCY

We will develop a series of tools that will help us better understand how the per-symbol redundancy behaves in a wide range of large alphabet cases. In particular, for i.i.d. sources, we completely characterize the asymptotic per-symbol redundancy in terms of single letter marginals. Fundamental to our analysis is the understanding of how much complexity lurks in the tails of distributions.

To this end, we define what we call the tail redundancy. We assert the basic definition below, but simplify several nuances around it in Section IV-A, eventually settling on a operationally workable characterization.

Definition 3: For a collection \(\mathcal{P}\) of distributions, define for all \(m \geq 1\)
\[
T_m(\mathcal{P}) \overset{\text{def}}{=} \inf_{q \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)},
\]
where the infimum is over all distributions \(q\) over \(\mathbb{N}\). We define the tail redundancy as
\[
\mathcal{T}(\mathcal{P}) \overset{\text{def}}{=} \limsup_{m \to \infty} T_m(\mathcal{P}).
\]
The above quantity, \(T_m(\mathcal{P})\) can be negative, and is not then redundancy of any collection of distributions as is conventionally understood. However, let \(S_m = \{x \in \mathbb{N} : x \geq m\}\). Then \(\overline{T}_m(\mathcal{P})\), which is defined as
\[
\inf_{q \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \left( \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} + p(S_m) \log \frac{1}{p(S_m)} \right)
\]
is always non-negative, and can be phrased in terms of a conventional redundancy. To see this, let \(p'\) be the distribution over numbers in \(S_m\) obtained from \(p\) as \(p'(x) = p(x)/p(S_m)\), and note that
\[
\overline{T}_m(\mathcal{P}) = \inf_{q \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)}.
\]

While we unravel the above definitions in detail in the next section, we first note that, as with redundancy, only tight classes can possibly have finite tail redundancy. In general if the single letter redundancy is infinite, so is the tail redundancy.

Proposition 7: For \(\mathcal{P} \subseteq \mathcal{P}(\mathbb{N})\), if \(R_1(\mathcal{P}) = \infty\), then for all \(m \in \mathbb{N}\) and for all distributions \(q \in \mathcal{P}(\mathbb{N})\),
\[
\sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} = \infty.
\]
and therefore, \(\mathcal{T}(\mathcal{P}) = \infty\). In particular if \(\mathcal{P}\) is not tight, \(\mathcal{T}(\mathcal{P}) = \infty\).

Proof: Suppose \(R_1(\mathcal{P}) = \infty\), and there exists \(m \in \mathbb{N}\) and a distribution \(q_m \in \mathcal{P}(\mathbb{N})\) and some \(M < \infty\) such that
\[
\sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q_m(x)} = M.
\]
Consider the distribution \(q_1 \in \mathcal{P}(\mathbb{N})\) that assigns probability \(1/(m-1)\) for all numbers from 1 through \(m-1\). Then the distribution \(q = (q_1 + q_m)/2\) satisfies
\[
\sup_{p \in \mathcal{P}} \sum_{x \in \mathbb{N}} p(x) \log \frac{p(x)}{q(x)} \leq M + \log(m-1) + 1,
\]
a contradiction that \(R_1(\mathcal{P}) = \infty\). Therefore, we can also conclude that \(\overline{T}_m(\mathcal{P}) = \infty\) for all \(m\), and therefore \(\mathcal{T}(\mathcal{P}) = \infty\).

For the last part, if \(\mathcal{P}\) is not tight, Lemma 1 implies that the single letter redundancy is infinite, and therefore the tail redundancy is infinite as well.

We add a comment about Kieffer’s characterization of weak universality, i.e., classes \(\mathcal{P}\) admitting a universal measure \(q\) such that
\[
\sup_{p \in \mathcal{P}} \limsup_{n \to \infty} \frac{1}{n} R_n(p, q) = 0.
\]
Kieffer characterized that a stationary ergodic class \(\mathcal{P}\) is weakly universal iff it admits a distribution \(q\) satisfying for all \(p \in \mathcal{P}\),
\[
E \log \frac{1}{q(X)} < \infty\] (note however \(\sup_p E \log \frac{1}{q(X)}\) could be infinite). Note the distinction from our case, we are characterizing strong compression, namely
\[
\limsup_{n \to \infty} \frac{1}{n} R_n(p, q) = 0.
\]
Therefore, Kieffer’s result, even restricted to i.i.d. classes will not imply the tail redundancy development here. Characterization of the strong result beyond i.i.d. classes remains open at this point.

A. Operational Characterization of Tail Redundancy

We refine the above definitions in several ways. First we prove that the sequence \(\overline{T}_m\) always has a limit and
\[
\mathcal{T}(\mathcal{P}) = \lim_{m \to \infty} \overline{T}_m(\mathcal{P})
\]
\[
= \lim_{m \to \infty} \inf_{q_m \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q_m(x)}. \quad (7)
\]
Next, we show that the limit and inf above can be interchanged, and in addition, that a minimizer exists—namely there is always a distribution over $\mathbb{N}$ that achieves the tail redundancy. This will let us operationally characterize the notions in the definitions above.

**Lemma 3:** $\hat{T}_m(\mathcal{P})$ is non-increasing in $m$.

**Proof:** Let $q$ be any distribution over $\mathbb{N}$ and as before, $S_m = \{x \in \mathbb{N} : x \geq m\}$. We show that

$$\sup_{p \in \mathcal{P}} \left( \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} + p(S_m) \log \frac{1}{p(S_m)} \right) \geq \hat{T}_{m+1}(\mathcal{P}),$$

thus proving the lemma.

To proceed, note that without loss of generality we can assume $\sum_{x \geq m} q_m(x) = 1$. For $x \geq m + 1$, let

$$q(x) = \frac{q_m(x)}{\sum_{x \geq m+1} q_m(x)} = \frac{q_m(x)}{1 - q_m(m)}.$$  \hspace{1cm} (8)

We have

$$\hat{T}_m = \sup_{p \in \mathcal{P}} \left( \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} + p(S_m) \log \frac{1}{p(S_m)} \right)$$

$$= \sup_{p \in \mathcal{P}} \left[ p(m) \log \frac{p(m)}{q(m)} + p(S_{m+1}) \log \frac{1}{1 - q_m(m)} \right. \right.$$

$$\left. + \sum_{x \geq m+1} p(x) \log \left( \frac{p(x)}{q(x)} \right) + p(S_m) \log \left( \frac{1}{p(S_m)} \right) \right]$$

$$\geq \hat{T}_{m+1}(\mathcal{P}),$$

where in (a), we use (8) for $x \geq m + 1$, in (b) we absorb the last term into the first two terms of the prior equation, noting that $p(S_m) - p(m) = p(S_{m+1})$, and in (c), the KL divergence term denotes the divergence between two Bernoulli random variables with parameters $\frac{p(m)}{p(S_m)}$ and $q_m(m)$ respectively. The last inequality follows from the non-negativity of KL-divergence and because

$$\sup_{p \in \mathcal{P}} \left[ \sum_{x \geq m+1} p(x) \log \left( \frac{p(x)}{q(x)} \right) + p(S_m) \log \left( \frac{1}{p(S_m)} \right) \right] \geq \hat{T}_{m+1}(\mathcal{P}).$$

**Corollary 3:** For all collections $\mathcal{P} \in \mathcal{P}(\mathbb{N})$, $\lim_{m \to \infty} \hat{T}_m(\mathcal{P})$ exists.

**Proof:** From Lemma 3 and the fact that $\hat{T}_m(\mathcal{P}) \geq 0$ for all $m$.

**Lemma 4:** For all collections $\mathcal{P} \in \mathcal{P}(\mathbb{N})$, the limit $\lim_{m \to \infty} T_m(\mathcal{P})$ exists and hence

$$T(\mathcal{P}) = \lim_{m \to \infty} T_m(\mathcal{P}) = \lim_{m \to \infty} \hat{T}_m(\mathcal{P}).$$

**Proof:** If $\mathcal{P}$ is not tight, the lemma holds vacuously from Proposition 7. Therefore, we suppose in the rest of the proof that $\mathcal{P}$ is tight. Observe from the definitions that

$$T_m(\mathcal{P}) \leq \hat{T}_m(\mathcal{P}).$$

Let $S_m = \{x \geq m\}$ as before and let $q$ be any distribution over $\mathbb{N}$. Then

$$\sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)}$$

$$\geq \sup_{p \in \mathcal{P}} \left( \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} + p(S_m) \log \frac{1}{p(S_m)} \right)$$

$$+ \inf_{\hat{p} \in \mathcal{P}} \max_{\hat{p} \in \mathcal{P}} \hat{p}(S_m) \log \hat{p}(S_m)$$

$$\geq \inf_{q \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \left( \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)} + p(S_m) \log \frac{1}{p(S_m)} \right)$$

$$+ \inf_{\hat{p} \in \mathcal{P}} \hat{p}(S_m) \log \hat{p}(S_m)$$

$$= \hat{T}_m(\mathcal{P}) + \inf_{\hat{p} \in \mathcal{P}} \hat{p}(S_m) \log \hat{p}(S_m)$$

Since $\mathcal{P}$ is tight, $\sup_{p \in \mathcal{P}} p(S_m) \to 0$ as $m \to \infty$ and hence $\inf_{\hat{p} \in \mathcal{P}} \hat{p}(S_m) \log \hat{p}(S_m) \to 0$ as $m \to \infty$. From Corollary 3, we know that the sequence $\{\hat{T}_m(\mathcal{P})\}$ has a limit. Therefore, the sequence $T_m(\mathcal{P})$ also has a limit and in particular we conclude

$$T(\mathcal{P}) = \lim_{m \to \infty} T_m(\mathcal{P}) = \lim_{m \to \infty} \hat{T}_m(\mathcal{P}).$$

Therefore, taking into account the above lemma, we can rephrase the definition of tail redundancy as in (7),

$$T(\mathcal{P}) \equiv \lim_{m \to \infty} \inf_{q \in \mathcal{P}(\mathbb{N})} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)}.$$

We now show that

$$T(\mathcal{P}) = \min_{q \in \mathcal{P}(\mathbb{N})} \lim_{m \to \infty} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)}.$$

Note that the limit above need not be finite for every $q$. We take the above equation to mean the minimization over all $q$ such that the limit exists. If no such $q$ exists, the term on the right is considered to be vacuously infinite.

**Lemma 5:** For a collection $\mathcal{P}$ of distributions over $\mathbb{N}$ with tail redundancy $T(\mathcal{P})$, there is a distribution $q^*$ over $\mathbb{N}$ that satisfies

$$\lim_{m \to \infty} \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)} = T(\mathcal{P}).$$

**Proof:** If $\mathcal{P}$ is not tight, the lemma is vacuously true and any $q$ is a “minimizer”.



Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Therefore, we suppose in the rest of the proof that \( \mathcal{P} \) is tight. From Lemma 1, we can pick a finite number \( m_r \) such that
\[
\sup_{p \in \mathcal{P}} p(x \geq m_r) \leq \frac{1}{2^r},
\]
and let \( q_r \) be any distribution that satisfies
\[
\sup_{p \in \mathcal{P}} \sum_{x \geq m_r} p(x) \log \frac{p(x)}{q_r(x)} \leq T_m(\mathcal{P}) + \frac{1}{r}.
\]
Since the limit of \( T_m(\mathcal{P}) \) as \( m \to \infty \) is \( T(\mathcal{P}) \), we have
\[
\lim_{r \to \infty} \sup_{p \in \mathcal{P}, x \geq m_r} p(x) \log \frac{p(x)}{q_r(x)} = T(\mathcal{P}).
\]
Following a well-known approach where we average distributions, take
\[
q^*(x) = \sum_{r \geq 1} \frac{q_r(x)}{r(r+1)}.
\]
Now we also have for \( r \geq 2 \) and any \( m_r < m < m_{r+1} \) that
\[
\sup_{p \in \mathcal{P}} \sum_{x \geq m_r} p(x) \log \frac{p(x)}{q^*(x)} \geq \sup_{p \in \mathcal{P}} \left( p(m_r \leq x < m) \log \frac{p(m_r \leq x < m)}{q^*(m_r \leq x < m)} + \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)} \right) \geq \sup_{p \in \mathcal{P}} \left( p(m_r \leq x < m) \log p(m_r \leq x < m) + \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)} \right) \geq -\frac{r}{2} + \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)},
\]
where the first inequality is the logsum inequality and the last inequality follows because \( p(m_r \leq x < m) \leq p(m_r \leq x) \leq \frac{1}{2^r} < \frac{1}{r} \) for \( r \geq 2 \). Similarly, for \( r \geq 2 \) and \( m_r < m < m_{r+1} \), we have
\[
\sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)} \geq \frac{r}{2} + \sup_{p \in \mathcal{P}} \sum_{x \geq m} p(x) \log \frac{p(x)}{q^*(x)}.
\]
Therefore,
\[
\lim_{m \to \infty} \sup_{p \in \mathcal{P}, x \geq m} p(x) \log \frac{p(x)}{q^*(x)} \leq \lim_{r \to \infty} \left[ \sup_{p \in \mathcal{P}, x \geq m_r} p(x) \log \frac{p(x)}{q_r(x)} + \frac{r}{2^r} + \log \frac{r(r+1)}{2^r} \right] = T(\mathcal{P}).
\]
Similarly,
\[
\lim_{m \to \infty} \sup_{p \in \mathcal{P}, x \geq m} p(x) \log \frac{p(x)}{q^*(x)} \geq T(\mathcal{P}),
\]
and the lemma follows.

Henceforth, we will describe any distribution \( q^* \) as in the lemma above as “\( q \) achieves the tail redundancy for \( \mathcal{P} \)”.

**Corollary 4:** If a collection \( \mathcal{P} \) of distributions is tight and has tail redundancy \( T(\mathcal{P}) \), then there is a distribution \( q^* \) over \( \mathbb{N} \) that satisfies
\[
\lim_{m \to \infty} \sup_{p \in \mathcal{P}, x \geq m} p(x) \log \frac{p(x)}{q^*(x)} = T(\mathcal{P}).
\]

**Proof:** The result follows using Lemma 5 and the fact that \( \mathcal{P} \) is tight. ■

**B. Properties of the Tail Redundancy**

We examine two properties of tail redundancy in this subsection. Note that the tail redundancy \( T(\mathcal{P}) \) is defined as the limit of \( T_m(\mathcal{P}) \) as \( m \to \infty \), however \( T_m(\mathcal{P}) \) need not always be non-negative. However, we show that \( T(\mathcal{P}) \) is always non-negative. The second property concerns the behavior of tail redundancy across finite unions of classes. This property, while interesting inherently, also helps us cleanly characterize the per-symbol redundancy of i.i.d. sources in Section V.

**Lemma 6:** For all \( \mathcal{P} \), \( T(\mathcal{P}) \geq 0 \).

**Proof:** Follows directly from Lemma 4, and from the fact that \( T_m(\mathcal{P}) \geq 0 \) for all \( m \).

We now show that the tail redundancy of a finite union of collections equals the maximum of the tail redundancies of the individual parts of the union.

**Lemma 7:** Let \( \mathcal{T}(\mathcal{P}_1), \mathcal{T}(\mathcal{P}_2), \ldots, \mathcal{T}(\mathcal{P}_k) \) be tail redundancy of collections \( \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_k \) respectively. Then
\[
\mathcal{T} \left( \bigcup_{i=1}^k \mathcal{P}_i \right) = \max_{1 \leq j \leq k} \mathcal{T}(\mathcal{P}_j).
\]

**Proof:** We first observe \( \mathcal{T} \left( \bigcup_{i=1}^k \mathcal{P}_i \right) \geq \max_{1 \leq j \leq k} \mathcal{T}(\mathcal{P}_j) \), since for all \( q \in \mathcal{P}(\mathbb{N}) \), and all \( 1 \leq j \leq k \), we have
\[
\sup_{p \in \bigcup_{i=1}^k \mathcal{P}_i, x \geq m} p(x) \log \frac{p(x)}{q(x)} \geq \sup_{p \in \mathcal{P}_j, x \geq m} p(x) \log \frac{p(x)}{q(x)} \geq \inf_{p \in \mathcal{P}_j} \sup_{x \geq m} p(x) \log \frac{p(x)}{q(x)} = T_m(\mathcal{P}_j)
\]
To show that \( \mathcal{T} \left( \bigcup_{i=1}^k \mathcal{P}_i \right) \leq \max_{1 \leq j \leq k} \mathcal{T}(\mathcal{P}_j) \), let \( \mathcal{q}_1, \mathcal{q}_2, \ldots, \mathcal{q}_k \) be distributions that achieve the tail redundancies \( \mathcal{T}(\mathcal{P}_1), \mathcal{T}(\mathcal{P}_2), \ldots, \mathcal{T}(\mathcal{P}_k) \) respectively. Furthermore, for all distributions \( q \in \mathcal{P}(\mathbb{N}) \) and collections of distributions \( \mathcal{P} \subseteq \mathcal{P}(\mathbb{N}) \), let
\[
T_m(\mathcal{P}, q) = \sup_{p \in \mathcal{P}, x \geq m} p(x) \log \frac{p(x)}{q(x)}.
\]
Clearly, we have
\[
\mathcal{T} \left( \bigcup_{i=1}^k \mathcal{P}_i \right) = \lim_{m \to \infty} \inf_{q} T_m \left( \bigcup_{i=1}^k \mathcal{P}_i, q \right).
\]
Let
\[
\hat{q}(x) = \frac{\sum_{i=1}^k q_i(x)}{k}
\]
for all $x \in \mathbb{N}$. We will attempt to understand the behavior of the sequence $T_m(\bigcup_{i=1}^k \mathcal{P}_i, \hat{q})$ first. Observe that

$$T_m(\bigcup_{i=1}^k \mathcal{P}_i, \hat{q}) = \max_{1 \leq j \leq k} \sup_{p \in \mathcal{P}_j} \sum_{x \geq m} p(x) \log \frac{p(x)}{q(x)}$$

$$= \max_{1 \leq j \leq k} T_m(\mathcal{P}_j, \hat{q}). \quad (9)$$

For all $1 \leq j \leq k$, the limit $\lim_{m \to \infty} T_m(\mathcal{P}_j, \hat{q})$ exists and both limits are strongly compressible.

Theorem 1: Let $\mathcal{P}$ be a collection of distributions over $\mathbb{N}$ and $\mathcal{P}_\infty$ be the collection of all measures over infinite sequences that can be obtained by i.i.d. sampling from a distribution in $\mathcal{P}$. Then

$$R(\mathcal{P}_\infty) = \limsup_{n \to \infty} \frac{1}{n} R_n(\mathcal{P}_\infty) = T(\mathcal{P}).$$

A couple of quick examples first. Note that any class $\mathcal{P}$ over a finite alphabet has tail redundancy 0, and therefore, the asymptotic per-symbol redundancy for any such class is 0 as is well known.

Example 3: Proposition 5 proved that $\mathcal{B}_\infty$ does not have zero asymptotic per-symbol redundancy. We show that $T(\mathcal{B}) = 1$. We first show that $T(\mathcal{B}) \geq 1$. Let $\mathcal{B}_m = p \in \mathcal{B} : p(T_m) > 0$.

We have for all $q \in \mathbb{P}(\mathbb{N})$ and $m \geq k$ that

$$T_2(B, q) = \sup_{p \in B_{x \geq 2^k}} \sum_{x \geq 2^k} p(x) \log \frac{p(x)}{q(x)} \geq \sup_{p \in B_m} \sum_{x \geq 2^k} p(x) \log \frac{p(x)}{q(x)}$$

$$= \sup_{p \in B_m} \max_{x \geq 2^k} \frac{p(x)}{q(x)} \geq 1.$$

Thus, for all $q$, $T_2(B, q) \geq 1$. Hence $T(\mathcal{B}) \geq 1$. Choose $q \in \mathbb{P}(\mathbb{N})$ that assigns $q(T_n) = \frac{1}{n^{1+\epsilon}}$ for all $n$, and $q(x(T_n)) = 1/2^n$ for all $x \in T_n$. For this distribution, we have

$$T_2(B, q) = 1,$$

and thus $T(\mathcal{B}) = 1$. Thus, we can also conclude that

$$R(\mathcal{B}_\infty) = 1,$$

improving the estimate $R(\mathcal{B}_\infty) \geq 1 - \frac{1}{7}$ in Example 2.

Example 4: Let $h > 0$. Let $\mathcal{M}_h$ be the collection of monotone distributions over $\mathbb{N}$ such that

$$E_p \left( \log \frac{1}{p(X)} \right)^2 < h.$$

Let $\mathcal{M}_\infty$ be the set of all i.i.d. distributions with one dimensional marginals from $\mathcal{M}_h$. We show that $T(\mathcal{M}_h) = 0$, and thus that $\mathcal{M}_\infty$ is strongly compressible. To see this, observe that since $p \in \mathcal{M}_h$ is monotone, we will have $p(i) \leq \frac{1}{7}$ and therefore $E_p(\log X)^2 \leq h$, and from Cauchy-Schwartz, that $E_p \log X \leq \sqrt{h}$. Hence for all $m \geq 2$,

$$p(x \geq m) \leq \frac{h}{(\log m)^2}.$$
Taking a supremum over \( p \) and limit as \( m \to \infty \), we conclude \( T(M_n) = 0 \).

VI. PROOF OF THEOREM 1

We first consider the case when \( P \) is not tight. Using Lemma 1, \( R_1 = \infty \) and using Lemma 7, the tail redundancy \( T(P) = \infty \) as well. Furthermore, \( R_1 = \infty \) implies from Proposition 1 that for all \( m \geq 1 \), \( \frac{1}{m} R_m(P^{\infty}) = \infty \). Therefore, if \( P \) is not tight, \( \frac{1}{m} R_m(P^{\infty}) = T(P) = \infty \) and the Theorem holds.

We now consider the case when \( P \) is tight. If \( P \) is tight we first show in Section VI-A that for all \( m \), \( \frac{1}{m} R_m(P^{\infty}) \geq T(P) \) and in Section VI-B that \( \frac{1}{m} R_m(P^{\infty}) \leq T(P) \).

A. Converse Part

We show that if \( P \) is tight, for all \( n \geq 1 \),

\[
\frac{1}{n} R_n(P^{\infty}) \geq T(P),
\]

thus also proving that \( R(P^{\infty}) \geq T(P) \). Since \( P \) is tight, for any \( c > 0 \), we can find a finite number \( m_n \) such that \( \forall p \in P \),

\[
p(x \geq m_n) < \frac{c}{n}.
\]

Let \( t_n^p \equiv p(x \geq m_n) \) be the tail probability past \( m_n \) under \( P \). Let \( Y = \{-1, 1, 2, \ldots, m_n - 1\} \). For each sequence \( x^n \in \mathbb{N}^n \), let the auxiliary sequence \( y^n = y_1, \ldots, y_n \in Y^n \) be defined by

\[
y_i(y^n) = \begin{cases} x_i & \text{if } x_i < m_n \\ -1 & \text{if } x_i \geq m_n. \end{cases}
\]

For \( x^n \in \mathbb{N}^n \) and \( y^n \in Y^n \), we say \( x^n \sim y^n \) if \( x^n \) and \( y^n \) are consistent (\( y^n \) would be the auxiliary sequence constructed from \( x^n \)). For all \( n \geq 1 \) and \( r_n \in \mathbb{P}([0,\infty)) \), we show that

\[
\frac{1}{n} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{r_n(x^n)} \geq T(P).
\]

proving also that

\[
R(P^{\infty}) = \lim_{n \to \infty} \frac{1}{n} R_n(P^{\infty})
\]

\[
= \sup_{n \to \infty} \inf_{r_n \in \mathbb{P}([0,\infty))} \frac{1}{n} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{r_n(x^n)}
\]

\[
\geq T(P).
\]

Fix any \( r_n \in \mathbb{P}([0,\infty)) \). Now for all \( x^n \in \mathbb{N}^n \) and \( y^n \) such that \( x^n \sim y^n \),

\[
r_n(x^n) = r_X(x^n, y^n) = r_{XY}(x^n | y^n) r_Y(y^n),
\]

where we use the subscripts \( XY \), \( X | Y \) and \( Y \) to denote the appropriate induced distributions, i.e., \( r_Y(y^n) = \sum_{x^n} p(x^n) r_n(x^n) \), \( r_{XY}(x^n, y^n) = r_n(x^n) \) if \( x^n \sim y^n \) and 0 else, and \( r_{X | Y}(x^n | y^n) = r_{n}(x^n) / r_Y(y^n) \) if \( x^n \sim y^n \) and 0 else. Define \( G \subset Y^n \), where

\[
G = \{ y^n \in Y^n : \text{exactly one element of } y^n \text{ is } -1 \}.
\]

We will focus our attention primarily on auxiliary sequences in \( G \), by noting

\[
\sup_{p \in P} D_n(p | r_n) = \sup_{p \in P} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{r_n(x^n)}
\]

\[
= \sup_{p \in P} \left( \sum_{y^n \in G} \sum_{x^n} p(x^n) \log \frac{p(x^n | y^n)}{r_{XY}(x^n | y^n) r_Y(y^n)} + \sum_{y^n \not\in G} \sum_{x^n} p(x^n) \log \frac{p(y^n)}{r_Y(y^n)} \right)
\]

\[
\geq \sup_{p \in P} \sum_{y^n \in G} \sum_{x^n} p(x^n) \log \frac{p(x^n | y^n)}{r_{XY}(x^n | y^n) r_Y(y^n)}.
\]

where (a) uses the fact that KL divergence is greater than or equal to 0, and in the above, we use the convention that if \( x^n \not\sim y^n \), both \( p(x^n | y^n) \) and \( r_{n}(x^n | y^n) \) are 0 and the corresponding contribution of such a pair to the summations above is 0.

For a given consistent pair \( x^n \sim y^n \) for \( y^n \in G \), it is easy to see that if \( y_i = -1 \), then \( x_i \) is the only symbol \( \geq m_n \) and \( p(x^n | y^n) \) is essentially written in terms of the single letter distribution on \( N \) which we denote \( p(\cdot | x \geq m_n) \), namely

\[
p(x_i | X \geq m_n) \equiv p(x_i | x^n) = \frac{p(x_i)}{\sum_{x \geq m_n} p(x)} = \frac{p(x_i)}{r^n_{\mu}}.
\]

Similarly, for any \( y^n \in G \), we can extract a single letter distribution \( r_{X | Y}(x^n | y^n) \) in a similar fashion. If \( i \) is the only number such that \( y_i = -1 \), then for all \( x^n \sim y^n \)

\[
r_{X | Y}(x_i | y_i) \equiv r_{X | Y}(x^n | y^n) = \frac{r_n(x_i y_{i+1}^{-1})}{\sum_{y_{i+1} \in Y} r_n(x_i y_{i+1}^{-1})}
\]

Therefore,

\[
\sup_{p \in P} \sum_{y^n \in G} \sum_{x^n \sim y^n} p(x^n) \log \frac{p(x^n | y^n)}{r_{XY}(x^n | y^n) r_Y(y^n)}
\]

\[
\geq \sup_{p \in P} \sum_{y^n \in G} \sum_{x \geq m_n} p(x | X \geq m_n) \log \frac{p(x | X \geq m_n)}{r_{X | Y}(x^n | y^n)}
\]

\[
= \sup_{p \in P} \sum_{y^n \in G} \sum_{x \geq m_n} p(x) \log \frac{p(x) / r^n_{\mu}}{r_{X | Y}(x^n | y^n)}.
\]

To reduce the above expression, for all \( p \in P \), let

\[
y(p) = \arg \min_{y^n \in G} \sum_{x \geq m_n} p(x) \log \frac{p(x)}{r_{X | Y}(x^n | y^n)}.
\]

Then,

\[
\sup_{p} \sum_{y^n \in G} \sum_{x \geq m_n} p(x) \log \frac{p(x) / r^n_{\mu}}{r_{X | Y}(x^n | y^n)}
\]

\[
\geq \sup_{p} \sum_{y^n \in G} \sum_{x \geq m_n} p(x) \log \frac{p(x) / r^n_{\mu}}{r_{X | Y}(x^n | y^n)}.
\]
Observing that for all \( p \in \mathcal{P} \), \( p(G) = n(1 - \tau^p_n)^{n-1} \tau^p_n \), we have therefore that
\[
\sup_{p} D_n(p(x) || r(x)) \geq \sup_{p} \frac{p(G)}{\tau^p_n} \sum_{x \leq m_n} p(x) \log \frac{p(x)}{r^p(x)}(x) + \tau^p_n \log \frac{1}{\tau^p_n}.
\]

(11)

For \( y \in G \), let \( \mathcal{P}_y = \{ p \in \mathcal{P} : y(p) = y \} \). Then \( \mathcal{P} \) can be written as the finite union,
\[
\mathcal{P} = \cup_{y \in G} \mathcal{P}_y.
\]

Therefore from Lemma 7,
\[
T(\mathcal{P}) = \max_{y \in G} T(\mathcal{P}_y).
\]

We then have
\[
\sup_{p \in \mathcal{P}} \left[ \sum_{x \leq m_n} p(x) \log \frac{p(x)}{r^p(x)}(x) + \tau^p_n \log \frac{1}{\tau^p_n} \right]
\]
\[
= \max_{y \in G} \sup_{p \in \mathcal{P}_y} \left[ \sum_{x \leq m_n} p(x) \log \frac{p(x)}{r^p(x)}(x) + \tau^p_n \log \frac{1}{\tau^p_n} \right]
\]
\[
= \max_{y \in G} \left( \inf_{q \in P(N)} \sup_{p \in \mathcal{P}_y} \left[ \sum_{x \leq m_n} p(x) \log \frac{p(x)}{q(x)} + \tau^p_n \log \frac{1}{\tau^p_n} \right] \right)
\]
\[
= \max_{y \in G} T_m(\mathcal{P}_y)
\]
\[
= T(\mathcal{P}),
\]
(12)

where (\( \ast \)) follows since for any collection, Lemmas 3 and 4 together imply that \( T_m \) monotonically decreases to the limit \( T \). Putting (11) and (12) together, we obtain
\[
\sup_{p \in \mathcal{P}} D_n(p(x) || r(x)) \geq n \left( 1 - \frac{c}{n} \right)^n T(\mathcal{P}).
\]

Since the inequality holds for all \( c > 0 \), we have
\[
\sup_{p \in \mathcal{P}} \frac{1}{n} D_n(p(x) || r(x)) \geq \left( 1 - \frac{c}{n} \right)^n T(\mathcal{P}) = T(\mathcal{P}).
\]

B. Direct Part

We now show that
\[
R(\mathcal{P}^\infty) = \lim_{n \to \infty} \frac{1}{n} R_n(\mathcal{P}^\infty) \leq T(\mathcal{P}).
\]

First, note from Proposition 1 that
\[
R(\mathcal{P}^\infty) \leq \lim_{n \to \infty} \frac{1}{n} R_n(\mathcal{P}^\infty) \leq R_1(\mathcal{P})
\]
so if \( R(\mathcal{P}^\infty) \) is infinite, then \( R_1(\mathcal{P}) \), and from Proposition 7, \( T(\mathcal{P}) \) is infinite as well, and vacuously, \( R(\mathcal{P}^\infty) \leq T(\mathcal{P}) \).

For the rest of the proof, we assume that \( R(\mathcal{P}^\infty) < \infty \).

Our proof will be constructive. We describe length \( n \) sequences from \( \mathbb{N}^n \) using distributions \( q_n \in \mathbb{P}(\mathbb{N}^n) \), constructed as follows. We first clip the sequences at a threshold \( m \), replacing all occurrences of numbers \( \geq m \) in the sequence with a new symbol, \( -1 \). To complete the description, we then describe the actual number that occurred corresponding to each \( -1 \) using a single letter distribution \( q^s \in \mathbb{P}(\mathbb{N}) \) that achieves the tail redundancy of \( \mathcal{P} \). The threshold \( m \) will be chosen to vary with the sequence length \( n \). This simple construction is enough to achieve asymptotically per-symbol redundancy of \( \leq T(\mathcal{P}) \).

While this approach will yield a sequence of distributions \( \{ q_n \in \mathbb{P}(\mathbb{N}^n), n \geq 1 \} \) that are not consistent (primarily because we vary the threshold \( m \) with the sequence length \( n \)), note that the general construction in Appendix A provides a way to construct a universal probability measure \( q \in \mathbb{P}(\mathbb{N}^\infty) \) which incurs the same asymptotic per-symbol redundancy as the sequence \( \{ q_n \in \mathbb{P}(\mathbb{N}^n), n \geq 1 \} \) of distributions.

We begin by noting that for any finite \( m \), there is a distribution \( r_m \) that achieves the minimax redundancy of encoding \( m \)-ary i.i.d. strings \([13]\). \(^3\) Let the redundancy of \( r_m \) against \( (m+1) \)-ary i.i.d. sequences of length \( n \) be \( \rho_{m,n} \). It is known that \( \rho_{m,n} \sim \frac{2}{m} \log n \) \([2], [14], [20]\). In particular, it is easy to see that encoding these sequences with the (suboptimal) add-1 (Laplace estimator, or the Bayesian mixture with the conjugate Dirichlet prior with all parameters 1) rule incurs redundancy \( \log (\frac{n+m-1}{m-1}) \), so
\[
\rho_{m,n} \leq \log (\frac{n+m-1}{m-1}) \leq (m-1) \log n.
\]

(13)

Let \( q^s \) be the distribution that achieves \( T(\mathcal{P}) \). With the benefit of hindsight, we set \( m = \sqrt{n} \). As before, let \( Y = \{ -1, 1, \ldots, m-1 \} \) and we construct an auxiliary sequence \( y^m \in \mathcal{Y}^m \) from \( x^m \) where
\[
y_i = \begin{cases} x_i & \text{if } x_i < m \\ -1 & \text{if } x_i \geq m \end{cases}
\]

As before, given any sequence \( y^m \in \mathcal{Y}^m \), and \( x^m \in \mathbb{N}^m \), we say \( x^m \sim y^m \) if \( y^m \) is consistent with \( x^m \) \((y^m \) would be constructed from \( x^m \)). Let \( q_m(x) = q^s(x)/\sum_{x \geq m} q^s(x') \) for \( x \geq m \). Then, we construct a distribution \( q_m \in \mathbb{P}(\mathbb{N}^m) \) by first specifying the probabilities of the auxiliary sequences \( y^m \in \mathcal{Y}^m \) using the \( m \)-ary minimax optimal distribution \( r_m \),
\[
q(y^m) = r_m(y^m),
\]
followed by describing \( x_i \) for each \( y_i \),
\[
q(x_i | y_i) = \begin{cases} q_m(x_i) & \text{if } y_i = -1 \\ 1 & \text{if } y_i \neq -1, x_i = y_i \\ 0 & \text{if } y_i \neq -1, x_i \neq y_i. \end{cases}
\]

\(^3\) Such a \( r_m \) is also a Bayesian mixture of the \( m \)-ary i.i.d. probability measures.
Finally for all $x^n$, $q(x^n) = \sum_{c \in \mathcal{Y}} q_Y(c^n) q(x^n|c^n)$, which will coincide with $q_Y(y^n) q(x^n|y^n)$ for the unique $y^n \in \mathcal{Y}^n$ that is consistent, i.e., $x^n \sim y^n$. Then,

$$
\frac{1}{n} \sum_{x^n \in \mathcal{X}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \frac{1}{n} \sum_{y^n \in \mathcal{Y}^n} p(y^n) \log \frac{p(y^n)}{q(y^n)} + \frac{1}{n} \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} p(x^n) \log \frac{p(x^n|y^n)}{q(x^n|y^n)}
$$

$$
= \frac{\rho_{m,n}}{n} + \frac{1}{n} \sum_{y^n \in \mathcal{Y}^n} p(y^n) \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} p(x^n|y^n) \log \frac{p(x^n|y^n)}{q(x^n|y^n)}. \tag{14}
$$

For any $y^n \in \mathcal{Y}^n$, let $k(y^n)$ be the number of occurrences of $-1$ in $y^n$. Let $\tau_{p,m} = \sum_{x \geq m} p(x)$, then for all $x^n \sim y^n$,

$$
p(x^n|y^n) = \prod_{i \subset j = -1} p(x_i). \tag{15}
$$

We can rewrite the second term in equation (14) as

$$
\frac{1}{n} \sum_{y^n \in \mathcal{Y}^n} p(y^n) \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} p(x^n|y^n) \log \frac{p(x^n|y^n)}{q(x^n|y^n)}
$$

$$
= \frac{1}{n} \sum_{y^n \in \mathcal{Y}^n} p(y^n) \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} \left( \prod_{i \subset j = -1} \frac{p(x_i)}{\tau_{p,m}} \right) \log \prod_{i \subset j = -1} \frac{p(x_i)/\tau_{p,m}}{q_m(x_i)}
$$

$$
def \frac{1}{n} \sum_{y^n} p(y^n) A(k(y^n)).
$$

For each $y^n$, we can bound $A(k(y^n))$ as follows,

$$
A(k(y^n)) = \sum_{j \subset i \subset -1} \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} = k(y^n) \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)}. \tag{16}
$$

Now we have $\sum_{y^n} p(y^n) A(k(y^n))$ equals

$$
\sum_{y^n} p(y^n) k(y^n) \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} = \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} \sum_{y^n} p(y^n) k(y^n).
$$

Combining equation (14) and (16), we have

$$
\frac{1}{n} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \leq \frac{\rho_{m,n}}{n} + \frac{1}{n} \mathbb{E} k(Y^n) \left( \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} \right)
$$

$$
\leq \frac{\rho_{m,n}}{n} + \frac{1}{n} \mathbb{E} k(Y^n) \left( \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} \right)
$$

$$
\leq \frac{\rho_{m,n}}{n} + \frac{1}{n} \mathbb{E} k(Y^n) \left( \sum_{x \geq m} \frac{p(x)}{\tau_{p,m}} \log \frac{p(x)/\tau_{p,m}}{q_m(x)} \right)
$$

where the second to last inequality follows since $\mathbb{E}(Y^n) = n\tau_{p,m}$ and the last inequality because $q_m(x) = q^{*}(x)$.

\[\sum_{i \geq m} q^{*}(x') \geq q^{*}(x).\] Now, taking the supremum over all $p$ and the limsup as $n \to \infty$, we have

\[\limsup \sup_{n \to \infty} \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \leq \limsup \sup_{n \to \infty} p \left( \frac{\rho_{m,n}}{n} + \sum_{x \geq m} p(x) \log \frac{p(x)}{q^{*}(x)} + \tau_{p,m} \frac{1}{\tau_{p,m}} \log \frac{1}{\tau_{p,m}} \right).\]

We claim that the limit above is $T(\mathcal{P})$. Now from (13), for all $m$ and $n$,

\[\frac{\rho_{m,n}}{n} \leq \frac{m-1}{n} \log n,\]

so as $n \to \infty$ while $m = \sqrt{n}$, $\frac{\rho_{m,n}}{n} \to 0$. The second term in the parenthesis goes to $T(\mathcal{P})$ as $n \to \infty$, since $m = \sqrt{n} \to \infty$ and since $q^{*}$ achieves the tail redundancy for the collection $\mathcal{P}$. For the last term, recall that when the asymptotic per-symbol redundancy is finite, so is the single letter redundancy. Therefore, the collection $\mathcal{P}$ is tight, and hence as $n \to \infty$, while $m = \sqrt{n}$,

\[\lim_{n \to \infty} \sup_{p \in \mathcal{P}, \sqrt{n}} \log \frac{1}{\tau_{p,\sqrt{n}}} \to 0.\]

VII. CONCLUSION

The paper establishes the scaling of the asymptotic per-symbol i.i.d. redundancy of compressing sources from a countable alphabet, capturing it by a single letter characterization called the tail redundancy.

APPENDIX A

Lemma 8: Let $a^{(j)}$, $1 \leq j \leq k$ be $k$ different sequences with limits $a^{(j)}$ respectively. For all $i$, let

\[\hat{a}_i = \max a^{(i)}.\]

Then the sequence $\{\hat{a}_i\}$ has a limit and the limit equals $\max a^{(i)}$.

Proof: Wolog, let the sequences be such that the limits are $a^{(1)} \geq a^{(2)} \geq \ldots \geq a^{(k)}$. Consider any $0 < \epsilon < \frac{a^{(1)} - a^{(k)}}{2}$. Then for all $1 \leq j \leq k$, there exist $N_j$ such that for all $n \geq N_j$, $|a^{(j)} - a^{(i)}| \leq \epsilon$. Let $N = \max N_j$. We now have that for all $i \geq N$,

\[\hat{a}_i = \max a^{(j)} = a^{(1)};\]

and therefore, the sequence $\{\hat{a}_i\}$ has a limit, and is equal to $a^{(1)} = \max_{1 \leq j \leq k} \lim_{i \to \infty} a^{(j)}$.

APPENDIX B

Our definitions of redundancy in (4) and (5) conform to the standard definitions, though they look non-standard. This Appendix proves the equivalence for completeness.

In standard parlance, we usually adopt

\[\rho_n \overset{\text{def}}{=} \inf_{q_n \in \mathcal{P}(\mathcal{X}^n)} \sup_{p \in \mathcal{P}} \frac{1}{n} \mathbb{E} \log \frac{p(X^n)}{q_n(X^n)}\]
where $q_n$ is any distribution over $\mathbb{N}^n$ as the length-$n$ redundancy, while the asymptotic per-symbol redundancy is $\rho \triangleq \limsup_{n \to \infty} \rho_n$. In this Appendix, we show that

$$\rho_n = \inf_{q \in \mathbb{P}(\mathbb{N}^n)} \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} = \frac{1}{n} \log (\rho_n)$$

and

$$\rho = \inf_{q \in \mathbb{P}(\mathbb{N}^n)} \limsup_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} = \frac{1}{n} \log (\rho),$$

where $R_n$ and $R$ are our definitions from (4) and (5), which require consistency of the estimator across sequence lengths unlike $\rho_n$ and $\rho$.

Claim 1: For all $n \geq 1$, $\rho_n = R_n$.

Proof: Follows directly from [27, Lemma 32].

Claim 2: $\rho = R$.

Proof: We get for free that

$$\rho \leq \inf_{q} \limsup_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} = R$$

We will now show that $R < \rho$ to establish the claim. From the standard definition of $\rho$, we know that for each $\epsilon > 0$, there is a sequence $\{q_n \in \mathbb{P}(\mathbb{N}^n) : n \geq 1\}$ of distributions over $\mathbb{N}^n$ respectively satisfying $\limsup_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q_n(X^n)} < \rho + \epsilon$.

Now each $q_n \in \mathbb{P}(\mathbb{N}^n)$ can be extended to a probability measure $q_n^* \in \mathbb{P}(\mathbb{N}^\infty)$ as in [27, Lemma 32]. Define the measure $q_n \in \mathbb{P}(\mathbb{N}^n)$ by assigning to each finite sequence $x$ of natural numbers, $q_n(x) = \sum_{m \geq 1} q_n^*(x) \frac{1}{m(m+1)}$, and extending it to a probability measure $\mathbb{P}(\mathbb{N}^\infty)$. Now we have

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q_n(X^n)} = \lim_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q_n^*(X^n)} \leq \lim_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q_n^*(X^n)} + \log(n(n+1)) < \rho + \epsilon,$$

thus proving that for all $\epsilon > 0$,

$$R = \inf_{q \in \mathbb{P}(\mathbb{N}^n)} \limsup_{n \to \infty} \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} < \rho + \epsilon.$$