Data-derived weak universal consistency: the case of universal compression

Narayana Santhanam Venkat Anantharam Wojciech Szpankowski

Abstract

Many current applications in data science need rich model classes to adequately represent the statistics that may be driving the observations. But rich model classes may be too complex to admit estimators that converge to the truth with convergence rates that can be uniformly bounded over the entire collection of probability distributions comprising the model class, i.e. it may be impossible to guarantee uniform consistency of such estimators as the sample size increases. In such cases, it is conventional to settle for estimators with guarantees on convergence rate where the performance can be bounded in a model-dependent way, i.e. pointwise consistent estimators. But this viewpoint has the serious drawback that estimator performance is a function of the unknown model within the model class that is being estimated, and is therefore unknown. Even if an estimator is consistent, how well it is doing at any given time may not be clear, no matter what the sample size of the observations.

Departing from the classical uniform/pointwise consistency dichotomy that leads to this impasse, a new analysis framework is explored by studying rich model classes that may only admit pointwise consistency guarantees, yet all the information about the unknown model driving the observations that is needed to gauge estimator accuracy can be inferred from the sample at hand. We expect that this *data-derived* estimation framework will be broadly applicable to a wide range of estimation problems by providing a methodology to deal with much richer model classes. In this paper we analyze the lossless compression problem in detail in this novel data-derived framework.

I. Introduction and Motivation

Many of the most challenging problems in the data sciences stem from one or more of the following characteristics associated with data: extreme scale (typically requiring that the data reside on multiple storage nodes); high dimensionality and sparsity; patterns in the data that manifest at multiple scales; dynamic, temporal, and heterogeneous structure; complex dependencies between different parts of the data; and noise/ missing data. Tasks such as image recognition, classification, control and many others, which are built on such data sources, depend on estimating the relevant underlying structure in the data. Rich model classes, i.e. rich collections of probabilistic models, such as the collection of all probability distributions over a large or countably infinite support, or the set of long memory, slowly mixing Markov processes are often required to adequately model the complex characteristics of these data sources. A comprehensive approach to address these key challenges is critically needed.

N. Santhanam is with the Department of Electrical Engineering, University of Hawaii, Manoa. Email: nsanthan@hawaii.edu.

V. Anantharam is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley. Email: ananth@eecs.berkeley.edu.

W. Szpankowski is with the Department of Computer Science, Purdue University. Email: szpan@purdue.edu.

Indeed, in bringing rigorous theory to bear on data science, the first question we face is related to model selection. There is often a tension between the need for rich model classes to better represent data and our ability to handle these collections from a mathematical point of view. Many applications, particularly in the big data regime, force us to consider model collections that are too complex to admit estimators with traditional model-agnostic *uniformly* consistent guarantees. These new collections often only admit *pointwise* convergent estimators [1] – i.e. convergence is only guaranteed individually for each model in the model class – which often are difficult to use predictively as their convergence cannot be verified. In this paper we depart from this dichotomy, and we propose a new analysis framework by characterizing rich model classes that may only admit pointwise guarantees, yet all the information about the unknown model needed to gauge estimator accuracy can be inferred from the observations. More precisely, we introduce here *data-derived* consistency, a new framework to analyze these rich model collections. To retain focus, in this document we concentrate on universal compression to bring out the salient features of this framework. We also make connections to a related prediction problem that was analyzed by us earlier in [2], and is now seen to fit into this broader framework.

The richness of a model class is often quantified by metrics such as its VC-dimension [3], Rademacher complexity [4], [5], [6], or – what is most relevant in the context of universal compression – its asymptotic per-symbol redundancy [7], [8], [9], [10], [11], [12]. Ideally, one would want an estimation algorithm with a model-agnostic guarantee on its perfomance, depending only on the sample size – this is the *uniform* consistency dogma that underlies most formulations of engineering applications today. But requiring such uniform consistency restricts the richness of the model classes we can deal with. Generally speaking, the more complex a model is, the less one could expect to be able to provide such uniform consistency guarantees.

When the model classes we are interested in are too complex to admit *uniformly consistent estimators*, the common belief is that the best we can do is to have estimators with convergence guarantees dependent on not just the sample size but on the underlying model in the model class that governs the statistics of the observations. These are called *pointwise consistent* estimators. It is well-understood that this viewpoint may not always be particularly useful, the problem now being that our gauge of the performance of the estimation scheme is dependent on the unknown underlying model – the very ambiguity we are addressing! Even if we have a pointwise consistent estimator, which is eventually almost surely accurate under the underlying model, for any fixed sample size we may never know how well the estimator is doing no matter how large the sample size is.

We illustrate this issue with a simple example below. Before doing so, we first introduce some of the notational conventions that will be used throughout this document. The symbol $:=$, and occasionally $=:$, is used to denote equality by definition. We write log for logarithms to base 2 and ln for logarithms to the natural base. The set of natural numbers, denoted $\mathbb{N}$, is the set $\{1, 2, \ldots\}$, thought of as endowed with its usual σ-algebra comprised of all subsets of $\mathbb{N}$. For $n \geq 1$, we use $\mathbb{N}^n$ to denote the set of strings of length n of natural numbers, with the product σ-algebra. The set of infinite sequences of natural numbers is denoted $\mathbb{N}^\infty$, and is thought of as endowed with the corresponding product σ-algebra. We will adopt the convention of thinking of a probability measure on $\mathbb{N}$ as defined by a distribution, which assigns a probability to each natural number. A string of integers $(x_1, \ldots, x_n) \in \mathbb{N}^n$ will be denoted by $\mathbf{x}$, or by x^n when it seems important to emphasize the specific length of the string. For a string of integers $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{N}^n$, its empirical distribution or *type* is the sequence of unnormalized fractions on $\mathbb{N}$ assigning the fraction $\frac{m}{n}$ to $x \in \mathbb{N}$ if x shows up m times in the string $\mathbf{x}$. It is conventional to think of the type as a probability distribution on $\mathbb{N}$ and we will do so when convenient, but it is important at some places in the document to think of it as comprised of unnormalized fractions. $\mathbb{N}^*$ denotes the set of strings of naturals of finite length, including the empty string. For the purposes of this paper it suffices to think of $\mathbb{N}^*$ as a set with no additional structure.

Example 1. (Hiding entropy)

For $\epsilon > 0$ and $M \in \mathbb{N}$, let $p_{\epsilon,M}$ be the probability distribution that assigns probability $1-\epsilon$ to the natural number 1 and assigns probability ϵ/M to the natural numbers 2 through $M+1$. Denote the probability distribution that assigns probability 1 to the natural number 1 by p_0. Let $\mathcal{W}$ be the set comprised of the probability distributions $p_{\epsilon,M}$ for $\epsilon > 0$ and $M \in \mathbb{N}$, as well as p_0.

Our task is to estimate the Shannon entropy of a probability distribution in $\mathcal{W}$ using *i.i.d.* samples from it. However, we do not know which probability distribution in $\mathcal{W}$ is governing the law of the observed samples. The natural *plug-in estimator* assigns to a sample $X_1, \ldots, X_n$ the entropy of its empirical distribution. Since every probability distribution in $\mathcal{W}$ has finite support, the plug-in estimate is consistent almost surely, no matter which underlying distribution from $\mathcal{W}$ is generating the observations. But at what point do we know that the plug-in estimate is close to the correct answer? Indeed, can we, at any point, get an upper bound for the true entropy using the plug-in estimate with, say, a confidence probability $3/4$, regardless of what the true probability distribution in $\mathcal{W}$ is?

It turns out that it is *impossible* to provide such guarantees for $\mathcal{W}$. To see why, suppose we see a sequence of n successive 1s. This could have come from p_0, or, with high probability, from any probability distribution $p_{\epsilon,M}$ with $0 < \epsilon \ll \frac{1}{n}$. What is worse, for any upper bound $\hat{h}$ we may provide, however large, even if $0 < \epsilon \ll \frac{1}{n}$, the entropy of $p_{\epsilon,M}$ where $M \geq 2^{\hat{h}/\epsilon}$ is $h(\epsilon) + \epsilon \log M \geq \hat{h}$. Every such $p_{\epsilon,M}$ gives the sample of n successive 1s a probability of at least $> 3/4$ if ϵ is sufficiently small, so our upper bound fails.

This argument applies whether we obtained $\hat{h}$ from the plug-in estimator or *any* other estimator of the entropy. No upper bound that we propose on the entropy based on any finite sequence of 1s can hold with confidence probability $3/4$ under all probability distributions in $\mathcal{W}$. To make matters worse, the sequence of all 1s occurs with probability 1 when the underlying model in force is p_0. Therefore, even when we could estimate the entropy consistently, we could never obtain even a trivial upper bound on the entropy with a confidence probability $\geq 3/4$. □

We therefore challenge the dichotomy of *uniform* and *pointwise* consistency in the analysis of statistical estimators. This paper considers a new paradigm positioned in between these two extremes. We modify the definition of pointwise consistent estimators, keeping as far as possible the richness of the model class but ensuring that all the information needed about the unknown model to evaluate estimator accuracy can be gleaned from the observations. We call this modified notion of pointwise consistency *data-derived* consistency. The crux of the *data-derived* framework is to provide a mechanism that allows us to gauge from the observations how well we are doing.

We bring out the salient features of the data-derived framework in this document in the framework of universal compression. In the context of providing efficient compressed representations of samples from a data source, the goal of universal compression is to be able to work with a rich class of models for the source being compressed. Universal compression posits that we have a model class of source probability measures, while we are required to come up with a universal probability measure that attempts to compress any source in the model class as well as possible without prior knowledge of the source. Since the universal probability measure is not exactly matched to any single source probability measure in the model class it incurs a redundancy, measured using the Kullback-Leibler (KL) divergence, against any source in the model class when compressing a sequence of observed samples whose statistics are governed by this source. The uniform consistency setup in this case corresponds to what is commonly known as the *strong* compression formulation, where we find universal probability measures whose per-symbol redundancy incurred against any source in the model class can be uniformly bounded over the entire model class and, in addition, diminishes to 0 as the sample size grows to infinity. The pointwise consistency setup in this case corresponds to what is commonly known as the *weak* compression formulation and is one where the universal probability measure incurs asymptotically zero per-symbol redundancy against each source in the model class, but the convergence to zero is not necessarily uniform over the entire model class.

We propose and study the *data-derived* weak compression formulation (*d.w.c.*) which identifies when, in the weak compression setup, we can also estimate the redundancy of the universal probability measure relative to the underlying source model generating the data. Broadly speaking, we aim to find a universal estimator/encoding with a given accuracy as well as a corresponding stopping rule that allows us to find out at what point the KL divergence from the true source becomes (and remains) small, for a predetermined sequence length. To characterize the classes of probability distributions on $\mathbb{N}$ that are data-derived weakly compressible, we shall introduce the notion of what it means for a probability distribution in the class to be *deceptive* relative to the class. At a high level, a source probability distribution, viewed as a member of a collection of probability distributions, is *deceptive* if the asymptotic per-symbol redundancy of neighborhoods of the source within the model class is bounded away from 0, in the limit as the neighborhood shrinks to 0. Then, in our main finding, Theorem 17, we show that *a collection of probability measures is data-derived weakly compressible iff no source in the model class is deceptive.*

As we delve deeper into this formulation, we will see that data-derived consistency changes how we think of model classes. It shifts the focus away from the global complexity of the model class to some form of local complexity of each model within the model class, viewed as a member of the model class.

Our notion of data-derived consistency is closely related to other formulations in compression and learning theory – in particular hierarchical universal compression [13] and data-dependent structural risk minimization [14], as well as its subsequent development via the *luckiness* framework [15]. Fundamental to all these approaches is to balance the sample complexity of learning with the desire for richer model collections (or hypothesis collections as the case may be).

The paper is organized as follows. In the next section we develop our data-derived approach. Section III recalls some of the central prior results on universal compression that we build on in our work. Section IV discusses our main result (Theorem 17), which completely characterizes *d.w.c.* model classes of *i.i.d.* probability distributions on a countable set. We then illustrate several nuances in our formulation and results using several examples in Section V. Sections VI and VII are devoted to proving the main result. The main thread of the discussion is supported by several appendices. Appendix I reconciles the traditional definitions of strong and weak compressibility with those we work with in this paper. Appendix II gathers several basic results on entropy, redundancy and the Jensen-Shannon divergence that we draw upon throughout the paper. Appendix III contains the details of the proof for the claims made regarding one of the examples in Section V. Appendix IV proves a lemma needed for the proof the sufficiency part of the main theorem. The last bit of the proof of the necessity part of the main theorem is in Appendix V and that of the sufficiency part in Appendix VI. Finally, Appendix VII corrects an erroneous claim made in passing in the concluding remarks in [2] (which does not in any way affect the rest of that paper).

II. Formulation of the problem

Let $\mathcal{P}$ be a collection of probability distribution over $\mathbb{N}$. Given $\mathcal{P}$, we let $\mathcal{P}^\infty$ denote the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. We will use the term *source* to denote either $p \in \mathcal{P}$ or $p^\infty \in \mathcal{P}^\infty$ as appropriate. For notational simplicity and following the convention in literature, we will also often drop the superscript in p^∞ and use p both for the probability distribution on $\mathbb{N}$ and the corresponding *i.i.d.* probability measure induced on $\mathbb{N}^\infty$. Further, for $n \geq 1$ and a string of natural numbers $\mathbf{x} := (x_1, \ldots, x_n) =: x^n \in \mathbb{N}^n$, we will write $p(\mathbf{x})$ or $p(x^n)$ for $\prod_{i=1}^n p(x_i)$. Here p can be thought of as a simplified notation for the product probability measure p^n on $\mathbb{N}^n$ corresponding to the probability distribution p on $\mathbb{N}$.

We consider here the lossless compression problem for collections of large alphabet *i.i.d.* sources. The main contribution of this work is to propose and develop the data-derived framework for estimation problems. The large alphabet *i.i.d.* compression problem is the vehicle we have used to do this, but the reader can no doubt easily come up with her or his favorite estimation problem where this framework might lead to interesting developments. In Example 8 we consider the problem of estimation of percentiles of the probability distribution defining the source – this has been studied in depth in [2], so here all we show is that this estimation task lies in the data-derived framework proposed in this document. Another example, which we have not studied in depth, but which seems to us to be particularly interesting, is that of entropy estimation, see Example 9.

Before embarking on the discussion, we introduce some additional notational conventions. For $1 \leq m \leq n$ and strings $\mathbf{y} \in \mathbb{N}^m$ and $\mathbf{x} \in \mathbb{N}^n$, we write $\mathbf{y} \preceq \mathbf{x}$ to denote that $\mathbf{y}$ is a prefix of $\mathbf{x}$. We can also use this notation when $\mathbf{y} \in \mathbb{N}^m$ and $x \in \mathbb{N}^\infty$. The length of a finite string $\mathbf{x} \in \mathbb{N}^n$ is denoted by $|\mathbf{x}|$.

For a probability measure q on $\mathbb{N}^\infty$, given $n \geq 1$ and a string $\mathbf{x} \in \mathbb{N}^n$, we write $q(\mathbf{x})$ for the probability under q of the set of strings in $\mathbb{N}^\infty$ whose prefix of length n is $\mathbf{x}$. In effect, we are treating $\mathbf{x}$ as also denoting an event in $\mathbb{N}^\infty$. Note that, for $p \in \mathcal{P}$, $n \geq 1$, and $\mathbf{x} \in \mathbb{N}^n$, this notational convention is consistent with the earlier conventions of writing p for both $p^\infty \in \mathcal{P}^\infty$ and for the product probability measure on $\mathbb{N}^n$ corresponding to p.

It is a standard fact that a probability measure q on $\mathbb{N}^\infty$ is completely specified by $q(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{N}^n$ for all $n \geq 1$, subject to the consistency conditions $q(\mathbf{x}) = \sum_{\mathbf{y} \in \mathbb{N}^m \,:\, \mathbf{x} \preceq \mathbf{y}} q(\mathbf{y})$ for all $1 \leq n \leq m$ and $\mathbf{x} \in \mathbb{N}^n$.

$\{0,1\}^*$ denotes the set of binary strings of finite length. The notation $\{0,1\}^* \backslash \emptyset$ is used for the set of binary strings of finite length, excluding the empty string. For $\mathbf{b} \in \{0,1\}^* \backslash \emptyset$, the length of $\mathbf{b}$ is denoted by $l(\mathbf{b})$.

We write $\mathbb{1}(A)$ to denote the indicator of an event A.

It is convenient to state some of the supporting results in this document at a level of generality where the underlying set is a countable set, in which case we denote such a set by $\mathcal{X}$. Also, we will state some results that apply to arbitrary collections of probability measures on $\mathbb{N}^\infty$, i.e. not necessarily of the form $\mathcal{P}^\infty$ for some collection of probability distributions $\mathcal{P}$ on $\mathbb{N}$. In such cases, we denote such a collection of probability measures on $\mathbb{N}^\infty$ by Λ.

If q and r are arbitrary probability measures on $\mathbb{N}^\infty$, then

$$D_n(q||r) := E_q \log \frac{q(X^n)}{r(X^n)},$$

denotes the KL divergence over length n strings of q with respect to r. If p and $\tilde{p}$ are probability distributions on $\mathbb{N}$, then $D(p||\tilde{p})$ denotes the KL divergence of p with respect to $\tilde{p}$, which is $E_p \log \frac{p(X)}{\tilde{p}(X)}$. Note that, with our conventions, the expression $D_n(p||\tilde{p})$ is also well-defined, and can be viewed as a shorthand notation for $D_n(p^\infty||\tilde{p}^\infty)$. We thus have $D_n(p||\tilde{p}) = nD(p||\tilde{p})$ for all $n \in \mathbb{N}$, since p^∞ and $\tilde{p}^\infty$ are *i.i.d.* probability measures on $\mathbb{N}^\infty$. KL divergence is also called *relative entropy*.

For probability distributions p and $\tilde{p}$ on $\mathbb{N}$, their ℓ_1 distance is

$$|p - \tilde{p}|_1 := \sum_{i \in \mathbb{N}} |p(i) - \tilde{p}(i)|.$$

II-A. Strong compressibility and weak compressibility

In the lossless data compression problem for the collection of probability measures $\mathcal{P}^\infty$ on $\mathbb{N}^\infty$ corresponding to a collection of probability distributions $\mathcal{P}$ on $\mathbb{N}$, our estimator is a probability measure q on $\mathbb{N}^\infty$. [1] The problem formulation can be understood by thinking of the *loss* $L(p, q, \mathbf{x})$ incurred by the estimator q against a source p, given the length n observation $\mathbf{x} \in \mathbb{N}^n$, as being the *excess codelength*,

$$L(p, q, \mathbf{x}) := \log \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

The terminology is justified by thinking of $\log \frac{1}{p(\mathbf{x})}$ as an indication of the length of the binary string one would want to use to represent $\mathbf{x}$ in an ideal prefix-free scheme for compressing strings of length n from the source p if one knew what p was, and thinking of $\log \frac{1}{q(\mathbf{x})}$ as the length of the binary string one would be led to use for representing $\mathbf{x}$ in the prefix-free compression scheme suggested by the estimator q. For more on this, see the discussion in Appendix I on how strong and weak compressibility is typically defined in the literature.

With this loss function in mind, we now make the following definitions.

Definition 2. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$, and $\mathcal{P}^\infty$ the corresponding collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. Then $\mathcal{P}^\infty$, or equivalently $\mathcal{P}$, is called *strongly compressible* if there is a probability measure q on $\mathbb{N}^\infty$ satisfying

$$\limsup_{n\to\infty} \sup_{p \in \mathcal{P}^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} = 0. \tag{1}$$

$\square$

The preceding definition may seem unusal relative to the definition of strong compressibility that is traditionally encountered in the literature on data compression [8], [1]. In Appendix I we establish that it is identical to the traditional definition.

[1] It is not required that the probability measure q be generated by *i.i.d.* sampling.

Discussions of data compression in the literature are often framed in the language of *redundancy*. We formalize this notion in the following definition.

Definition 3. Let Λ be any collection of probability measures on $\mathbb{N}^\infty$. The *length-n redundancy* of Λ is defined to be

$$R_n(\Lambda) := \inf_q \sup_{r \in \Lambda} E_r \log \frac{r(X^n)}{q(X^n)}, \tag{2}$$

where the outer infimum is taken over all probability measures on $\mathbb{N}^\infty$, or equivalently over all probability measures on $\mathbb{N}^n$. The redundancy in the special case $n = 1$ is called the *single letter redundancy* of Λ, and $R_n(\Lambda)/n$ is called the *per-symbol length-n redundancy* of Λ. The *asymptotic per-symbol redundancy* of Λ is $\limsup_{n\to\infty} R_n(\Lambda)/n$.

More generally, given a probability measure $\hat{q}_n$ on $\mathbb{N}^n$ one can define the length-n redundancy of Λ with respect to $\hat{q}_n$ to be $\sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{\hat{q}_n(X^n)}$ and similarly for the per-symbol length-n redundancy of Λ with respect to $\hat{q}_n$. Given a probability measure q on $\mathbb{N}^\infty$, one can define the asymptotic-per-symbol redundancy of Λ with respect to q to be $\limsup_{n\to\infty} \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)}$.

Even more generally, given a probability measure $\hat{q}_n$ on $\mathbb{N}^n$ one can define the length-n redundancy of $r \in \Lambda$ with respect to $\hat{q}_n$ to be $E_r \log \frac{r(X^n)}{\hat{q}_n(X^n)}$ and define the per-symbol length-n redundancy of $r \in \Lambda$ with respect to $\hat{q}_n$ similarly. Given a probability measure q on $\mathbb{N}^\infty$, one can define the asymptotic-per-symbol redundancy of $r \in \Lambda$ with respect to q to be $\limsup_{n\to\infty} \frac{1}{n} E_r \log \frac{r(X^n)}{q(X^n)}$.

When $\mathcal{P}$ is a collection of probability distributions on $\mathbb{N}$, and $\mathcal{P}^\infty$ the corresponding collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$, we will talk about each of the redundancy quantities as properties of $\mathcal{P}$ when in fact they are defined for $\mathcal{P}^\infty$. Similarly, given a probability measure $\hat{q}_n$ on $\mathbb{N}^n$ or a probability measure q on $\mathbb{N}^\infty$ we will talk about each of the redundancy quantities for a given $p \in \mathcal{P}$ with respect to $\hat{q}_n$ or q (as appropriate) when we mean the corresponding quantities for the $p^\infty \in \mathcal{P}^\infty$ corresponding to p. □

It is worth noting that a collection of probability distributions on $\mathbb{N}$ is strongly compressible iff its asymptotic per-symbol redundancy is zero. For completeness, we give a proof of this claim in Lemma 31 in Appendix I. We also observe that the asymptotic per-symbol redundancy of a collection of probability measures Λ on $\mathbb{N}^\infty$ can also be written as

$$\limsup_{n\to\infty} R_n(\Lambda)/n = \limsup_{n\to\infty} \frac{1}{n} \inf_q \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)} = \inf_q \limsup_{n\to\infty} \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)},$$

where the infimum on both sides of the equality is over probability measures q on $\mathbb{N}^\infty$. Namely, the $\limsup_{n\to\infty}$ can be interchanged with the $\inf_q$. A proof of this is given in Lemma 36 in Appendix II.

We can allow for much richer collections of probability distributions if we work with a weaker notion of compressibility.

Definition 4. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$, and $\mathcal{P}^\infty$ the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. Then $\mathcal{P}^\infty$, or equivalently $\mathcal{P}$, is called *weakly compressible* if there exists a probability measure q over $\mathbb{N}^\infty$ such that, for all $p \in \mathcal{P}^\infty$ with finite entropy rate, we have

$$\limsup_{n\to\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} = 0. \tag{3}$$

□

One artifact of the above definition is that any collection of probability distributions on $\mathbb{N}$ where every source has infinite entropy is vacuously weakly compressible. In Appendix I we establish that this definition of weak compressibility is identical to the definition of weak compressibility commonly encountered in the literature on data compression, see e.g. Kieffer [16]. Also, in Lemma 32 of Appendix I we formally establish the essentially tautological fact that a collection of probability distributions $\mathcal{P}$ on $\mathbb{N}$ is weakly compressible iff there exists a probability measure q on $\mathbb{N}^\infty$ such that every $p \in \mathcal{P}$ with finite entropy has vanishing asymptotic per-symbol redundancy with respect to q.

II-B. Compression in the data-derived sense

Working with collections of probability distributions on $\mathbb{N}$ that are weakly compressible gives us a richer class of models than working with those that are strongly compressible. Weak compressiblity of a collection $\mathcal{P}$ of probability distributions on $\mathbb{N}$ ensures that there is a probability measure q on $\mathbb{N}^\infty$ such that q is essentially as good an encoder as the underlying p for long enough strings of natural numbers drawn *i.i.d.* from p, where goodness is measured in terms of the number of bits used per symbol encoded. This is what it means to say that the asymptotic per-symbol redundancy of every $p^\infty \in \mathcal{P}^\infty$ with respect to q is 0,

But observe that what one means by "long enough" depends on the unknown p, since convergence to the limit in (3) need not be uniform over $p \in \mathcal{P}$. The main contribution of our work is to come to grips with this issue without having to back off all the way to being able to deal only with strongly compressible collections of probability distributions.

Our ideas are built around the notion of a *universal stopping rule*, which we introduce next. Recall that a stopping rule is a function of observed strings where the decision to *stop* or not at any given time is based only on what has been observed thus far. We formalize a stopping rule by a function τ from $\mathbb{N}^*$, the set of all finite strings of naturals, to the set $\{0, 1\}$,

$$\tau : \mathbb{N}^* \to \{0, 1\}.$$

When τ assigns value 0 on a finite string x^n, possibly the empty string, it indicates that the stopping rule is still waiting after having observed x^n. A string x^n, possibly the empty string, is assigned 1 if the stopping rule has stopped on any prefix of x^n. From a notational point of view, since τ quantifies a stopping rule, we will have for all strings x^n with prefix x^m that $\tau(x^n) \geq \tau(x^m)$.

The stopping rule τ is required to be universal for $\mathcal{P}$. In other words, the stopping rule cannot change depending on the unknown probabilistic model $p \in \mathcal{P}$ that is generating the observations. In the formulation that we will develop in this paper, for a given a threshold $\delta > 0$, a stopping rule, call it τ for now, will be based on some fixed probability measure q on $\mathbb{N}^\infty$, and will signify when the length is "long enough" that the normalized KL divergence between the underlying source distribution and the probability measure q has fallen below δ and will remain below δ henceforth. We will insist that τ stops at a finite time for all $p \in \mathcal{P}$, *i.e.,*

$$p(\lim_{n\to\infty} \tau(X^n) = 1) = 1, \text{ for all } p \in \mathcal{P}. \tag{4}$$

We will include the condition in (4) in the concept of what we mean by a universal stopping rule.

To understand this requirement better, fix a probability measure q on $\mathbb{N}^\infty$, and for $p \in \mathcal{P}$ let

$$\mathcal{N}_{p,\delta;q} := \{n : \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} > \delta\}.$$

Thus $\mathcal{N}_{p,\delta;q}$ is the set of all lengths $n \geq 1$ such that the length-n KL divergence of the *i.i.d.* probability measure p^∞ corresponding to p with respect to the probability measure q is worse than the accuracy required. Now consider the set

$$\mathcal{N}_{\delta;q} := \cup_{p \in \mathcal{P}} \mathcal{N}_{p,\delta;q}.$$

In the trivial case where $\mathcal{N}_{\delta;q}$ is a finite set, let N denote the largest element in $\mathcal{N}_{\delta;q}$. Then, for all $n \geq N$, we have

$$\sup_{p \in \mathcal{P}} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} \leq \delta.$$

Clearly we can choose the stopping rule to be 0 for all sequences with length $n \leq N$ and 1 for all sequences with length $> N$, and this is universal.

It is more interesting when $\mathcal{N}_{\delta;q}$ defined above is not a finite set. Even in this case, the stopping rule τ has to stop at a finite time almost surely no matter which source is governing the observations. Naturally, no matter when τ stops waiting, the sequence length may not be long enough for some sources in $\mathcal{P}$, so τ fails on such sequences. More formally, for $\delta > 0$, τ fails with respect to q or is *δ-premature with respect to q* for a source $p \in \mathcal{P}$ and at time i if there is some string x^i such that

$$\tau(x_1^i) = 1 \text{ and } \frac{1}{i} E_p \log \frac{p(X^i)}{q(X^i)} > \delta. \tag{5}$$

For $p \in \mathcal{P}$, consider the subset of $\mathbb{N}^\infty$ defined as

$$\left\{ x \in \mathbb{N}^\infty : \exists\, i \text{ such that } \tau(x^i) = 1 \text{ and } \frac{1}{i} \sum_{y^i \in \mathbb{N}^i} p(y^i) \log \frac{p(y^i)}{q(y^i)} > \delta \right\}. \tag{6}$$

For $p \in \mathcal{P}$, the above set is the set of strings on which τ is $\delta-$premature with respect to q. While this set depends on which $p \in \mathcal{P}$ is driving the observations, this set is an event in the product σ-algebra on $\mathbb{N}^\infty$ whatever the underlying $p \in \mathcal{P}$. To see this, note that it is a countable union of sets of the form $\left\{x \in \mathbb{N}^\infty : \tau(x^i) = 1\right\}$, $i \geq 1$ (which of the components sets lie in the union is determined, for the fixed probability measure q on $\mathbb{N}^\infty$, by the underlying source probability distribution p).

While the set in (6) may not be an empty set, we can at least try to ensure that its probability under p is small. This thought process leads to what we mean by a collection of probability distributions on $\mathbb{N}$ being weakly compressible in the data-derived sense, formalized below. This is the central concept investigated in this paper.

Definition 5. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the associated collection of probability measures on $\mathbb{N}^\infty$ got by *i.i.d.* sampling from the individual distributions in $\mathcal{P}$. We say that $\mathcal{P}^\infty$, or equivalently $\mathcal{P}$, is *weakly compressible in the data-derived sense* or *data-derived weakly compressible (*d.w.c.*)* if there is a probability measure q on $\mathbb{N}^\infty$ such that, for any accuracy $\delta > 0$ and confidence probability $0 < 1 - \eta < 1$, there is a universal stopping rule $\tau_{\delta,\eta}$ with the property that, no matter what $p^\infty \in \mathcal{P}^\infty$ is in force, we have

$$\begin{aligned} &p(\tau_{\delta,\eta} \text{ is } \delta-\text{premature with respect to } q \text{ for } p) \\ &:= p(\exists\, i \text{ such that } \tau_{\delta,\eta}(X^i) = 1 \text{ and } \frac{1}{i} \sum_{y^i \in \mathbb{N}^i} p(y^i) \log \frac{p(y^i)}{q(y^i)} > \delta) < \eta. \end{aligned} \tag{7}$$

□

Claim 6. (Strongly compressible implies *d.w.c.*) Suppose $\mathcal{P}$ is a collection of probability distributions on $\mathbb{N}$ that is strongly compressible, namely there exists a probability measure q on $\mathbb{N}^\infty$ that satisfies (1). It follows then that, for all $\delta > 0$, the sets

$$N_{\delta;q} := \{n : \sup_{p \in \mathcal{P}^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} > \delta\}$$

are finite. For any $\eta > 0$, suppose we set $\tau_{\delta,\eta}(x^i) = 1$ if $i > \max N_{\delta;q}$ and 0 else, we obtain for all $p \in \mathcal{P}^\infty$ that $p(\tau_{\delta,\eta}$ is $\delta-$premature with respect to $q) = 0$. Thus every strongly compressible collection of probability distributions on $\mathbb{N}$ is *d.w.c.*. □

Claim 7. (*d.w.c.* implies weakly compressible) Suppose $\mathcal{P}$ is a collection of probability distributions on $\mathbb{N}$ that is *d.w.c.*, as in Definition 5. Let q be a probability measure on $\mathbb{N}^\infty$ such that, for every accuracy $\delta > 0$ and confidence probability $0 < 1 - \eta < 1$ there is a universal stopping time $\tau_{\delta,\eta}$ satisfying (7) for every $p \in \mathcal{P}$. Fix $p \in \mathcal{P}$. From (7) we conclude that, for all $i \geq 1$, we have

$$p(\tau_{\delta,\eta}(X^i) = 1) \mathbb{1} \left(\frac{1}{i} \sum_{y^i \in \mathbb{N}^i} p(y^i) \log \frac{p(y^i)}{q(y^i)} > \delta \right) < \eta.$$

However, since the stopping time $\tau_{\delta,\eta}$ is universal, it must satisfy (4), i.e. it stops eventually. Hence we have

$$\lim_{i \to \infty} p(\tau_{\delta,\eta}(X^i) = 1) = 1.$$

From this, it follows that

$$\limsup_{i \to \infty} \frac{1}{i} \sum_{y^i \in \mathbb{N}^i} p(y^i) \log \frac{p(y^i)}{q(y^i)} \leq \delta,$$

(in fact, for this to hold, it suffices to have the condition in (7) hold for some $0 < 1 - \eta < 1$ and not necessarily for all $\eta > 0$, for the given $\delta > 0$). Letting $\delta \to 0$, we see that the condition in (3) holds, for the given probability measure q on $\mathbb{N}^\infty$, for all $p \in \mathcal{P}$. This means, by definition, that $\mathcal{P}$ is weakly compressible. □

Claims 6 and 7 imply that

$$\text{Strongly compressible} \subseteq \textit{d.w.c.} \subseteq \text{ weakly compressible.}$$

In Section V-A we will see examples of model classes demonstrating that each of these inclusions is strict.

As can be seen from the preceding discussion, our formulation of *d.w.c.* model classes is aimed at addressing the most interesting case from a statistical modeling viewpoint, which is the case where $\mathcal{P}^\infty$ is weakly compressible, but not strongly compressible. Typically, we need global constraints on the collection of sources that comprise a model class to render the model class strongly compressible – for example, that the square root of the Fisher information be integrable over the model class for a class to be strongly compressible [10]. By contrast, as we will see, data-derived weak compressibility does not depend on controlling the entire class $\mathcal{P}^\infty$, but requires only that local neighborhoods of each $p \in \mathcal{P}$, viewed as a member of $\mathcal{P}$, be simple. Indeed, one of the main contributions of this paper is to obtain a condition that is both necessary and sufficient for an *i.i.d.* collection $\mathcal{P}^\infty$ to be *d.w.c.*.

The operational interpretation for our formulation of *d.w.c.* model classes comprised of *i.i.d.* sources can be articulated as follows. Given such a model class, let q be any measure on $\mathbb{N}^\infty$ that verifies the definition, i.e. such that for every $\delta > 0$ and $\eta > 0$ there is some universal stopping rule $\tau_{\delta,\eta} : \mathbb{N}^* \mapsto \{0,1\}$ for which the probability under every p in the model class that $\tau_{\delta,\eta}$ is δ-premature with respect to q for p is less than η.

As we observe the realization of the *i.i.d.* data samples from the (unknown) source p in the model class, we will eventually see a string of some (random) length $n = n(\delta, \eta, p)$ (say x^n) such that $\tau_{\delta,\eta}(x^n) = 1$. Now, even though we do not know p, we get the guarantee (with confidence probability $\geq 1 - \eta$) that using q to compress any subsequent length-n or longer sequence of symbols in the usual way (*i.e.,* $-\log q(x^k)$ bits for a sequence x^k) incurs an expected per-symbol redundancy $\leq \delta$.

II-C. *Other examples of data-derived problem formulations*

To clarify that the ideas in our framework have the potential to apply much more broadly to estimation problems other than the lossless compression problem that we have focused on in this document, we highlight in this section data-derived formulations for two other estimation problems. The first is a prediction task from [2], which we call the *insurance* problem, while the second is an entropy estimation task. In later sections, we will also make some comparisons between the insurance problem and the universal lossless compression problem studied here.

Example 8. **(Insurability)** Suppose we have a collection $\mathcal{P}^\infty$ of *i.i.d.* measures over $\mathbb{N}^\infty$. Given a finite sample $(X_1, \ldots, X_n)$ with *i.i.d.* marginals from an unknown $p \in \mathcal{P}$ we want to estimate a finite upper bound on the next symbol X_{n+1} in a data-derived sense. If there are $p \in \mathcal{P}$ with unbounded support then for any finite upper bound we propose there is a probability under such p that it may not be valid. In our data-derived formulation, we therefore want to provide an estimated upper bound $\Phi(X_1^n)$, and a universal stopping rule τ that tells us from what point we should believe that our estimates $\Phi(X_1^n)$ are at least as big as X_{n+1}, while allowing for some probability of being wrong.

Formally, given a confidence probability $0 < 1-\eta < 1$, we seek to come up with a mapping $\Phi : \mathbb{N}^* \to \mathbb{R}$ and a stopping rule τ such that, for all $p \in \mathcal{P}$, we have

$$p\big(\,\exists i \in \mathbb{N} \text{ such that } \Phi(X^i) < X_{i+1} \text{ and } \tau(X^i) = 1\big) < \eta.$$

If this is possible, we say that the model class $\mathcal{P}^\infty$ is *insurable.* In prior work, in [2], the collections $\mathcal{P}^\infty$ that are insurable were completely characterized. See Corollary 19 and Corollary 20 for more details and connections with the results developed in this document. □

Example 9. **(Entropy estimation)** Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$. Given a finite sample $(X_1, \ldots, X_n)$ sampled *i.i.d.* from an unknown $p \in \mathcal{P}$, we want to provide a data-derived finite upper bound $\hat{H}$ on the entropy of p. Formally, given a confidence probability $0 < 1 - \eta < 1$, we would like to come up with a mapping $\hat{H} : \mathbb{N}^* \to \mathbb{R}$ and a universal stopping rule τ such that, for all $p \in \mathcal{P}$, we have

$$p\Big(\,\exists i \in \mathbb{N} \text{ such that } \hat{H} < H(p) \text{ and } \tau(X^i) = 1\Big) < \eta. \qquad \square$$

We do not know the answer to this question, in the sense that we do not know a simple intuitive necessary and sufficient condition that will characterize which collections $\mathcal{P}$ of probability distributions on $\mathbb{N}$ admit data-derived estimates of entropy and which do not.

III. Background

This section highlights some interesting prior results on universal compression that will be used in this paper. Readers can skip the proofs in this section if they are willing to take the results here at face value when they are referred to. We have collected in this section the more interesting prior results we use. Other, more basic, prior results that we also use are collected in Appendix II.

III-A. *Weak compression*

Let $\mathcal{P}$ be a collection of probability distribution on $\mathbb{N}$ and $\mathcal{P}^\infty$ the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. In Appendix I we have demonstrated that the notion of weak compressibility of $\mathcal{P}^\infty$ in the sense of Kieffer [16] is identical to the definition of weak compressibility of $\mathcal{P}^\infty$ that we have made in Definition 4.

The following lemma gives a useful characterization of weak compressibility.

Lemma 10. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the associated set of *i.i.d.* probability measures on $\mathbb{N}^\infty$. Then $\mathcal{P}^\infty$ is weakly compressible iff there exists a distribution q on $\mathbb{N}$ such that for all $p \in \mathcal{P}$ with finite entropy we have

$$\sum_{x\in\mathbb{N}} p(x) \log \frac{1}{q(x)} < \infty. \tag{8}$$

Proof From [16, Theorem 1] we know that $\mathcal{P}^\infty$ is weakly compressible iff there is a countable set $\mathcal{Q} := \{q_1, q_2, \ldots\}$ of probability distributions on $\mathbb{N}$ such that for all $p \in \mathcal{P}$ with finite entropy there is some $q_i \in \mathcal{Q}$ satisfying

$$\sum_{x\in\mathbb{N}} p(x) \log \frac{1}{q_i(x)} < \infty.$$

Therefore, if there is a probability distribution q on $\mathbb{N}$ satisfying (8) for all $p \in \mathcal{P}$, we can immediately conclude that $\mathcal{P}^\infty$ is weakly compressible. It remains to show the converse.

To do this, suppose that $\mathcal{P}^\infty$ is weakly compressible and let $\mathcal{Q}$ be a choice of the countable set of probability distributions on $\mathbb{N}$ guaranteed by [16, Theorem 1]. Fix some enumeration of $\mathcal{Q}$ as $\mathcal{Q} = \{q_1, q_2, \ldots\}$.

Consider the probability distribution q on $\mathbb{N}$ given by

$$q(n) := \frac{\sum_{i=1}^{|\mathcal{Q}|} \frac{q_i(n)}{i(i+1)}}{\sum_{j=1}^{|\mathcal{Q}|} \frac{1}{j(j+1)}}, \quad n \in \mathbb{N},$$

where the upper limit of the summation is understood to be ∞ if $\mathcal{Q}$ is countably infinite. Observe that, for all i and for all n, we have

$$q(n) \geq \frac{q_i(n)}{i(i+1)}.$$

Therefore, for all $p \in \mathcal{P}$ with finite entropy and all $q_i \in \mathcal{Q}$, we have

$$\sum_{x\in\mathbb{N}} p(x) \log \frac{1}{q(x)} \leq \sum_{x\in\mathbb{N}} p(x) \log \frac{i(i+1)}{q_i(x)}.$$

Since the right hand side of the preceding equation is finite for at least one $q_i \in \mathcal{Q}$, this completes the proof. $\square$

III-B. Finite redundancy implies tightness

Let us recall the definition of *tightness* of a collection of probability distributions on $\mathbb{N}$.

Definition 11. A collection $\mathcal{P}$ of probability distributions on $\mathbb{N}$ is said to be tight if for every $\gamma > 0$ there is a natural number M_γ such that

$$\sup_{p\in\mathcal{P}} p(X > M_\gamma) < \gamma.$$

□

We now show that tightness of a collection of probability distributions on $\mathbb{N}$ is implied by finiteness of the single letter redundancy of the collection. The result we present is a well-known folk theorem, see for example [17, Lemma 4]. Here we give an elementary proof of this result.

Lemma 12. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$. If the single letter redundancy of $\mathcal{P}$ is finite, then $\mathcal{P}$ is tight.

Proof We reproduce the proof from [18, Lemma 1]. Since $\mathcal{P}$ has finite single letter redundancy, there is a probability distribution q on $\mathbb{N}$ such that

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

Proposition 33 in Appendix II implies that for all $p \in \mathcal{P}$ we have

$$E_p|\log\frac{p(X)}{q(X)}| \le \dot{R} + 2(\log e)/e.$$

Hence, for all $p \in \mathcal{P}$ and all integers $m \ge 1$, we have

$$p(\left|\log\frac{p(X)}{q(X)}\right| > m) \le (\dot{R} + (2\log e)/e)/m. \tag{9}$$

To complete the argument, we need to define the *linearly interpolated cumulative distribution function* of a probability distribution on $\mathbb{N}$.

Definition 13. For a probability distribution q on $\mathbb{N}$, the linearly interpolated cumulative distribution $\dot{F}_q(n)$ for $n \in \mathbb{N}\cup\{0\}$ follows the standard definition of the cumulative distribution function, i.e.

$$\dot{F}_q(n) := F_q(n) = \mathbb{P}(X \le n) \tag{10}$$

where X is a random variable distributed according to q. For $n \in \mathbb{N}\cup\{0\}$ and a real number $n \le x \le n+1$, however, we define

$$\dot{F}_q(x) := (n+1-x)\dot{F}_q(n) + (x-n)\dot{F}_q(n+1).$$

Note that $\dot{F}_q$ is a nondecreasing function with domain the nonnegative real numbers and range either $[0,1]$ or $[0,1)$. For $t \in [0,1)$, we define $\dot{F}_q^{-1}(t)$ to be the right continuous inverse of $\dot{F}_q$, i.e.

$$\dot{F}_q^{-1}(t) := \sup\{x \ge 0 : \dot{F}_q(x) \le t\}.$$

□

Given $\gamma > 0$, pick m so large that $(\dot{R} + (\log e)/e)/m < \gamma/2$. For all $p \in \mathcal{P}$, we then have

$$\begin{aligned} p(X > \dot{F}_q^{-1}(1-\gamma/2^{m+1})) &= p(\log \frac{p(X)}{q(X)} > m, X > \dot{F}_q^{-1}(1-\gamma/2^{m+1})) + p(\log \frac{p(X)}{q(X)} \leq m, X > \dot{F}_q^{-1}(1-\gamma/2^{m+1})) \\ &\leq p(|\log \frac{p(X)}{q(X)}| > m) + 2^m q(X > \dot{F}_q^{-1}(1-\gamma/2^{m+1})) \\ &< (\dot{R} + (\log e)/e)/m + \gamma/2 \\ &< \gamma. \end{aligned}$$

This establishes that $\mathcal{P}$ is tight. □

III-C. Bounds on redundancy

The following technical lemma is used in Example 23 and in Example 27. Its roots go back to [13].

Lemma 14. Let $\mathcal{X}$ be a countable set, and $\mathcal{P}$ be a collection of probability distributions on $\mathcal{X}$. For i ranging over the finite set of indices $\{1, \ldots, M\}$ or over all indices $i \geq 1$, let $S_i \subset \mathcal{X}$ be a subset of $\mathcal{X}$, and assume that these sets are pairwise disjoint. Suppose that for each i there exists $p_i \in \mathcal{P}$ such that

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

Then, for all probability distributions q on $\mathcal{X}$, we have

$$\sup_{p \in \mathcal{P}} D(p||q) \geq \delta \log(M) - 1,$$

if the number of subsets in the collection is finite, equal to M, and

$$\sup_{p \in \mathcal{P}} D(p||q) = \infty,$$

if the number of subsets in the collection is infinite.

Proof This is a simplified formulation of the *distinguishability* concept in [13]. To prove the claim, note that for any m at most equal to the number of subsets in the collection, we must have $q(S_i) \leq 1/m$ for some i. For such a choice of i we can write

$$\begin{aligned} D(p_i||q) &= \sum_{x \in S_i} p_i(x) \log \frac{p_i(x)}{q(x)} + \sum_{x \in S_i^c} p_i(x) \log \frac{p_i(x)}{q(x)} \\ &\overset{(a)}{\geq} p_i(S_i) \log \frac{p_i(S_i)}{q(S_i)} + p_i(S_i^c) \log \frac{p_i(S_i^c)}{q(S_i^c)} \\ &\geq p_i(S_i) \log \frac{1}{q(S_i)} + p_i(S_i^c) \log \frac{1}{q(S_i^c)} - 1 \\ &\geq \delta \log m - 1, \end{aligned}$$

where step (a) is from the log sum inequality. This completes the proof. □

IV. Characterization of *d.w.c.* model classes

In this section we state our primary result, which is a necessary and sufficient condition for a model class comprised of a collection of probability distributions $\mathcal{P}$ on $\mathbb{N}$ to be data-derived weak compressible.

We will see that what decides whether a model class $\mathcal{P}$ is *d.w.c.* or not is a *local* property of the probability distributions in $\mathcal{P}$, viewed as members of $\mathcal{P}$. Namely, the characterization of data-derived weak compressibility is based on considering a property of local neighborhoods, as defined in Section IV-A, of the individual probability distributions in the model class. Distributions having bad local neighborhoods are what we call *deceptive* distributions, defined and studied in detail in Section IV-B. The notion of deceptive distributions lies at the heart of our characterization, in Theorem 17, of which model classes are *d.w.c.*.

IV-A. Local neighborhoods

We will see in this section that what makes the local neighborhoods of a probability distribution $p \in \mathcal{P}$ bad and kills *d.w.c.* is that when a stopping rule is forced by $p^\infty \in \mathcal{P}^\infty$ into certifying the accuracy of the estimate at some time (which will have to be the case, since the stopping rule has to stop with probability 1 under p), it will nevertheless be the case that there are other probability distributions in $\mathcal{P}$, potentially arbitrarily close to p, which induce inadequate performance on the estimator. We now proceed to make this vague description of the underlying ideas precise.

For probability distributions p and $\tilde{p}$ on $\mathbb{N}$, we define

$$\mathcal{J}(p, \tilde{p}) := D\left(p || \frac{p+\tilde{p}}{2}\right) + D\left(\tilde{p} || \frac{p+\tilde{p}}{2}\right), \tag{11}$$

which, up to a scaling factor, is a Jensen-Shannon divergence between p and $\tilde{p}$ [19]. The primary reason we use the Jensen-Shannon divergence in place of the KL divergence is that $\mathcal{J}$ satisfies a pseudo-triangle inequality, as shown in Lemma 37 in Appendix II, while still retaining much of the statistical interpretation that a KL divergence has. Lemma 37, reproduced from [2], also shows that $\mathcal{J}$ is intimately connected with the ℓ_1-distance between probability distributions on $\mathbb{N}$. Indeed, $\mathcal{J}$ generates the same topology on the set of probability distributions on $\mathbb{N}$ that the ℓ_1-distance does.

More generally, for probability measures q and r on $\mathbb{N}^\infty$, we use the notation

$$\mathcal{J}(q, r) := D_1\left(q || \frac{q+r}{2}\right) + D_1\left(r || \frac{q+r}{2}\right),$$

where, in the above, the KL divergences are taken between the single letter marginals of q and r on the first sample. Note that in this case it is no longer necessary that q should equal r when $\mathcal{J}(q, r) = 0$. Also note that this notation is consistent with our convention of using the notation p to represent both a probability distribution on $\mathbb{N}$ and the corresponding $p^\infty \in \mathcal{P}^\infty$.

Definition 15. An $\epsilon-$*neighborhood* of $p \in \mathcal{P}$ is the set $B(p, \epsilon; \mathcal{P})$ of all $p' \in \mathcal{P}$ such that $\mathcal{J}(p, p') < \epsilon$. □

For technical reasons, we will also make use of ℓ_1-neighborhoods in the paper in addition to the $\epsilon-$neighborhoods defined above (which we will refer to simply as *neighborhoods*). The ℓ_1-neighborhood of radius $\epsilon > 0$ around $p \in \mathcal{P}$ is comprised of all $p' \in \mathcal{P}$ such that $|p - p'|_1 < \epsilon$.

IV-B. Deceptive distributions

Definition 16. $p^\infty \in \mathcal{P}^\infty$ is said to be *deceptive* if the asymptotic per-symbol redundancy of neighborhoods of p is bounded away from 0 in the limit as the neighborhood shrinks to 0. More precisely, we define $p^\infty \in \mathcal{P}^\infty$, or equivalently $p \in \mathcal{P}$, to be deceptive if

$$\lim_{\epsilon\to 0} \inf_q \limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon;\mathcal{P})} \frac{1}{n} D_n(p'||q) > 0. \tag{12}$$

In the above, the infimum is over all q that are probability measures on $\mathbb{N}^\infty$ (not necessarily obtained by *i.i.d.* sampling). The verbal description of this condition in terms of the asymptotic per-symbol redundancy of the neighborhoods of p is justified by Lemma 36, which is proved in Appendix II. $\square$

Our main result is the following Theorem 17. The necessity part of this theorem is proved in Section VI and the sufficiency part in Section VII.

Theorem 17. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the associated collection of probability measures on $\mathbb{N}^\infty$ got by *i.i.d.* sampling. Then $\mathcal{P}^\infty$ is *d.w.c.* iff no $p \in \mathcal{P}$ is deceptive. $\square$

In the rest of this section we explore the concept of deceptive distributions to flesh out a few properties of such distributions and their neighborhoods. This will help to better understand Definition (12) and will set the stage for understanding the proof of Theorem 17.

IV-B.1) A simpler characterization of deceptive distributions

In determining whether a source $p \in \mathcal{P}$ is deceptive, (12) allows us to choose q depending on ϵ. We now show that this degree of freedom is unnecessary.

Lemma 18. If $p \in \mathcal{P}$ is not deceptive, then there is a single probability measure q^* on $\mathbb{N}^\infty$ such that

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon;\mathcal{P})} \frac{1}{n} D_n(p'||q^*) = 0.$$

Hence, we have that p is deceptive iff for all probability measures q on $\mathbb{N}^\infty$ we have

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon;\mathcal{P})} \frac{1}{n} D_n(p'||q) > 0.$$

Proof Because p is not deceptive, there exists a sequence $(\delta_m > 0, m \geq 1)$, with $\lim_{m\to\infty} \delta_m \to 0$, and a sequence of probability measures $(q_m, m \geq 1)$ on $\mathbb{N}^\infty$ such that, for all sufficiently large $m \geq 1$, we have

$$\limsup_{n\to\infty} \sup_{p' \in B(p,1/m;\mathcal{P})} \frac{1}{n} D_n(p'||q_m) \leq \delta_m.$$

Define the probability measure q^* on $\mathbb{N}^\infty$ that, for each $n \geq 1$ and $\mathbf{x} \in \mathbb{N}^n$, assigns to the string $\mathbf{x}$ the probability

$$q^*(\mathbf{x}) := \sum_{m\geq 1} \frac{q_m(\mathbf{x})}{m(m+1)}.$$

For all $m \geq 1$, $n \geq 1$ and $p' \in B(p, 1/m; \mathcal{P})$, we have

$$\frac{1}{n} D_n(p'||q^*) \leq \frac{1}{n} D_n(p'||q_m) + \frac{\log(m(m+1))}{n}.$$

This implies that

$$\limsup_{n\to\infty} \sup_{p'\in B(p,1/m;\mathcal{P})} \frac{1}{n} D_n(p'||q^*) \leq \delta_m + \lim_{n\to\infty} \frac{\log(m(m+1))}{n} = \delta_m,$$

and so

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \sup_{p'\in B(p,\epsilon;\mathcal{P})} \frac{1}{n} D_n(p'||q^*) = \lim_{m\to\infty} \limsup_{n\to\infty} \sup_{p'\in B(p,1/m;\mathcal{P})} \frac{1}{n} D_n(p'||q^*) \leq \lim_{m\to\infty} \delta_m = 0.$$

This concludes the proof. □

IV-B.2) Neighborhoods of non-deceptive distributions are tight

Recall the definition of *tightness* of a collection of probability distributions on $\mathbb{N}$ from Definition 11. The following corollary is immediate.

Corollary 19. If $p \in \mathcal{P}$ is not deceptive, then some neighborhood of p is tight.

Proof If $p \in \mathcal{P}$ is not deceptive then, for some $\epsilon > 0$, there exists $n \geq 1$ and a probability measure q on $\mathbb{N}^\infty$ such that

$$\sup_{p'\in B(p,\epsilon)} D_n(p'||q) < \infty.$$

From Proposition 34 in Appendix II, it follows that the single letter redundancy of the neighborhood $B(p,\epsilon)$ is finite, which implies that $B(p,\epsilon)$ is tight, from Lemma 12. □

The above corollary helps to make a connection between two data-derived formulations – *d.w.c.*, which is considered in this document, and insurability, from Example 8. We showed in [2] that a collection of *i.i.d.* probability measures $\mathcal{P}^\infty$ on $\mathbb{N}^\infty$ is insurable iff some neighborhood, exactly as defined here, of every $p \in \mathcal{P}$ is tight. We therefore obtain

Corollary 20. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and let $\mathcal{P}^\infty$ denote the associated collection of *i.i.d.* probability measures on $\mathbb{N}^\infty$. If $\mathcal{P}^\infty$ is *d.w.c.*, then $\mathcal{P}^\infty$ is insurable. □

In both cases, note that the condition relies on some neighborhood within the model class of every model being simple. We expect this kind of locality to appear as a feature of the characterization of which model classes admit data-derived estimators in most data-derived formulations.

V. Examples

We now discuss a series of examples that highlight various aspects of our formulation. These examples also help flesh out the notion of what it means for a probability distribution to be deceptive.

V-A. Strongly compressible $\subsetneq$ d.w.c. $\subsetneq$ *weakly compressible*

We first give examples showing that weakly compressible collections of probability distribution on $\mathbb{N}$ are a strictly richer class of models than *d.w.c.* collections. We also show that there are collections of probability distributions on $\mathbb{N}$ that are *d.w.c.* but are not strongly compressible.

Weakly compressible but not *d.w.c.*

We consider two examples in this category.

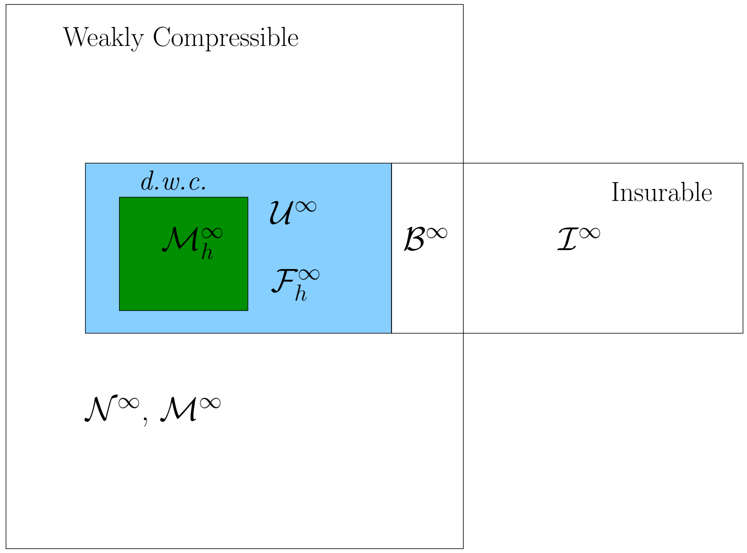

Fig. 1. Summary of examples: $\mathcal{M}_h^\infty$ is strongly compressible (hence *d.w.c.*, insurable and weakly compressible), $\mathcal{U}^\infty$ and $\mathcal{F}_h^\infty$ are *d.w.c.* (hence insurable and weakly compressible), $\mathcal{B}^\infty$ is weakly compressible and insurable but not *d.w.c.*, $\mathcal{N}^\infty$ and $\mathcal{M}^\infty$ are weakly compressible, but not insurable nor *d.w.c.*, while $\mathcal{I}^\infty$ is insurable but not weakly compressible. Note that Corollary 20 shows that all *d.w.c.* collections are insurable, while Claim 6 and Claim 7 show that strong compressibility implies *d.w.c.* and that *d.w.c.* implies weak compressiblity respectively.

A monotone probability distribution p on $\mathbb{N}$ is one that satisfies $p(y) \geq p(y+1)$ for all $y \in \mathbb{N}$. Let $\mathcal{M}$ denote the collection of all monotone probability distributions on $\mathbb{N}$ and $\mathcal{M}^\infty$ be the corresponding collection of *i.i.d.* probability measures on $\mathbb{N}^\infty$.

Example 21. **($\mathcal{M}^\infty$ is weakly compressible but not *d.w.c.*.)**

To see that $\mathcal{M}^\infty$ is weakly compressible [20] note that, for all $p \in \mathcal{M}$ and all $n \in \mathbb{N}$, we have

$$p(n) \leq \frac{1}{n}.$$

It follows that every $p \in \mathcal{M}$ with finite entropy must satisfy

$$\sum_{n \geq 1} p(n) \log n \leq \sum_{n \geq 1} p(n) \log \frac{1}{p(n)} < \infty. \tag{13}$$

Now consider the probability distribution q on $\mathbb{N}$ assigning probability $q(n) = \frac{6}{\pi^2 n^2}$ to $n \in \mathbb{N}$. From (13) we see that, for all $p \in \mathcal{M}$ with finite entropy, we have

$$\sum_{n \geq 1} p(n) \log \frac{1}{q(n)} < \infty.$$

From Lemma 10 we conclude that $\mathcal{M}^\infty$ is weakly compressible.

It turns out that all the probability distributions $p \in \mathcal{M}$ are deceptive. To conclude this, we show that no neighborhood around any $p \in \mathcal{M}$ is tight and then appeal to Corollary 19. This would then imply, by Theorem 17, that $\mathcal{M}^\infty$ is not *d.w.c.*. In fact, it would have been enough to show that there exists some $p \in \mathcal{M}$ such that that no neighborhood of p is tight.

Let $\mathcal{U}$ denote the collection of all uniform distributions over finite supports of form $\{m, m+1, \ldots, M\}$ where m and M are positive integers with $m \leq M$. For $p \in \mathcal{M}$ and $\epsilon > 0$, consider the collection

$$\mathcal{M}(p, \epsilon) := \{p' : p' = (1-\alpha)p + \alpha q \text{ for } q \in \mathcal{U} \cap \mathcal{M} \text{ and } 0 \leq \alpha < \epsilon\}. \tag{14}$$

In (14) q can be any monotone uniform distribution, namely a uniform distribution with support $\{1, \ldots, M\}$ for some $M > 0$. Clearly $\mathcal{M}(p, \epsilon) \subset \mathcal{M}$. Note also that $\mathcal{M}(p, \epsilon)$ is a subset of an ℓ_1-neighborhood of p corresponding to ℓ_1-distance 2ϵ. We will show that $\mathcal{M}(p, \epsilon)$ is not tight for all p and all $\epsilon > 0$. By Lemma 37 and the definition of neighborhoods in Definition 15, it follows that no neighborhood of any $p \in \mathcal{M}$ is tight.

For $0 < \alpha < \epsilon$, let $0 < \delta < \alpha$ and $n \geq 1$. Observe that if the support $\{1, \ldots, M\}$ of a uniform distribution $q' \in \mathcal{U} \cap \mathcal{M}$ satisfies $M \geq \frac{n}{1-\frac{\delta}{\alpha}}$, then we have

$$q'\{j : j > n\} = 1 - \frac{n}{M} \geq \frac{\delta}{\alpha}.$$

Thus, given any $p \in \mathcal{M}$, we have a distribution $p' = (1-\alpha)p + \alpha q' \in \mathcal{M}(p, \epsilon)$ that satisfies $p'\{j : j > n\} \geq \delta$. Therefore, $\mathcal{M}(p, \epsilon)$ is not tight. This completes the argument. □

For our second example, we consider the set $\mathcal{N}_1^\infty$ of all *i.i.d.* probability measures on $\mathbb{N}^\infty$ corresponding to the set of all probability distributions p on $\mathbb{N}$ such that $E_p X < \infty$, denoted $\mathcal{N}_1$.

Example 22. **($\mathcal{N}_1^\infty$ is weakly compressible but not *d.w.c.*.)**

Note that every $p \in \mathcal{N}_1$ has finite entropy. Also, by definition, all $p \in \mathcal{N}_1$ satisfy $\sum_{i \geq 1} i p_i < \infty$. Therefore the simplified version of Kieffer's condition for weak compressibility, as stated in Lemma 10, is satisfied by the distribution $q(i) := 1/2^i$ $(i \geq 1)$. Thus we conclude that $\mathcal{N}_1$ is weakly compressible.

We can show that every $p \in \mathcal{N}_1$ is deceptive by showing that no neighborhood of any $p \in \mathcal{N}_1$ is tight. The approach is similar to that in Example 21. Given $\epsilon > 0$, consider distributions of the form $p' = (1-\alpha)p + \alpha q$, where $q \in \mathcal{U}$ is a uniform distribution over a support of the form $\{m, m+1, \ldots, M\}$, and $0 < \alpha < \epsilon$. Since q has finite support, we have $p' \in \mathcal{N}_1$.

As in Example 21 we observe that (i) the ℓ_1 distance between p' and q is strictly less than 2ϵ; (ii) for all $0 < \delta < \alpha$ and $n \geq 1$, we can pick $q' \in \mathcal{U}$, with $\mathcal{U}$ defined as in Example 21, whose support satisfies $M \geq \frac{n}{1-\frac{\delta}{\alpha}}$, which then implies that the $(1-\delta)-$percentile of $p' := (1-\alpha)p + \alpha q'$ can be made to lie above n. Since the above construction works for arbitrary $n \geq 1$ and in view of Lemma 37 and the way in which neighborhoods are defined in Definition 15, no neighborhood of any $p \in \mathcal{N}_1$ is tight, which shows that every $p \in \mathcal{N}_1$ is deceptive and hence, by Theorem 17, that $\mathcal{N}_1$ cannot be *d.w.c.*. As in Example 21, to apply Theorem 17 it would have been enough to show that there is at least one $p \in \mathcal{N}_1$ which is deceptive. □

d.w.c. but not strongly compressible

The example we consider in this category is $\mathcal{U}$, which is defined in Example 21. Let $\mathcal{U}^\infty$ denote the collection of all *i.i.d.* probability measures on $\mathbb{N}^\infty$ corresponding to $\mathcal{U}$.

Example 23. **($\mathcal{U}^\infty$ is not strongly compressible but is *d.w.c.*.)**

We first show that $\mathcal{U}$ has infinite single letter redundancy. To see this, we partition $\mathbb{N}$ into disjoint subsets $(T_i, i \geq 0)$, where $T_i := \{2^i, \ldots, 2^{i+1} - 1\}$. For each T_i there is an associated distribution $p_i \in \mathcal{U}$

such that $p_i(T_i) = 1$. Since the number of these disjoint sets T_i is infinite, we conclude from Lemma 14 that the single redundancy of $\mathcal{U}$ is ∞.

From the second part of Proposition 34 we can now conclude that the length-n redundancy of $\mathcal{U}$ is ∞ for all $n \geq 1$, so its asymptotic per-symbol redundancy is also ∞, which means, by Lemma 31, that $\mathcal{U}$ is not strongly compressible.

To see that $\mathcal{U}$ is *d.w.c.*, note that around each probability distribution $p \in \mathcal{U}$ there is an ℓ_1-neighborhood that contains no other probability distribution in $\mathcal{U}$. Such a neighborhood has length-n redundancy equal to 0 for all n because the only possible distribution in the neighborhood is p. Hence the asymptotic per-symbol redundancy of all sufficient small neighborhoods of each $p \in \mathcal{U}$ is zero, which means, by definition, that each $p \in \mathcal{U}$ is not deceptive, see Definition 16. □

Strongly compressible and *d.w.c.*

For completeness we next give an example of a collection of probability distributions on $\mathbb{N}$ which is strongly compressible, hence automatically *d.w.c.*.

For $h > 0$, we consider the set $\mathcal{M}_h \subset \mathcal{M}$ of all monotone probability distributions on $\mathbb{N}$ where the second moment of the self information satisfies the bound

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

Let $\mathcal{M}_h^\infty$ denote the set of all *i.i.d.* probability measures on $\mathbb{N}^\infty$ corresponding to $\mathcal{M}_h$.

Example 24. **($\mathcal{M}_h^\infty$ is strongly compressible, hence *d.w.c.*.)**

Note that for any monotone probability distribution p on $\mathbb{N}$ and all $i \geq 1$ we have $p(i) \leq 1/i$. Therefore for any $p \in \mathcal{M}_h$, if X is a random variable taking values in $\mathbb{N}$ with the probability distribution p, we have

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

Therefore, for all $p \in \mathcal{M}_h$, we have by the Cauchy-Schwarz inequality that $E_p \log X \leq \sqrt{h}$. Now, for the probability distribution q on $\mathbb{N}$ given by $q(i) = \frac{1}{i(i+1)}, i \geq 1$, we have

$$\sup_{p\in\mathcal{M}_h} E_p\left(\lceil\log\frac{1}{q(X)}\rceil\right)^2 \leq \sup_{p\in\mathcal{M}_h} E_p\big(\log(X^2+X)+1\big)^2 \leq \sup_{p\in\mathcal{M}_h} E_p(2\log X+2)^2 \leq 4(\sqrt{h}+1)^2,$$

where the last inequality follows because, for all $p \in \mathcal{M}_h$, we have

$$E_p(2\log(X)+2)^2 = 4E_p\big(\log^2(X)+2\log X+1\big) \leq 4(h+2\sqrt{h}+1) = 4(\sqrt{h}+1)^2.$$

Therefore (see Appendix III for a proof), we can construct a probability measure q^* on $\mathbb{N}^\infty$ such that

$$\sup_{p\in\mathcal{M}_h^\infty} \frac{1}{n} D_n(p||q^*) \leq \frac{2h^{\frac{1}{4}}(\sqrt{h}+1)}{\sqrt{\ln n}} + \pi\sqrt{\frac{2}{3n}}\log e.$$

From this it follows that the collection $\mathcal{M}_h^\infty$ is strongly compressible, and therefore *d.w.c.* trivially from Claim 6. □

Comparing Examples 21 and 24, we observe, that countable unions of *d.w.c.* model classes need not be *d.w.c.*. In fact, as we will see in Example 27, even finite unions of *d.w.c.* model classes need not be *d.w.c.*.

V-B. d.w.c. *collections*

Thus far, we have seen two *d.w.c.* class – $\mathcal{U}^\infty$ and $\mathcal{M}_h^\infty$. But neither is completely satisfying. In the collection $\mathcal{U}$ above, there was a neighborhood around each probability measure $p \in \mathcal{U}$ with no other element of $\mathcal{U}$. Thus $\mathcal{U}$ trivially satisfied the local condition characterizing *d.w.c.* in Theorem 17. The $\mathcal{M}_h$ case falls into another extreme – the entire model collection $\mathcal{M}_h$ is strongly compressible, and therefore the condition characterizing *d.w.c.* in Theorem 17 was again satisfied in a trivial way.

We now therefore construct two additional examples of *d.w.c.* model classes that are much more interesting. Our first example is of *d.w.c.* model classes $\mathcal{F}_h$, where neither of the two extreme situations mentioned above holds. Our second example is of a *d.w.c.* model class $\mathcal{H}$ with a source none of whose neighborhoods are strongly compressible, but where the asymptotic per-symbol redundancy diminishes to 0 as the neighborhood shrinks to the defining probability distribution.

More interesting *d.w.c.* model classes

For a probability distribution p on $\mathbb{N}$ and a number $M > 0$, define the probability measure

$$p^{(M)}(n) := \begin{cases} p(n-M) & n \geq M+1 \\ 0 & \text{else.} \end{cases}$$

Namely, $p^{(M)}$ shifts p to the right by M. Furthermore, let the *span* of any probability distribution p on $\mathbb{N}$ having finite support be defined to be the largest natural number which has non-zero probability under p.

For $h > 0$, we consider the model classes

$$\mathcal{F}_h := \left\{ (1-\epsilon)p_1 + \epsilon p_2^{(\mathrm{span}(p_1)+1)} : p_1 \in \mathcal{U}, p_2 \in \mathcal{M}_h \text{ and } 0 < \epsilon < 1 \right\}.$$

As usual, let $\mathcal{F}_h^\infty$ denote the set of *i.i.d.* probability measures on $\mathbb{N}^\infty$ associated to $\mathcal{F}_h$. Note that the initial uniform component of any $p \in \mathcal{F}_h$ is uniquely determined.

Example 25. ($\mathcal{F}_h^\infty$ **is *d.w.c.*.**)

Proof Let the *base* of any probability distribution over the naturals be the smallest natural number which has non-zero probability. Consider any probability distribution $p = (1-\epsilon)p_1 + \epsilon p_2^{(\mathrm{span}(p_1)+1)} \in \mathcal{F}_h$ with $p_1 \in \mathcal{U}$, $p_2 \in \mathcal{M}_h$, and $0 < \epsilon < 1$. Let m denote base(p) (which clearly equals base(p_1)), and let $m + M - 1$ denote the span(p_1), where $M \geq 1$. Thus $|\mathrm{support}(p_1)| = M$.

Consider any probability distribution $u \in \mathcal{F}_h$, written as $u = (1-\epsilon')u_1 + \epsilon' u_2^{(\mathrm{span}(q_1)+1)}$, where $u_1 \in \mathcal{U}$, $u_2 \in \mathcal{M}_h$, and $0 < \epsilon' < 1$. Suppose that u is within ℓ_1 distance $\frac{(1-\epsilon)^2}{M(M+1)}$ from p. We show that

$$|\mathrm{span}(u_1)| \leq m + \left\lceil \frac{M}{1-\epsilon} \right\rceil.$$

To see this, suppose to the contrary that we have

$$|\mathrm{span}(u_1)| \geq m + \left\lceil \frac{M}{1-\epsilon} \right\rceil + 1.$$

If base(u_1)$\leq m$, all elements in the support of p_1 are assigned probability $\leq \frac{1}{\frac{M}{1-\epsilon}+1}$ from u. If base(u_1)$>$ m, then u(base(p_1))=0. Thus, in either case, we have $u(\mathrm{base}(p_1)) \leq \frac{1}{\frac{M}{1-\epsilon}+1}$.

We can now lower bound the ℓ_1 distance between p and u by

$$\frac{(1-\epsilon)}{M} - \frac{1}{\frac{M}{1-\epsilon}+1} = \frac{(1-\epsilon)^2}{M(M+1-\epsilon)} > \frac{(1-\epsilon)^2}{M(M+1)}.$$

This contradiction proves the claim.

Now, for fixed numbers m' and M', consider the collection $\mathcal{P}_{m',M'} \subseteq \mathcal{F}_h$ of all probability distributions with base m', and whose support of the initial uniform component is M'. Recall that $\mathcal{M}_h$ was shown to be strongly compressible in Example 24. Observe that the redundancy of $\mathcal{P}_{m',M'}$ will be at most the redundancy of $\mathcal{M}_h$ plus 1. Therefore we must also have that $\mathcal{P}_{m',M'}$ is strongly compressible.

The set of all probability distributions in the ℓ_1-neighborhood of $p \in \mathcal{F}_h$ with radius $\frac{(1-\epsilon)^2}{M(M+1)}$ can be decomposed into the finite union

$$\bigcup_{\substack{m',M' \\ m'+M' \leq \lceil m+\frac{M}{1-\epsilon} \rceil}} \mathcal{P}_{m',M'}.$$

Each component of the finite union is strongly compressible. Therefore it follows that this neighborhood of $p \in \mathcal{F}_h$ is strongly compressible. Thus no $p \in \mathcal{F}_h$ is deceptive and the collection is *d.w.c.*. □

We construct a *d.w.c.* collection $\mathcal{H}$ where one of the probability distributions in $\mathcal{H}$ has no non-zero neighborhood that is also strongly compressible.

We again partition $\mathbb{N}$ into $(T_i, i \geq 0)$ as before, where $T_i = \{2^i, \dots, 2^{i+1}-1\}$ for $i \geq 0$. Let $\mathcal{H}$ contain the probability distribution p_0 that assigns probability $\frac{1}{(i+1)(i+2)}$ to 2^i for all $i \geq 0$. We will construct $\mathcal{H}$ in such a way that while p_0 is not going to be deceptive in $\mathcal{H}$, no neighborhood of p_0 in $\mathcal{H}$ will be strongly compressible.

We construct $\mathcal{H}$ in several steps. We first fix a sequence $(\epsilon_m, m \geq 2)$ such that $0 < \epsilon_m < \frac{1}{2}$ and

$$\lim_{m\to\infty} \epsilon_m = 0.$$

Next, for $m \geq 2$, $k \geq m$, and $j \in \left\{2^k+1, , \dots, 2^k + 2^{\lceil k\epsilon_m \rceil}\right\}$, we define the probability distribution

$$p_{_{m,k,j}}(r) := \begin{cases} p_0(r), & \text{if } 1 \leq r \leq 2^{m-1}-1, \\ \frac{1}{m} - \frac{1}{k+1}, & \text{if } r = 2^{m-1}+1, \\ \frac{1}{k+1}, & \text{if } r = j, \\ 0, & \text{else.} \end{cases}$$

Now, for $m \geq 2$ and $k \geq m$, let

$$\mathcal{H}_{m,k} := \left\{ p_{_{m,k,j}} : 2^k + 1 \leq j \leq 2^k + 2^{\lceil k\epsilon_m \rceil} \right\},$$

let

$$\mathcal{H}_m := \cup_{k \geq m} \mathcal{H}_{m,k},$$

and, finally, let

$$\mathcal{H} := \{p_0\} \cup \left(\cup_{m\geq 2} \mathcal{H}_m\right).$$

A few observations about our construction. For all $m \geq 2$, all the probability distributions in $\mathcal{H}_m$ assign probabilities exactly as p_0 does to every element in $\cup_{i=0}^{m-2} T_i$, and the rest of their support is disjoint from that of p_0. It follows that, for all $m \geq 2$. for all $p \in \mathcal{H}_m$, we have

$$||p - p_0||_1 = \frac{2}{m}.$$

Hence, for all $m \geq 2$, the set of probability distributions in $\mathcal{H}$ within ℓ_1 distance $\leq \frac{2}{m}$ from p_0 is precisely $\{p_0\} \cup (\cup_{r \geq m} \mathcal{H}_r)$. Around any probability distribution in $\mathcal{H}$ other than p_0, there is a non-zero neighborhood containing no other probability distribution that belongs to $\mathcal{H}$. Therefore, none of the probability distributions in $\mathcal{H}$ other than p_0 can possibly be deceptive. Hence, to show that $\mathcal{H}$ is *d.w.c.*, we have to prove that p_0 is not deceptive.

Example 26. **(None of the neighborhoods of $p_0 \in \mathcal{H}$ is strongly compressible.)**
We show that for all $m \geq 2$ the collection of probability distributions $\mathcal{H}_m$ is not strongly compressible, *i.e.,* its asymptotic per-symbol redundancy is bounded away from zero.

To see this, for $2^k + 1 \leq j \leq 2^k + 2^{\lceil k\epsilon_m \rceil}$, let $S_j \subset \mathbb{N}^{k+1}$ be the set of all length-$(k+1)$ sequences all of whose symbols but one are from $\cup_{i=0}^{m-1} T_i$, and there is exactly one occurrence of the number j in the sequence. Clearly, for distinct j, S_j are disjoint. Observe that

$$p_{m,k,j}(S_j) = \left(1 - \frac{1}{k+1}\right)^k \geq \frac{1}{e}.$$

Therefore, from Lemma 14, we have that the length-$(k+1)$ redundancy of $\mathcal{H}_{m,k}$, which we denote by $R_{k+1}(\mathcal{H}_{m,k})$, satisfies

$$\frac{R_{k+1}(\mathcal{H}_{m,k})}{k+1} \geq \frac{1}{k+1}\left(\frac{\log|\mathcal{H}_{k,m}|}{e} - 1\right) = \frac{1}{k+1}\left(\frac{\lceil k\epsilon_m \rceil}{e} - 1\right).$$

Since for all $k \geq m \geq 2$ we have $\mathcal{H}_{m,k} \subset \mathcal{H}_m$, it follows that for $m \geq 2$ the length-n redundancy of $\mathcal{H}_m$, for $n \geq m+1$, which we denote by $R_n(\mathcal{H}_m)$, satisfies

$$\frac{R_n(\mathcal{H}_m)}{n} \geq \frac{R_n(\mathcal{H}_{m,n-1})}{n} \geq \frac{1}{n}\left(\frac{\lceil (n-1)\epsilon_m \rceil}{e} - 1\right).$$

Hence, the asymptotic per-symbol redundancy of $\mathcal{H}_m$ satisfies

$$\limsup_{n\to\infty} \frac{R_n(\mathcal{H}_m)}{n} \geq \frac{\epsilon_m}{e}. \tag{15}$$

Thus $\mathcal{H}_m$ is not strongly compressible and, in particular, neither is any ℓ_1 neighborhood of p_0.

Nevertheless, we can show that p_0 is not deceptive. We will verify that, as $m \to \infty$, the asymptotic per-symbol redundancy of an ℓ_1 neighborhood of radius $\frac{2(m+1)}{m^2}$ around p_0 goes to 0. [2]

To do so, observe from Proposition 35 that the asymptotic per-symbol redundancy of any collection of probability distributions on $\mathbb{N}$ is upper bounded by the single-letter redundancy of the collection. Recall that for $m \geq 2$ the ℓ_1 neighborhood of radius $\frac{2(m+1)}{m^2}$ around p_0 is the collection $\{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)$. We will verify that the single-letter redundancy of $\{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)$ diminishes to 0 as $m \to \infty$, which will then imply that p_0 is not deceptive, using Proposition 35.

[2]The choice of radius $\frac{2(m+1)}{m^2}$ is made since it satisfies $\frac{2}{m} < \frac{2(m+1)}{m^2} < \frac{2}{m-1}$ for $m \geq 2$, and we defined ℓ_1 neighborhoods to be open sets.

For $m \geq 2$, let q_m be the probability distribution on $\mathbb{N}$ defined by

$$q_m(r) := \begin{cases} p_0(r), & \text{if } 1 \leq r \leq 2^{m-1}-1, \\ \frac{1}{m} - \frac{1}{m+1}, & \text{if } r = 2^{m-1}+1, \\ \frac{1}{(k+1)(k+2)} \frac{1}{2^{\lceil k\epsilon_m \rceil}}, & \text{if } r \in \left\{2^k+1, , \ldots, 2^{\lceil k\epsilon_m \rceil}\right\}, k \geq m, \\ 0, & \text{else.} \end{cases}$$

Let $l \geq m \geq 2$. Then, for every $k \geq l$ and $j \in \left\{2^k+1, , \ldots, 2^k + 2^{\lceil k\epsilon_l \rceil}\right\}$, note that $p_{l,k,j} \in \mathcal{H}_{l,k}$ and q_l assign the same probabilities as those assigned by p_0 to every number $\leq 2^{l-1}-1$. It follows that

$$\begin{aligned} D(p_{l,k,j}||q_l) &= p_{l,k,j}(2^{l-1}+1) \log \frac{p_{l,k,j}(2^{l-1}+1)}{q_l(2^{l-1}+1)} + p_{l,k,j}(j) \log \frac{p_{l,k,j}(j)}{q_l(j)} \\ &\leq \frac{1}{l} \log(l+1) + \frac{1}{k+1} \log(k+2) + \frac{1}{k+1} \log 2^{\lceil k\epsilon_l \rceil} \\ &\leq \epsilon_l + \frac{2}{l} \log(l+1) + \frac{1}{l+1}. \end{aligned} \tag{16}$$

Now, for $m \geq 2$, consider the mixture probability distribution $\bar{q}_m$ on $\mathbb{N}$ given by

$$\bar{q}_m(r) := \sum_{l \geq m} \frac{m}{l(l+1)} q_l(r).$$

Fix $m \geq 2$. We have seen that any probability distribution in $\mathcal{H}$ in the ℓ_1 neighborhood of radius $\frac{2(m+1)}{m^2}$ around p_0 must belong to $\{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)$. For every $k \geq l \geq m$, and $j \in \left\{2^k+1, , \ldots, 2^k + 2^{\lceil k\epsilon_l \rceil}\right\}$, we observe that $p_{l,k,j} \in \mathcal{H}_{l,k}$ and $\bar{q}_m$ assign the same probabilities as those assigned by p_0 to every number $\leq 2^{m-1}-1$. Also, p_0 and $\bar{q}_m$ assign the same probabilities as those assigned by p_0 to every number $\leq 2^{m-1}-1$. We will now use this observation to find upper bounds for $D(p_{m,k,j}||\bar{q}_m)$ for $k \geq m$ and $j \in \left\{2^k+1, , \ldots, 2^k + 2^{\lceil k\epsilon_m \rceil}\right\}$, then for $D(p_{l,k,j}||\bar{q}_m)$ for $k \geq l \geq m+1$ and $j \in \left\{2^k+1, , \ldots, 2^k + 2^{\lceil k\epsilon_l \rceil}\right\}$, and finally for $D(p_0||\bar{q}_m)$.

For $k \geq m$ and $j \in \left\{2^k+1, , \ldots, 2^k + 2^{\lceil k\epsilon_m \rceil}\right\}$, we write

$$\begin{aligned} D(p_{m,k,j}||\bar{q}_m) &= p_{m,k,j}(2^{m-1}+1) \log \frac{p_{m,k,j}(2^{m-1}+1)}{\bar{q}_m(2^{m-1}+1)} + p_{m,k,j}(j) \log \frac{p_{m,k,j}(j)}{q_m(j)} \\ &\leq p_{m,k,j}(2^{m-1}+1) \log \frac{(m+1)p_{m,k,j}(2^{m-1}+1)}{q_m(2^{m-1}+1)} + p_{m,k,j}(j) \log \frac{(m+1)p_{m,k,j}(j)}{q_m(j)} \\ &\leq \epsilon_m + \frac{4}{m} \log(m+1) + \frac{1}{m+1}, \end{aligned} \tag{17}$$

where the last step uses (16) for the choice $l = m$.

For $k \geq l \geq m+1$ and $j \in \{2^k+1, , \ldots, 2^k+2^{\lceil k\epsilon_l \rceil}\}$, we write

$$\begin{aligned}
D(p_{l,k,j}||\bar{q}_m) &= \sum_{n=m-1}^{l-2} p_{l,k,j}(2^n) \log \frac{p_{l,k,j}(2^n)}{\bar{q}_m(2^n)} + \sum_{r=2^{l-1}}^{\infty} p_{l,k,j}(r) \log \frac{p_{l,k,j}(r)}{\bar{q}_m(r)} \\
&\leq \sum_{n=m-1}^{l-2} p_{l,k,j}(2^n) \log \frac{p_{l,k,j}(2^n)}{\frac{mq_{n+2}(2^n)}{(n+2)(n+3)}} + \sum_{r=2^{l-1}}^{\infty} p_{l,k,j}(r) \log \frac{p_{l,k,j}(n) l(l+1)}{mq_l(n)} \\
&\leq \sum_{n=m-1}^{l-2} p_{l,k,j}(2^n) \log \frac{p_{l,k,j}(2^n)}{\frac{q_{n+2}(2^n)}{(n+2)(n+3)}} + \sum_{r=2^{l-1}}^{\infty} p_{l,k,j}(r) \log \frac{p_{l,k,j}(r)}{q_l(r)} + \frac{\log(\frac{l(l+1)}{m})}{l} \\
&= \sum_{n=m-1}^{l-2} \frac{\log\left((n+2)(n+3)\right)}{(n+1)(n+2)} + \sum_{r=2^{l-1}}^{\infty} p_{l,k,j}(r) \log \frac{p_{l,k,j}(r)}{q_l(r)} + \frac{\log(\frac{l(l+1)}{m})}{l} \\
&\overset{(a)}{\leq} \sum_{n=m-1}^{l-2} \frac{\log\left((n+2)(n+3)\right)}{(n+1)(n+2)} + \epsilon_l + 4\frac{\log(l+1)}{l} + \frac{1}{l+1} \\
&\leq \sum_{n=m-1}^{\infty} \frac{\log\left((n+2)(n+3)\right)}{(n+1)(n+2)} + \epsilon_m + \frac{4\log(m+1)}{m} + \frac{1}{m+1}, \qquad (18)
\end{aligned}$$

where (a) uses the bound $\log(l(l+1)/m) \leq 2\log(l+1)$, observes that $q_{n+2}(2^n) = p_0(2^n) = p_{l,k,j}(2^n)$, and uses (16).

To bound $D(p_0||\bar{q}_m)$ from above, note that $\bar{q}_m(2^n) = \frac{m}{n+2} p_0(2^n)$ for $n \geq m-1$. Therefore we have

$$\begin{aligned}
D(p_0||\bar{q}_m) &= \sum_{n=m-1}^{\infty} p_0(2^n) \log \frac{p_0(2^n)}{\bar{q}_m(2^n)} \\
&\leq \sum_{n=m-1}^{\infty} \frac{\log(n+1)}{(n+1)(n+2)}. \qquad (19)
\end{aligned}$$

From (17), (18), and (19), the single letter redundancy of all sources around p_0 within ℓ_1 distance $\frac{2(m+1)}{m^2}$ of p_0 satisfies the upper bound

$$\sup_{p \in \{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)} D(p||\bar{q}_m) \leq \sum_{n=m-1}^{\infty} \frac{\log\left((n+2)(n+3)\right)}{(n+1)(n+2)} + \epsilon_m + \frac{4\log(m+2)}{m+1} + \frac{1}{m+1}. \qquad (20)$$

Note that

$$\sum_{n=1}^{\infty} \frac{\log\left((n+2)(n+3)\right)}{(n+1)(n+2)} < \infty.$$

Hence, as $m \to \infty$, each of the terms on the right side of (20) converges to 0. Since the single letter redundancy of $\{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)$ diminishes to 0 as $m \to \infty$, from Proposition 35, the asymptotic per-symbol redundancy of $\{p_0\} \cup (\cup_{l \geq m} \mathcal{H}_l)$ also diminishes to zero as $m \to \infty$. Therefore p_0 is not deceptive.

In conclusion, none of the neighborhoods of p_0 is strongly compressible, from (15), since the asymptotic per-symbol redundancy of a $\frac{2(m+1)}{m^2}$ size ℓ_1 neighborhood of p_0 is lower bounded by $\epsilon_m/e > 0$. Yet, as we showed above, p_0 is not deceptive. As noted above, no other probability distribution in $\mathcal{H}$ can possibly be deceptive since it has a neighborhood of nonzero radius around it containing no other probability distribution from $\mathcal{H}$. Therefore, $\mathcal{H}$ is *d.w.c.*.

$\square$

*V-C. Non-*d.w.c. *collections*

We now construct two examples of non-*d.w.c.* model classes to illustrate some additional points.

In Example 27 we define a model class $\mathcal{B}$ where exactly one source in the model class is deceptive. This would mean that $\mathcal{B}$ is not *d.w.c.*. However, even though $\mathcal{B}$ is not *d.w.c.*, removing the single deceptive source renders the rest of the model class *d.w.c.*. Put another way, adding a single source to a *d.w.c.* model class may make the resulting bigger model class not *d.w.c.*. Since a model class with one source is trivially *d.w.c.*, it follows that even finite unions of *d.w.c.* classes may not be *d.w.c.*.

The second example we give here is of an insurable model class $\mathcal{I}$ that is not *d.w.c.*. See Example 8 for the definition of insurability of a model class.

Partition $\mathbb{N}$ into $(T_i, i \ge 0)$, where $T_i := \{2^i, \dots, 2^{i+1}-1\}$, $i \ge 0$. For $0 < \epsilon < 1$, let $n_\epsilon = \lceil \frac{1}{\epsilon} \rceil$. Note that ϵ lies in the range $[\frac{1}{n_\epsilon}, \frac{1}{n_\epsilon - 1})$. For $1 \le j \le 2^{n_\epsilon}$, let $p_{\epsilon,j}$ be the probability distribution on $\mathbb{N}$ that assigns probability $1-\epsilon$ to the natural number 1 (or equivalently, to the set T_0), and ϵ to the natural number $2^{n_\epsilon} + j - 1$. Finally, let p_0 be a singleton probability distribution assigning probability 1 to the natural number 1.

Now, let $\mathcal{B}$ (mnemonic for binary, since every probability distribution in $\mathcal{B}$ has support of cardinality at most 2) be the collection of probability distributions on $\mathbb{N}$ defined by

$$\mathcal{B} := \{p_{\epsilon,j} : 0 < \epsilon < 1,\ 1 \le j \le 2^{n_\epsilon}\} \cup \{p_o\}.$$

As usual, $\mathcal{B}^\infty$ denotes the set of *i.i.d.* probability measures on $\mathbb{N}^\infty$ corresponding to $\mathcal{B}$.

Example 27. **(p_0 is the unique probability distribution in $\mathcal{B}$ that is deceptive.)**

An ℓ_1 neighborhood of radius δ around p_0 is comprised of p_0 and the $p_{\epsilon,j}$ for all $0 < \epsilon < \delta/2$, and all $1 \le j \le 2^{n_\epsilon}$. For all $n \ge 1$ and $j \in \mathcal{T}_n$, let $S_{n,j}$ denote the set of all length n strings of natural numbers with exactly one appearance of j and the remaining $n-1$ elements of the string being 1. Then, we have

$$p_{\frac{1}{n},j}(S_{n,j}) = \left(1 - \frac{1}{n}\right)^{n-1} \ge \frac{1}{e}.$$

For each $n \ge 1$, the sets $S_{n,j}$ are disjoint as j ranges over $\mathcal{T}_n$. Further, they are subsets of $\mathbb{N}^n$. Therefore, Lemma 14 implies that the length-n redundancy of the collection $\{p_{\frac{1}{n},j} : j \in \mathcal{T}_n\}$ is lower bounded by

$$\frac{n}{e} - 1.$$

Therefore, for all $n > \frac{2}{\delta}$, the length-n redundancy of the ℓ_1 neighborhood of radius δ is bounded below by $\frac{n}{e} - 1$. This implies that the asymptotic per-symbol redundancy of the ℓ_1 neighborhood of size δ is bounded below by $\frac{1}{e}$. From the second part of Lemma 18, we conclude that p_0 is deceptive.

On the other hand, for $0 < \epsilon < 1$, around every other probability distribution $p_{\epsilon,j} \in \mathcal{B}$, there is an ℓ_1-neighborhood of radius $\frac{1}{n_\epsilon}$ that contains only probability distributions in $\mathcal{B}$ that have support equal to $\{1, 2^{n_\epsilon} + j - 1\}$. For $n \ge 1$, let $\hat{r}_n$ denote the probability measure on $\mathbb{N}^n$ giving probability $\frac{1}{(n+1)\binom{n}{k}}$ to each of the strings in $\mathbb{N}^n$ comprised of k occurrences of $2^{n_\epsilon} + j - 1$ and $n-k$ occurrences of 1, $0 \le k \le n$.

Let r_n be the probability measure corresponding to $\hat{r}_n$, as in Lemma 29. Then, for all $p \in \mathcal{B}$ in this ℓ_1-neighborhood of $p_{\epsilon,j} \in \mathcal{B}$, we have for all n

$$D_n(p||r_n) \leq \log(n+1).$$

Noting that the measure r on $\mathbb{N}^\infty$ that assigns probability

$$r(\mathbf{x}) = \sum_{m\geq 1} \frac{r_m(\mathbf{x})}{m(m+1)}$$

satisfies

$$\limsup_{n\to\infty} \sup_{p:|p-p_{\epsilon,j}|<\frac{1}{n_\epsilon}} \frac{1}{n} D_n(p||q) \leq \lim_{n\to\infty} \frac{\log n}{n} = 0,$$

we conclude that for every $p_{\epsilon,j} \in \mathcal{B}$ there is an ℓ_1-neighborhood of $p_{\epsilon,j}$ that has zero asymptotic per-symbol redundancy. Hence, by Lemma 37, there is a neighborhood of $p_{\epsilon,j}$ that has zero asymptotic per-symbol redundancy. We conclude that, while p_0 is deceptive, no other probability distribution in $\mathcal{B}$ is deceptive.

Indeed, this is quite intuitive when we think about what is involved operationally in compressing strings of integers whose statistics are *i.i.d.* and governed by a probability distribution in $\mathcal{B}$. If at any point we see two distinct symbols in such a string, there is no ambiguity about what the underlying distribution is from that point on, and very little ambiguity in the probabilities of the two distinct symbols seen, of which one must be the symbol 1. But if we see a string of all 1s we can never be sure (no matter what the length of the string) what the underlying source is. One possibility is that the source is p_0.

But having seen a string of 1s of length m, there is also a reasonable chance that the underlying source could be $p_{\epsilon,j}$ for some $\epsilon \ll \frac{1}{m}$ and any $j \in T_{n_\epsilon}$. There are 2^{n_ϵ} such possible values j can take in T_{n_ϵ}, so any description of j requires an additional n_ϵ bits or $\gg m$ bits.

However, if we remove p_0 from the collection, we have no such trouble. We have no obligation to stop on any finite length string of all 1s, no matter how long it is, since the sequence of all 1s has probability 0 under every source in $\mathcal{B}$ other than p_0. □

The last example is a collection $\mathcal{I}$ of probability measures over $\mathbb{N}$ that is insurable but not *d.w.c.*. In fact $\mathcal{I}$ is not even weakly compressible.

Partition $\mathbb{N}$ into the sets $(T_i, i \geq 0)$ as before, where $T_i := \{2^i, \ldots, 2^{i+1}-1\}$. For each $i \geq 1$, pick exactly one element of T_i and assign it probability $1/(i(i+1))$. We define $\mathcal{I}$ to be the collection of all probability distributions on $\mathbb{N}$ that can be formed in this way. $\mathcal{I}^\infty$ denotes the set of *i.i.d.* probability measures on $\mathbb{N}^\infty$ corresponding to $\mathcal{I}$.

Example 28. **($\mathcal{I}$ is insurable but not weakly compressible, hence not *d.w.c.*)**

For all $p \in \mathcal{I}$ and all $k \geq 1$, we have

$$\sum_{n\geq 2^k} p(n) = \frac{1}{k}.$$

This means that the entire set $\mathcal{I}$ is tight. By [2, Theorem 1], we can therefore conclude that $\mathcal{I}$ is insurable.

On the other hand, for every probability distribution q on $\mathbb{N}$, for all $i \geq 1$ there is $x_i \in T_i$ such that

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

By the definition of $\mathcal{I}$, there is a probability distribution $p \in \mathcal{I}$ that has support $\{x_i : i \geq 1\}$. Note that $D(p||q) = \infty$. Since every probability distribution in $\mathcal{I}$ has finite entropy (in fact they all have the same entropy), from Lemma 10 we conclude that $\mathcal{I}$ is not weakly compressible. In particular, $\mathcal{I}$ is not *d.w.c.*. □

VI. Necessity part of Theorem 17

In this section we prove the necessity part of Theorem 17. Namely, we prove that the existence of deceptive distributions kills *d.w.c.*. More precisely, we prove that if $\mathcal{P}$ is a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the associated collection of *i.i.d.* probability measures on $\mathbb{N}^\infty$, then $\mathcal{P}^\infty$ is *d.w.c.* only if no $p \in \mathcal{P}^\infty$ is deceptive.

To prove this, suppose $p \in \mathcal{P}$ is deceptive. Then, by the second part of Lemma 18, for every probability measure q on $\mathbb{N}^\infty$ we can find $\delta > 0$ such that

$$\lim_{\epsilon' \to 0} \limsup_{n \to \infty} \sup_{p' \in B(p,\epsilon';\mathcal{P})} \frac{1}{n} D_n(p'||q) > \delta.$$

Pick any $0 < \eta < 1$, and let τ be a stopping rule. We will demonstrate that there is some $\tilde{p} \in \mathcal{P}$ such that

$$\tilde{p}(\tau \text{ is } \delta\text{−premature with respect to } q \text{ for } \tilde{p}) > \eta,$$

where we refer to the discussion around (5) to recall what it means for a stopping time to be $\delta-$premature for the probability distribution $\tilde{p} \in \mathcal{P}$, with respect to the probability measure q on $\mathbb{N}^\infty$.

In order to do this, for all $n \geq 1$ let

$$A_n := \{x^n \in \mathbb{N}^n : \tau(x^n) = 1\}$$

denote the set of sequences of length n on which τ has entered. Note that $p(A_n)$ is increasing with n and $\lim_{n\to\infty} p(A_n) = 1$. We can therefore pick $n \geq 4/(1-\eta)$ large enough such that $p(A_n) \geq (1+\eta)/2$.

Let [3] $\epsilon := \frac{\log e}{16n^8}$. Applying Lemma 39 in Appendix II to *i.i.d.* probability measures over length-n strings, we see that for all $\tilde{p} \in \mathcal{P}$ such that $\mathcal{J}(p, \tilde{p}) \leq \epsilon$, we have

$$\tilde{p}(A_n) > (1+\eta)/2 - \frac{2}{n} \geq \eta.$$

Since $\limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon';\mathcal{P})} \frac{1}{n} D_n(p'||q)$ is nondecreasing in ϵ', we can choose $\tilde{p} \in B(p, \epsilon; \mathcal{P})$ such that for some $n \geq 1$ we have

$$\tilde{p}(A_n) > \eta \text{ and } \frac{1}{n} D_n(\tilde{p}||q) > \delta.$$

This in turn means, for the choice of η and δ above, that $\tilde{p}(\tau$ is $\delta-$premature with respect to q for $\tilde{p}) > \eta$. This completes the proof of the necessity part of Theorem 17.

As a caveat regarding the structure of this proof, we remark that the presence of a deceptive distribution $p \in \mathcal{P}$ does not automatically imply that any other probability distribution in any neighborhood of the deceptive distribution p is also deceptive. For example, the class $\mathcal{B}$ in Example 27 has only p_0 deceptive, while no other distribution in its neighborhood is.

[3]Please note that in the interest of simplicity, we have not attempted to provide the best scaling for ϵ or the tightest possible bounds.

VII. Sufficiency part of Theorem 17

In this section we prove the sufficiency part of Theorem 17. Namely, we prove that if a collection $\mathcal{P}$ of probability distributions on $\mathbb{N}$ does not contain any deceptive distributions, then $\mathcal{P}$ is *d.w.c.*. We do this by explicitly constructing a probability measure q^* on $\mathbb{N}^\infty$ such that, given any desired confidence probability $0 < 1-\eta < 1$ and accuracy $\delta > 0$, there is a stopping rule τ such that, for every $p \in \mathcal{P}$, under p, τ is $\delta-$premature with respect to q^* for p, as defined in (5), with probability at most η. Note that it suffices to prove this for all δ of the form $\frac{1}{m}$ for $m \geq 1$, so will restrict attention to this case, and denote the corresponding stopping rule we construct by $\tau_{\eta,m}$.

Suppose $p \in \mathcal{P}$ is not deceptive. From Lemma 18, there is a probability measure q_p on $\mathbb{N}^\infty$ such that for all $m \geq 1$ we can pick $\epsilon_{p,m} > 0$ satisfying

$$\limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon_{p,m};\mathcal{P})} \frac{1}{n} D_n(p'||q_p) < \frac{1}{m}. \tag{21}$$

We fix such an $\epsilon_{p,m} > 0$ for each $p \in \mathcal{P}$ and $m \geq 1$, satisfying the additional technical requirement that $\epsilon_{p,m} < 16 \log e$.

For $\delta \geq 1$, let $m = 1$ and for $0 < \delta < 1$ let $m = \lceil 1/\delta \rceil$. Therefore m is the natural number such that $\frac{1}{m} \leq \delta < \frac{1}{m-1}$. For any $\delta > 0$, we call $\epsilon_{p,\lceil 1/\delta \rceil}$ the $\delta-$*reach* of p. In particular, $\epsilon_{p,m} > 0$ is the $\frac{1}{m}-$reach of p.

The intuitive meaning of the reach $\epsilon_{p,\delta}$ of a probability distribution $p \in \mathcal{P}$ is that, even if the statistics of the observations are being determined by some probability distribution in $\mathcal{P}$ within the reach of p that is not necessarily p, we have control, by waiting long enough, over the amount of harm, as determined by δ, that will be done if we decide instead that the statistics of the observations are being determined by p. This rough heuristic will be made more precise in what follows. Note that, for any $m \geq 1$, we do not require any regularity over $p \in \mathcal{P}$ of $\epsilon_{p,m}$. The reason this does not matter will also soon become apparent, and is basically because, for each $m \geq 1$, it will suffice to focus on only a countable collection of $p \in \mathcal{P}$.

Given $m \geq 1$, the *zone* $Q_{p,m}$ of a probability distribution $p \in \mathcal{P}$ is defined to be the set of probability distributions u on $\mathbb{N}$ given by

$$Q_{p,m} \stackrel{\text{def}}{=} \left\{ u : |p-u|_1 < \frac{{\epsilon_{p,m}}^2 (\ln 2)^2}{16} \right\}, \tag{22}$$

where, $\epsilon_{p,m}$ is the $\frac{1}{m}$-reach of p. Note that the probability distributions in $Q_{p,m}$ are not necessarily in $\mathcal{P}$.

Note that for all $p \in \mathcal{P}$, because we have assumed that $\epsilon_{p,m} < 16 \log e$, Lemma 37 in Appendix II implies that the zone $Q_{p,m}$ satisfies $Q_{p,m} \cap \mathcal{P} \subseteq B(p, \epsilon_{p,m}; \mathcal{P})$. Trivially $p \in Q_{p,m} \cap \mathcal{P}$. Therefore we have we have

$$\mathcal{P} = \cup_{p \in \mathcal{P}} (Q_{p,m} \cap \mathcal{P}).$$

Further, since $Q_{p,m}$ is open in the ℓ_1 topology, each of the intersections $Q_{p,m} \cap \mathcal{P}$ is relatively open in the ℓ_1 topology on $\mathcal{P}$. Since $\mathcal{P}$ is Lindelöf under the ℓ_1 topology (see [2, Sec. 6.1] for a proof), there is a countable set $\tilde{\mathcal{P}}_m \subseteq \mathcal{P}$, such that $\mathcal{P}$ is covered by the collection of relatively open sets $(Q_{\tilde{p},m} \cap \mathcal{P}, \tilde{p} \in \tilde{\mathcal{P}}_m)$, i.e. we have

$$\mathcal{P} = \cup_{\tilde{p} \in \tilde{\mathcal{P}}_m} (Q_{\tilde{p},m} \cap \mathcal{P}). \tag{23}$$

For any fixed $m \geq 1$, we will make a choice of such a $\tilde{\mathcal{P}}_m$ and refer to it as the *quantization* of $\mathcal{P}$ and to elements of $\tilde{\mathcal{P}}_m$ as the *centroids* of the quantization, borrowing from commonly used literature in classification. We index the countable set of centroids, $\tilde{\mathcal{P}}_m$ by $\iota_m : \tilde{\mathcal{P}}_m \to \mathbb{N}$.

We now construct a probability measure q^* on $\mathbb{N}^\infty$ and, for each $0 < \eta < 1$ and $m \geq 1$, a stopping rule $\tau_{\eta,m}$, such that the pair q^* and $\tau_{\eta,m}$ will together satisfy the required guarantee that for every $p \in \mathcal{P}$, the probability that the stopping time $\tau_{\eta,m}$ is $\frac{1}{m}-$premature with respect to q^* for p is at most η.

a) Construction of the probability measure q^* on $\mathbb{N}^\infty$: For each $\tilde{p} \in \tilde{\mathcal{P}}_m$ there is a probability measure $q_{\tilde{p}}$ on $\mathbb{N}^\infty$ satisfying (21) for $\tilde{p}$, with $\epsilon_{\tilde{p},m}$ denoting the $\frac{1}{m}-$reach of $\tilde{p}$. Let

$$\tilde{Q}_m := \{q_{\tilde{p}} : \tilde{p} \in \tilde{\mathcal{P}}_m\}$$

denote the collection of these probability measures as $\tilde{p}$ ranges over $\tilde{\mathcal{P}}_m$. Note that $\tilde{Q}_m$ is countable and is a collection of not necessarily *i.i.d.* probability measures on $\mathbb{N}^\infty$. For $\tilde{q} \in \tilde{Q}_m$, set the index $\iota_m(\tilde{q})$ to be equal the index assigned to the corresponding centroid $\tilde{p}$ in the enumeration of $\tilde{\mathcal{P}}_m$. Then define a probability measure q_m on $\mathbb{N}^\infty$ by setting, for each $n \geq 1$ and each $\mathbf{x} \in \mathbb{N}^n$, the probability

$$q_m(\mathbf{x}) := \sum_{\tilde{q} \in \tilde{Q}_m} \frac{\tilde{q}(\mathbf{x})}{\iota_m(\tilde{q})(\iota_m(\tilde{q})+1)}.$$

Finally, let q^* be the probability measure on $\mathbb{N}^\infty$ defined by letting

$$q^*(\mathbf{x}) := \sum_{m \geq 1} \frac{q_m(\mathbf{x})}{m(m+1)},$$

for each $n \geq 1$ and $\mathbf{x} \in \mathbb{N}^*$,

Now, for all $\tilde{p} \in \tilde{\mathcal{P}}_m$, we have

$$\begin{aligned}
\limsup_{n\to\infty} \sup_{p' \in B(\tilde{p}, \epsilon_{\tilde{p},\frac{1}{m}}; \mathcal{P})} \frac{1}{n} D_n(p' || q^*) &= \limsup_{n\to\infty} \sup_{p' \in B(\tilde{p}, \epsilon_{\tilde{p},\frac{1}{m}}; \mathcal{P})} \frac{1}{n} D_n(p' || q_m) \\
&= \limsup_{n\to\infty} \sup_{p' \in B(\tilde{p}, \epsilon_{\tilde{p},\frac{1}{m}}; \mathcal{P})} \frac{1}{n} D_n(p' || q_{\tilde{p}}) \\
&< \frac{1}{m}. \qquad (24)
\end{aligned}$$

We turn next to construct a stopping rule $\tau_{\eta,m}$ having the property that, for all $p \in \mathcal{P}$, we have

$$p\big(\tau_{\eta,m} \text{ is } \tfrac{1}{m}-\text{premature with respect to } q^* \text{ for } p\big) < \eta.$$

b) Description of the stopping rule $\tau_{\eta,m}$: Fix $0 < \eta < 1$ and $m \geq 1$. Let $p \in \mathcal{P}$ be the probability distribution in force, which is unknown. The idea is that we want sequences generated by the (unknown) $p \in \mathcal{P}$ to be captured by one of the centroids of the quantization $\tilde{\mathcal{P}}_m$ that have p in their $\frac{1}{m}-$reach.

Consider a length-n sequence x^n on which we have not yet decided the value of $\tau_{\eta,m}(x^m)$ for any $1 \leq m \leq n$. Let x^n have type (*i.e.,* empirical distribution) t, which we now insist on thinking of as a sequence of unnormalized fractions on $\mathbb{N}$, in order to ensure that t determines the length of the sequence x^n that defines it. The set of centroids in $\tilde{\mathcal{P}}_m$ that can potentially *capture* t is defined to be

$$\tilde{\mathcal{P}}_{m,t} := \{\tilde{p} \in \tilde{\mathcal{P}}_m : t \in Q_{\tilde{p},m}\}.$$

Since $\cup_{\tilde{p}\in\tilde{\mathcal{P}}_m}(Q_{\tilde{p},m}\cap\mathcal{P})$ is an open set containing $\mathcal{P}$, the probability under p of the set of all sequences of length n whose type is captured by some centroid in $\tilde{\mathcal{P}}_m$ approaches 1 as $n\to\infty$.

Not every centroid in $\tilde{\mathcal{P}}_{m,t}$ is necessarily benign, since some of these centroids may not have the generating probability measure p within their $\frac{1}{m}$-reach. Given that the number of centroids is countably infinite, there is no easy union bound based approach that could be invoked to resolve the issue. Therefore, when $\tilde{\mathcal{P}}_{m,t}\neq\emptyset$, we refine $\tilde{\mathcal{P}}_{m,t}$ further to $\hat{\mathcal{P}}_{m,t}\subset\tilde{\mathcal{P}}_{m,t}$ in a way that will allow us to use Lemma 40 to bound the probability of wrong capture.

To do so, for every $\tilde{p}\in\tilde{\mathcal{P}}_m$, with $\frac{1}{m}-$reach $\epsilon_{\tilde{p},m}$, let

$$D_{\tilde{p},m} := \frac{{\epsilon_{\tilde{p},m}}^4(\ln 2)^4}{256}\ .$$

The quantity above plays the role of γ when using Lemma 40.

To understand the core of our sufficiency proof, consider what happens when the underlying p happens to be outside the $\frac{1}{m}-$reach of $p'\in\tilde{\mathcal{P}}_{m,t}$. Since p is far from p' (out of its $\frac{1}{m}-$reach), but p' is close to the empirical distribution, t, of the observed sequence, our pseudo-triangle inequality from Lemma 37 will use the quantity $D_{p',m}$ to lower bound the distance of t from the underlying p, which allows us to conclude that sequences with type t have a small probability under p.

The centroids in $\tilde{\mathcal{P}}_{m,t}$ that get placed into $\hat{\mathcal{P}}_{m,t}$ are those that satisfy (26) and (27) below. In what follows, the quantity $\log C(p',m)$ of a centroid $p'\in\tilde{\mathcal{P}}_{m,t}$ plays the role of the "effective size" of the support size of p', corresponding to the number k of Lemma 40. Given $\tilde{p}\in\tilde{\mathcal{P}}_m$, we define $C(\tilde{p},m)$ via

$$C(\tilde{p},m) := 2^{3\left(\sup_{r\in B(\tilde{p},\epsilon_{\tilde{p},m};\mathcal{P})}\dot{F}_r^{-1}(1-\sqrt{D_{\tilde{p},m}}/6)\right)}, \tag{25}$$

and we note that $C(\tilde{p},m)$ is finite from the tightness result in Lemma 12. This is because we have

$$\limsup_{n\to\infty}\sup_{r\in B(\tilde{p},\epsilon_{\tilde{p},m};\mathcal{P})}\frac{1}{n}D_n(r||q^*)<\frac{1}{m},$$

from (24), which implies that for sufficiently large n the single letter redundancy of the family of n-fold product measures on $\mathbb{N}^n$ corresponding to the probability distributions in $B(\tilde{p},\epsilon_{\tilde{p},m};\mathcal{P})$ is finite, which, by Lemma 12, implies that this family of n-fold product measures on $\mathbb{N}^n$ is tight, which implies that the family of product distributions $B(\tilde{p},\epsilon_{\tilde{p},m};\mathcal{P})$ is tight.

With $C(p',m)$ for $p'\in\tilde{\mathcal{P}}_{m,t}$ defined as in (25), the conditions we require on $p'\in\tilde{\mathcal{P}}_{m,t}$ in order to place it in $\hat{\mathcal{P}}_{m,t}$ are

$$\exp\left(-nD_{p',m}/18\right)\leq\frac{\eta}{2C(p',m)\iota(p')^2n(n+1)}, \tag{26}$$

and

$$2\dot{F}_t^{-1}(1-\sqrt{D_{p',m}}/6)\leq\log C(p',m). \tag{27}$$

Note that given $\tilde{p}\in\tilde{\mathcal{P}}_m$ and a type t (which we recall determines the length n of the sequence defining it), one could ask if the conditions analogous to (26) and 27 hold or not for the pair $(\tilde{p},t)$; this observation will become important in Appendix V. It is also worth remarking that the proof of sufficiency of the necessary and sufficient condition for the insurability of a model class in [2, Thm. 1] also uses a similar criterion to bound the probability of wrong capture.

We are now in a position to specify the stopping rule $\tau_{\eta,m}$. Consider a sequence of natural numbers, x^n, having type t, which we recall determines the length n of the sequence, and assume that we have not yet specified $\tau_{\eta,m}$ for any prefix x^l of the sequence x^n for $1 \leq l \leq n$.

If $\hat{\mathcal{P}}_{m,t} = \emptyset$ there is no way to assign any element of $\hat{\mathcal{P}}_{m,t}$ to this sequence and its suffixes and so we move on to all the possible single letter extensions of the sequence x^n, without for the moment deciding what $\tau_{\eta,m}(x^n)$ is, although we it will eventually turn out to be 0.

If $\hat{\mathcal{P}}_{m,t} \neq \emptyset$, let $\hat{p}$ denote the probability distribution in $\hat{\mathcal{P}}_{m,t}$ with the smallest index. All suffixes of x^n are then said to be *trapped* by $\hat{p}$, which means that they are assigned to $\hat{p} \in \hat{\mathcal{P}}_{m,t}$. From (24), we have

$$\limsup_{n\to\infty} \sup_{r\in B(\hat{p},\epsilon_{\hat{p},m};\mathcal{P})} \frac{1}{n} D_n(r||q^*) < \frac{1}{m}.$$

This means that the set

$$N_{\hat{p}} := \{n : \sup_{r\in B(\hat{p},\epsilon_{\hat{p},m};\mathcal{P})} \frac{1}{n} D_n(r||q^*) \geq \frac{1}{m}\} \tag{28}$$

is finite. For any suffix x^N of x^n, when $N > \max N_{p^*}$, we set $\tau_{\eta,m}(x^N) = 1$, 0 else.

Finally for each finite string x^n for which the value of $\tau_{\eta,m}(x^n)$ has not yet been decided, we set this value to be 0. It can be checked that $\tau_{\eta,m}$ so defined is a stopping time. This is because if $\tau_{\eta,m}(x^n) = 0$ for any sequence $x^n \in \mathbb{N}^n$, then we also have $\tau_{\eta,m}(x^m) = 0$ for $1 \leq m \leq n$, i.e. for all its prefixes.

c) $\tau_{\eta,m}$ enters with probability 1: This is proved in Appendix V, using an argument similar to that used in the sufficiency proof in [2].

d) Probability under any $p \in \mathcal{P}$ that $\tau_{\eta,m}$ is $\frac{1}{m}-$premature with respect to $q*$ for p is strictly less than η: Consider any $p \in \mathcal{P}$. Among sequences of natural numbers on which $\tau_{\eta,m}$ has entered, we will distinguish between those that are in *good* traps and those in *bad* traps. If a sequence x^n is trapped by $\hat{p} \in \tilde{\mathcal{P}}_m$ such that $p \in B(\hat{p},\epsilon_{\hat{p},m};\mathcal{P})$, we call $\hat{p}$ is a good trap for that sequence. Conversely, if $p \notin B(\hat{p},\epsilon_{\hat{p},m};\mathcal{P})$, $\hat{p}$ is called a bad trap for that sequence.

(Good traps) Suppose a length-n sequence x^n is in a good trap. Namely, it is trapped by a probability distribution $\hat{p} \in \tilde{\mathcal{P}}_m$ such that $p \in B(\hat{p},\epsilon_{\hat{p},m};\mathcal{P})$. Then, if $\tau_{\eta,m}(x^n) = 1$ it must be the case that $\frac{1}{n}D(p||q^*) < \frac{1}{m}$. Thus such sequences cannot contribute to the probability under p of $\tau_{\eta,m}$ being $\frac{1}{m}-$premature with respect to q^* for p.

(Bad traps) We can show that the probability with which sequences generated by p fall into bad traps is strictly less than η using an argument, which is essentially identical to the one used in [2], based on the pseudo-triangle inequality from Lemma 37. This argument is reproduced in Appendix VI for the sake of completeness. Pessimistically, we assume that $\tau_{\eta,m}$ is $\frac{1}{m}-$premature with respect to q^* for p on every sequence that falls into a bad trap.

This completes the proof of the sufficiency part of Theorem 17.

Acknowledgments

This work was in part supported by the NSF Science & Technology Center for Science of Information Grant number CCF-0939370. In addition, Santhanam was also supported by NSF Grants CCF-1065632 and CCF-1619452; Anantharam was also supported by the ARO MURI grant W911NF- 08-1-0233, "Tools for the Analysis and Design of Complex Multi-Scale Networks", Marvell Semiconductor Inc., the U.C. Discovery program, the William and Flora Hewlett Foundation supported Center for Long Term Cybersecurity at Berkeley, and the NSF grants CNS-0910702, ECCS-1343998, CNS–1527846, CCF–1618145 and CCF-1901004; Szpankowski was also supported by NSF Grants CCF-1524312, CCF-2006440, and CCF-2007238.

References

[1] L.D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783—795, November 1973.

[2] N. Santhanam and V. Anantharam. Agnostic insurance of model classes. *Journal of Machine Learning Research*, pages 2329–2355, 2015. Full version available from arXiv doc id: 1212:3866.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

[5] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[6] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[7] Y.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3—17, 1987.

[8] B. Fittingoff. Universal methods of coding for the case of unknown statistics. In *Proceedings of the 5th Symposium on Information Theory*, pages 129—135. Moscow-Gorky, 1972.

[9] R.E. Krichevsky and V.K. Trofimov. The performance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199—207, March 1981.

[10] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629—636, July 1984.

[11] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743—2760, October 1998.

[12] M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Information Theory*, 50:2686–2707, 2004.

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

[14] S. Ben-David and S. Shalev-Schwartz. *Understanding Machine Learning*. Cambridge University Press, 2012.

[15] P. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[16] J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674—682, November 1978.

[17] David Haussler. A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280, 1997.

[18] M. Hosseini and N. Santhanam. Single letter characterization of average-case strong redundancy of compressing memoryless sequences. In *Allerton conference on communication, control and computing*, 2015.

[19] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[20] P. Elias. Universal codeword sets and representations of integers. *IEEE Transactions on Information Theory*, 21(2):194—203, March 1975.

[21] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and sons., 1991.

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

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

[24] S. Ho and R. Yeung. On information divergence measures and joint typicality. *IEEE Transactions on Information Theory*, 56(12):5893–5905, 2010.

[25] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger. Universal discrete denoising: known channel. *IEEE Transactions on Information Theory*, 51(1):5–28, 2005. See also HP Labs Tech Report HPL-2003-29, Feb 2003.

[26] K.L. Chung. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612—614, 1961.

Appendix I
Alternate definitions of strong and weak compressibility

We first establish the following elementary result.

Lemma 29. For $n \geq 1$, let $\hat{q}_n$ be a probability measure on $\mathbb{N}^n$. Then there is a probability measure q_n on $\mathbb{N}^\infty$ such that, for all $\mathbf{x} \in \mathbb{N}^n$, we have $q_n(\mathbf{x}) = \hat{q}_n(\mathbf{x})$.

Proof We define q_n by specifying $q_n(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{N}^m$ for all $m \geq 1$. If $1 \leq m \leq n$ and $\mathbf{y} \in \mathbb{N}^m$, let

$$q_n(\mathbf{y}) := \sum_{\mathbf{x}' \in \mathbb{N}^n : \mathbf{y} \preceq \mathbf{x}'} \hat{q}_n(\mathbf{x}').$$

For $m \geq n$ and $\mathbf{y} \in \mathbb{N}^m$, if $\mathbf{y}$ is $\mathbf{x}$ followed by a string of 1s, for some $\mathbf{x} \in \mathbb{N}^n$, let

$$q_n(\mathbf{y}) := \hat{q}_n(\mathbf{x}),$$

else let $q_n(\mathbf{y}) := 0$. It can be checked that q_n, defined in this way, satisfies the consistency conditions $q_n(\mathbf{z}) = \sum_{\mathbf{y} \in \mathbb{N}^m \,:\, \mathbf{z} \preceq \mathbf{y}} q_n(\mathbf{y})$ for all $1 \leq l \leq m$ and $\mathbf{z} \in \mathbb{N}^l$. Hence q_n defines a probability measure on $\mathbb{N}^\infty$. It can also be checked that q_n satisfies the requirement in the statement of the lemma. $\square$

Using Lemma 29, we now get the following result, which will help establish the equivalence of our definitions of strong and weak compressibility with those common in literature.

Lemma 30. Let Λ be any collection of probability measures on $\mathbb{N}^\infty$ (not necessarily *i.i.d.*). Suppose there exists a sequence of probability measures $\hat{q}_n$ on $\mathbb{N}^n$ such that

$$\limsup_{n\to\infty} \sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{\hat{q}_n(X^n)} = 0.$$

Then there is a probability measure q on $\mathbb{N}^\infty$ such that

$$\limsup_{n\to\infty} \sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{q(X^n)} = 0.$$

Proof For each $n \geq 1$, let the probability measure q_n on $\mathbb{N}^\infty$ be constructed to match the probability measure $\hat{q}_n$ on $\mathbb{N}^n$, as in Lemma 29. Define the probability measure q on $\mathbb{N}^\infty$ that, for each $n \geq 1$ and $\mathbf{x} \in \mathbb{N}^n$, assigns to $\mathbf{x}$ the probability

$$q(\mathbf{x}) := \sum_{i=1}^{\infty} \frac{q_i(\mathbf{x})}{i(i+1)}.$$

For all $n \geq 1$ we therefore have

$$\begin{aligned}
\sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{q(X^n)} &\leq \sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{q_n(X^n)} + \frac{\log(n(n+1))}{n} \\
&= \sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{\hat{q}_n(X^n)} + \frac{\log(n(n+1))}{n}.
\end{aligned}$$

Hence

$$\limsup_{n\to\infty} \sup_{r\in\Lambda} \frac{1}{n} E_r \log \frac{r(X^n)}{q(X^n)} = 0.$$

□

Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. In most prior work [8], [1], [16] the collection $\mathcal{P}$ is called strongly compressible if there is a sequence of probability measures $\hat{q}_n$ on $\mathbb{N}^n$ such that

$$\limsup_{n\to\infty} \sup_{p\in\mathcal{P}^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{\hat{q}_n(X^n)} = 0.$$

Lemma 30 immediately establishes that this definition is equivalent to the definition of strong compressibility that we have made in Definition 2.

The most commonly used definition of weak compressibility in prior work is due to Kieffer [16], and is framed in the language of length functions of compression schemes. Let Λ be any collection of stationary ergodic probability measures on $\mathbb{N}^\infty$ (not necessarily *i.i.d.*). A compression scheme is a sequence of mappings $\phi_n : \mathbb{N}^n \to \{0,1\}^* \backslash \emptyset$ whose image satisfies the prefix condition, i.e. for any two distinct elements in the domain the image of the first is not a prefix of the image of the second. The collection Λ is called weakly compressible if there is a compression scheme $(\phi_n, n \geq 1)$ such that, for all $r \in \Lambda$, we have

$$\lim_{n\to\infty} \frac{1}{n} E_r l(\phi_n(X^n)) = H(r),$$

where $H(r)$ denotes the entropy rate of r.

Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the corresponding collection of *i.i.d.* probability measures on $\mathbb{N}^\infty$ Note that $\mathcal{P}^\infty$ is a collection of stationary ergodic probability measures. We now show that the definition of weak compressibility of $\mathcal{P}^\infty$ in the sense of Kieffer [16] is identical to the definition of weak compressibility of $\mathcal{P}^\infty$ that we have made in Definition 4.

Suppose first that $\mathcal{P}^\infty$ is weakly compressible in the sense of Definition 4. If every probability distribution in $\mathcal{P}$ has infinite entropy, consider an arbitrary compression scheme $(\phi_n, n \geq 1)$, for instance by defining $\phi_n(x^n)$ by concatenating symbol by symbol the representation of $i \in \mathbb{N}$ by a bit string of length $\lceil \log \frac{1}{(i+1)(i+2)} \rceil$ coming from a prefix code for $\mathbb{N}$ corresponding to the probability distribution assigning probability $\frac{1}{(i+1)(i+2)}$ to $i \in \mathbb{N}$. Then we have

$$\frac{1}{n} E_p l(\phi_n(X^n)) \overset{(a)}{\geq} \frac{1}{n} E_p \log \frac{1}{p(X^n)} = \infty, \tag{29}$$

and so

$$\lim_{n\to\infty} \frac{1}{n} E_p l(\phi_n(X^n)) = H(p),$$

for all $p \in \mathcal{P}$. Here (a) in (29) can be seen by picking a probability measure q_n on $\mathbb{N}^n$ that satisfies $l(\phi_n(X^n)) \geq \log \frac{1}{q_n(x^n))}$ and observing that $E_p \log \frac{p(X^n)}{q_n(X^n)} \geq 0$. If there are probability distributions in $\mathcal{P}$ with finite entropy, let q be a probability measure on $\mathbb{N}^\infty$ verifying the requirements in Definition 4. For $n \geq 1$, let $\hat{q}_n$ denote the probability measure on $\mathbb{N}^n$ resulting from restricting q to $\mathbb{N}^n$. We can then define a compression scheme $(\phi_n, n \geq 1)$ such that $l(\phi_n(\mathbf{x})) = \lceil \log \frac{1}{\hat{q}_n(\mathbf{x})} \rceil$ for all $\mathbf{x} \in \mathbb{N}^n$ for all $n \geq 1$. Hence, for every $p \in \mathcal{P}$, we have

$$\frac{1}{n} E_p l(\phi_n(X^n)) = \frac{1}{n} E_p \lceil \log \frac{1}{\hat{q}_n(X^n)} \rceil = \frac{1}{n} E_p \lceil \log \frac{1}{q(X^n)} \rceil.$$

Suppose $H(p) = \infty$. By the same argument as that used in (29) we conclude that $\frac{1}{n} E_p l(\phi_n(X^n)) = \infty$ for all $n \geq 1$ and so, for all such p, we have

$$\lim_{n \to \infty} \frac{1}{n} E_p l(\phi_n(X^n)) = H(p).$$

On the other hand, if $H(p) < \infty$ we have

$$\begin{aligned} \frac{1}{n} E_p l(\phi_n(X^n)) &\leq \frac{1}{n} E_p \log \frac{1}{q(X^n)} + \frac{1}{n} \\ &= \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} + H(p) + \frac{1}{n}, \end{aligned}$$

and so, letting $n \to \infty$, we see that

$$\lim_{n \to \infty} \frac{1}{n} E_p l(\phi_n(X^n)) = H(p)$$

also holds for such p. We have established that $\mathcal{P}^\infty$ is also weakly compressible in the sense of Kieffer [16], irrespective of whether $\mathcal{P}$ is comprised entirely of probability distributions with infinite entropy or also contains probability distributions with finite entropy.

For the converse, suppose that $\mathcal{P}^\infty$ is weakly compressible in the sense of Kieffer [16]. For each $n \geq 1$ we can find a probability measure $\hat{q}_n$ on $\mathbb{N}^n$ such that $\hat{q_n}(\mathbf{x}) \geq 2^{-l(\phi_n(\mathbf{x}))}$ for all $\mathbf{x} \in \mathbb{N}^n$, where $(\phi_n, n \geq 1)$ is a compression scheme verifying the weak compressibility of $\mathcal{P}^\infty$ in the sense of Kieffer [16]. For each $n \geq 1$ we define the probability measure q_n on $\mathbb{N}^\infty$ in terms of $\hat{q}_n$ as in Lemma 29, and we define the probability measure q on $\mathbb{N}^\infty$ which, for each $n \geq 1$ and $\mathbf{x} \in \mathbb{N}^n$, assigns to $\mathbf{x}$ the probability

$$q(\mathbf{x}) := \sum_{i=1}^{\infty} \frac{q_i(\mathbf{x})}{i(i+1)}.$$

For each $p \in \mathcal{P}$ with finite entropy, we have

$$\begin{aligned} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} &\leq \frac{1}{n} E_p \log \frac{p(X^n)}{q_n(X^n)} + \frac{\log n(n+1)}{n} \\ &= \frac{1}{n} E_p \log \frac{p(X^n)}{\hat{q}_n(X^n)} + \frac{\log n(n+1)}{n} \\ &\leq -H(p) + \frac{1}{n} E_p l(\phi_n(X^n)) + \frac{\log n(n+1)}{n}, \end{aligned}$$

and so, from $\lim_{n \to \infty} \frac{1}{n} E_p l(\phi_n(X^n)) = H(p)$, we conclude that $\limsup_{n \to \infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} = 0$. This proves that $\mathcal{P}^\infty$ is weakly compressible in the sense of Definition 4.

To close this section, we give proofs of two statements that allow us to think about strong compressibility and weak compressibility respectively in terms of vanishing asymptotic per-symbol redundancy.

Lemma 31. Let $\mathcal{P}$ be a collection of probability distribution on $\mathbb{N}$ and $\mathcal{P}^\infty$ the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. Then $\mathcal{P}^\infty$ is strongly compressible iff it has zero asymptotic per-symbol redundancy.

Proof

If $\mathcal{P}^\infty$ is strongly compressible, then taking the probability measure q on $\mathbb{N}^\infty$ which verifies the strong compressibility condition in (1) from Definition 2 as the q in (2) from Definition 3 for each $n \geq 1$ immediately implies that $\mathcal{P}^\infty$ has zero asymptotic per-symbol redundancy.

Conversely, suppose $\mathcal{P}^\infty$ has zero asymptotic per-symbol redundancy. Given $\epsilon > 0$, for each $n \geq 1$ let q_n be a probability measure on $\mathbb{N}^\infty$ for which $\sup_{p\in\mathcal{P}^\infty} E_p \log \frac{p(X^n)}{q_n(X^n)} \leq R_n + \epsilon$, and define the probability measure q on $\mathbb{N}^\infty$ by

$$q(\mathbf{x}) := \sum_{i=1}^{\infty} \frac{q_i(\mathbf{x})}{i(i+1)}.$$

Then we have

$$\frac{1}{n} \sup_{p\in\mathcal{P}^\infty} E_p \log \frac{p(X^n)}{q(X^n)} \leq \frac{1}{n} \sup_{p\in\mathcal{P}^\infty} E_p \log \frac{r(X^n)}{q_n(X^n)} + \frac{\log(n(n+1))}{n},$$

and so

$$\limsup_{n\to\infty} \frac{1}{n} \sup_{p\in\mathcal{P}^\infty} E_p \log \frac{p(X^n)}{q(X^n)} \leq \epsilon.$$

Letting $\epsilon \to 0$ shows that $\mathcal{P}^\infty$ is strongly compressible. □

Lemma 32. Let $\mathcal{P}$ be a collection of probability distribution on $\mathbb{N}$ and $\mathcal{P}^\infty$ the collection of probability measures on $\mathbb{N}^\infty$ induced by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. Then $\mathcal{P}^\infty$ is weakly compressible iff there is a probability measure q on $\mathbb{N}^\infty$ such that for every $p \in \mathcal{P}$ with finite entropy the corresponding $p^\infty \in \mathcal{P}^\infty$ has zero asymptotic per-symbol redundancy with respect to q.

Proof

The claim is vacuously true if all the probability distributions in $\mathcal{P}$ have infinite entropy. If there are distributions in $\mathcal{P}$ with finite entropy and $\mathcal{P}^\infty$ is weakly compressible, then consider the probability measure q on $\mathbb{N}^\infty$ which verifies the weak compressibility condition in (3) from Definition 4. By definition, with respect to this q, every $p \in \mathcal{P}$ with finite entropy is such that the corresponding $p^\infty \in \mathcal{P}^\infty$ has zero asymptotic per-symbol redundancy with respect to q. Conversely, if there are distributions in $\mathcal{P}$ with finite entropy and there is a probability measure q on $\mathbb{N}^\infty$ such that for every $p \in \mathcal{P}$ the corresponding $p^\infty \in \mathcal{P}^\infty$ has zero asymptotic per-symbol redundancy with respect to q then, by definition, this q satisfies the condition in (3) from Definition 4 for all $p \in \mathcal{P}$ with finite entropy. This establishes that $\mathcal{P}^\infty$ is weakly compressible. □

Appendix II
Basic properties of relative entropy and redundancy

In this appendix we gather some basic results on the KL divergence and redundacy, which are used at various points in the document.

Proposition 33. Let p and q be two probability distributions on a countable set $\mathcal{X}$. Then

$$\sum_{x\in\mathcal{X}} p(x) \left| \log \frac{p(x)}{q(x)} \right| \leq D(p||q) + 2\frac{\log e}{e}.$$

Proof Let $S \subset \mathcal{X}$ be the set of all elements $x \in \mathcal{X}$ such that $p(x) \leq q(x)$. Note that $q(S) > 0$. We have

$$\begin{aligned} D(p||q) - \sum_{x\in\mathcal{X}} p(x) \left|\log \frac{p(x)}{q(x)}\right| &= 2\sum_{x\in S} p(x)\log\frac{p(x)}{q(x)} \\ &\overset{(a)}{\geq} 2p(S)\log\frac{p(S)}{q(S)} \\ &\geq 2p(S)\log p(S) \\ &\geq -2\frac{\log e}{e}, \end{aligned}$$

where step (a) is from the log sum inequality. The proposition follows. □

Proposition 34. For all probability measures r and q on $\mathbb{N}^\infty$ and all $1 \leq m \leq n$, we have

$$D_m(r||q) \leq D_n(r||q).$$

In particular, for any collection of probability distributions $\mathcal{P}$ on $\mathbb{N}$, if $\mathcal{P}^\infty$ denotes the associated collection of *i.i.d.* probability measures on $\mathbb{N}^\infty$, we will have

$$R_m(\mathcal{P}) := \inf_q \sup_{p\in\mathcal{P}} E_p \log\frac{p(X^m)}{q(X^m)} \leq \inf_q \sup_{p\in\mathcal{P}} E_p \log\frac{p(X^n)}{q(X^n)} = R_n(\mathcal{P}),$$

where the outer infimum on both sides is taken over all probability measures q on $\mathbb{N}^\infty$ and so $R_m(\mathcal{P})$ and $R_n(\mathcal{P})$ are the length-m redundancy and the length-n redundancy of $\mathcal{P}$, respectively.

Proof The first part of the claim follows from convexity, because, for all $y^m \in \mathbb{N}^m$, we have

$$r(y^m) = \sum_{x^n \,:\, y^m \preceq x^n} r(x^n) \text{ and } q(y^m) = \sum_{x^n \,:\, y^m \preceq x^n} q(x^n).$$

For the second part of the claim, for any $\epsilon > 0$ pick a probability measure q' on $\mathbb{N}^\infty$ such that

$$\sup_{p\in\mathcal{P}} E_p \log\frac{p(X^n)}{q'(X^n)} < R_n(\mathcal{P}) + \epsilon.$$

It then follows from the first part of the claim that

$$R_m(\mathcal{P}) \leq \sup_{p\in\mathcal{P}} E_p \log\frac{p(X^m)}{q'(X^m)} < R_n(\mathcal{P}) + \epsilon.$$

We let $\epsilon \to 0$ to complete the proof. □

Proposition 35. Let $\mathcal{P}$ be a collection of probability distributions on $\mathbb{N}$ and $\mathcal{P}^\infty$ the corresponding collection of probability measures on $\mathbb{N}^\infty$ got by *i.i.d.* sampling from the individual probability distributions in $\mathcal{P}$. For $n \geq 1$, let R_n denote the length-n redundancy of $\mathcal{P}^\infty$, as defined in (2). Then, for all $n \geq 1$, the per-symbol length-n redundancy of $\mathcal{P}^\infty$ satisfies $R_n/n \leq R_1$.

Proof Let $\epsilon > 0$. Let $\tilde{p}$ be a probability distribution on $\mathbb{N}$ such that the single letter redundancy of $\mathcal{P}^\infty$ with respect to $\tilde{p}$ is strictly less than $R_1 + \epsilon$. With the usual abuse of notation, let $\tilde{p}$ also denote the *i.i.d.* probability measure on $\mathbb{N}^\infty$ corresponding to $\tilde{p}$. Then, for all $p \in \mathcal{P}$, we have

$$\frac{1}{n} E_p \log\frac{p(X^n)}{\tilde{p}(X^n)} = E_p \log\frac{p(X)}{\tilde{p}(X)} < (R_1 + \epsilon).$$

By letting $\epsilon \to 0$, the proposition follows. □

Lemma 36. Let Λ be a collection of probability measures on $\mathbb{N}^\infty$. Then we have

$$\limsup_{n\to\infty} \frac{1}{n} \inf_q \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)} = \inf_q \limsup_{n\to\infty} \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)}, \tag{30}$$

where the infimum is taken over all probability measures q on $\mathbb{N}^\infty$. Namely, the $\limsup_{n\to\infty}$ can be interchanged with the $\inf_q$ in the definition of the asymptotic per-symbol redundancy of Λ.

Proof Fix $\epsilon > 0$. For $n \geq 1$, let q_n be a probability measure on $\mathbb{N}^\infty$ such that

$$\frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q_n(X^n)} < \frac{1}{n} R_n + \epsilon.$$

Define the probability measure $\bar{q}$ on $\mathbb{N}^\infty$ that, for each $n \geq 1$ and $\mathbf{x} \in \mathbb{N}^n$, assigns to $\mathbf{x}$ the probability

$$\bar{q}(\mathbf{x}) := \sum_{i=1}^{\infty} \frac{q_i(\mathbf{x})}{i(i+1)},$$

where, as usual, $q_i(\mathbf{x})$ is the probability under q_i of the event in $\mathbb{N}^\infty$ comprised of the sequences having the prefix $\mathbf{x}$. We then have

$$\frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{\bar{q}(X^n)} \leq \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q_n(X^n)} + \frac{\log(n(n+1)}{n} < \frac{1}{n} R_n + \epsilon + \frac{\log(n(n+1)}{n}.$$

Thus

$$\inf_q \limsup_{n\to\infty} \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)} \leq \limsup_{n\to\infty} \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{\bar{q}(X^n)} \leq \limsup_{n\to\infty} \frac{1}{n} R_n + \epsilon.$$

Letting $\epsilon \to 0$, we see that the term on the right hand side of (30) is no bigger than the term on its left hand side. Showing the inequality in the other direction is straightforward, since

$$\frac{1}{n} \inf_q \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)} \leq \frac{1}{n} \sup_{r\in\Lambda} E_r \log \frac{r(X^n)}{q(X^n)},$$

for each probability measure q on $\mathbb{N}^\infty$. This completes the proof. $\square$

For the following lemma, recall the definition of $\mathcal{J}(p, \tilde{p})$ for probability distributions p and $\tilde{p}$ on $\mathbb{N}$, made in (11).

Lemma 37. Let p and $\tilde{p}$ be probability distributions on $\mathbb{N}$. Then

$$\frac{\log e}{4} |p - \tilde{p}|_1^2 \leq \mathcal{J}(p, \tilde{p}) \leq |p - \tilde{p}|_1 \log e \ .$$

If, in addition, p' is a probability distribution on $\mathbb{N}$, then

$$\mathcal{J}(p, \tilde{p}) + \mathcal{J}(\tilde{p}, p') \geq \mathcal{J}^2(p, p') \frac{1}{8 \log e}.$$

Proof The lower bound in the first statement follows from Pinsker's inequality for the KL divergence, see [21] for example, from which we get

$$D\left(p || \frac{p + \tilde{p}}{2}\right) \geq \frac{\log e}{8} |p - \tilde{p}|_1^2,$$

and similarly for $D\Big(\tilde{p}||\frac{p+\tilde{p}}{2}\Big)$. For the upper bound in the first statement, since $\log(1+z) \leq z \log e$ for all $z \geq 0$, we may write

$$\begin{aligned}\frac{1}{\log e}\mathcal{J}(p,\tilde{p}) &\leq \sum_{x:p(x)\geq\tilde{p}(x)} p(x)\left(\frac{p(x)-\tilde{p}(x)}{p(x)+\tilde{p}(x)}\right) + \sum_{x:\tilde{p}(x)\geq p(x)} \tilde{p}(x)\left(\frac{\tilde{p}(x)-p(x)}{p(x)+\tilde{p}(x)}\right)\\ &\leq |p-\tilde{p}|_1.\end{aligned}$$

To prove the triangle-like inequality, note that

$$\begin{aligned}\mathcal{J}(p,\tilde{p}) + \mathcal{J}(\tilde{p},p') &\geq \frac{\log e}{4}\big(|p-\tilde{p}|_1^2 + |\tilde{p}-p'|_1^2\big)\\ &\geq \frac{\log e}{8}\big(|p-\tilde{p}|_1 + |\tilde{p}-p'|_1\big)^2\\ &\geq \frac{\log e}{8}\big(|p-p'|_1\big)^2\\ &\geq \frac{1}{8\log e}\mathcal{J}(p,p')^2,\end{aligned}$$

where the last inequality follows from the upper bound on $\mathcal{J}(p,p')$ already proved in the first part of the statement. □

Using Lemma 37, we can prove the following result, which is identical to [2, Lemma 6]. We reproduce the proof from [2] for completeness.

Lemma 38. Let $\epsilon_0 > 0$. If

$$|p_0 - q|_1 \leq \frac{\epsilon_0^2(\ln 2)^2}{16} \ ,$$

then for all $p \in \mathcal{P}$ with $\mathcal{J}(p,p_0) \geq \epsilon_0$, we have

$$\mathcal{J}(p,q) \geq \frac{\epsilon_0^2 \ln 2}{16}.$$ □

Proof Since

$$|p_0 - q|_1 \leq \frac{\epsilon_0^2(\ln 2)^2}{16},$$

Lemma 37 implies that

$$\mathcal{J}(p_0,q) \leq \frac{\epsilon_0^2 \ln 2}{16}.$$

Further, Lemma 37 then implies that

$$\mathcal{J}(p,q) + \frac{\epsilon_0^2 \ln 2}{16} \ \geq \ \mathcal{J}(p,q) + \mathcal{J}(p_0,q) \ \geq \ \frac{\mathcal{J}^2(p,p_0)\ln 2}{8} \ \geq \ \frac{\epsilon_0^2 \ln 2}{8},$$

where the last inequality follows since $\mathcal{J}(p,p_0) \geq \epsilon_0$. This completes the proof. □

The following result from [2] will be needed to prove the necessity part of Theorem 17.

Lemma 39. Fix $\epsilon > 0$. Let p and q be probability distributions on $\mathbb{N}$ with $\mathcal{J}(p,q) \leq \epsilon$. Fix $n \in \mathbb{N}$. Consider the probability measures on $\mathbb{N}^n$ obtained by *i.i.d.* sampling from p and q respectively, which we continue to denote by p and q respectively, following our convention.

Suppose $A_n \subset \mathbb{N}^n$ is subset for which $p(A_n) \geq 1 - \alpha$, for some $\alpha > 0$. Then we have

$$q(A_n) > 1 - \alpha - 2n^3 \sqrt{\frac{4\epsilon}{\log e}} - \frac{1}{n}. \qquad \square$$

Proof Let

$$\mathcal{B}_1 := \left\{ i \in \mathbb{N} : q(i) \leq p(i)\left(1 - \frac{1}{n^2}\right)\right\}, \text{ and } \mathcal{B}_2 := \left\{ i \in \mathbb{N} : p(i) \leq q(i)\left(1 - \frac{1}{n^2}\right)\right\}.$$

Since we have assumed that $\mathcal{J}(p,q) \leq \epsilon$ we have, from Lemma 37, that

$$|p - q|_1 \sqrt{\frac{\log e}{4}} \leq \sqrt{\mathcal{J}(p,q)} \leq \sqrt{\epsilon}.$$

Further, we have

$$|p - q|_1 \geq \sum_{x \in \mathcal{B}_1} (p(x) - q(x)) \geq \frac{p(\mathcal{B}_1)}{n^2} \geq \frac{q(\mathcal{B}_1)}{n^2},$$

and similarly

$$|p - q|_1 \geq \sum_{x \in \mathcal{B}_2} (q(x) - p(x)) \geq \frac{q(\mathcal{B}_2)}{n^2} \geq \frac{p(\mathcal{B}_2)}{n^2}.$$

From the preceding inequalities, it follows that

$$p(\mathcal{B}_1 \cup \mathcal{B}_2) \leq 2n^2 \sqrt{\frac{4\epsilon}{\log e}} \text{ and } q(\mathcal{B}_1 \cup \mathcal{B}_2) \leq 2n^2 \sqrt{\frac{4\epsilon}{\log e}}. \tag{31}$$

Let $S := \mathbb{N} - (\mathcal{B}_1 \cup \mathcal{B}_2)$. For all $x \in S$ we have

$$q(x) \geq p(x)\left(1 - \frac{1}{n^2}\right). \tag{32}$$

In addition, from (31) we have

$$p(S) \geq 1 - 2n^2 \sqrt{\frac{4\epsilon}{\log e}}.$$

Let $S_n \subset \mathbb{N}^n$ denote the set of all length-n strings of symbols from S. Clearly

$$p(S_n) \geq (1 - 2n^2 \sqrt{\frac{4\epsilon}{\log e}})^n > 1 - 2n^3 \sqrt{\frac{4\epsilon}{\log e}}.$$

Thus we have

$$p(A_n \cap S_n) > 1 - 2n^3 \sqrt{\frac{4\epsilon}{\log e}} - \alpha.$$

From (32), for all $x^n \in S_n$, we have

$$q(x^n) \geq p(x^n)\left(1 - \frac{1}{n^2}\right)^n > p(x^n)\left(1 - \frac{1}{n}\right).$$

Therefore,

$$q(A_n) \geq q(A_n \cap S_n) > (1 - 2n^3 \sqrt{\frac{4\epsilon}{\log e}} - \alpha)\left(1 - \frac{1}{n}\right) > 1 - \alpha - 2n^3 \sqrt{\frac{4\epsilon}{\log e}} - \frac{1}{n}. \qquad \square$$

Appendix III
Length-n per-symbol redundancy of $\mathcal{M}_h$

We construct a probability measure q^* on $\mathbb{N}^\infty$ such that for $\mathcal{M}_h$ we have

$$\sup_{p\in\mathcal{M}_h^n} \frac{1}{n} D_n(p||q) \leq \frac{2h^{\frac{1}{4}}(\sqrt{h}+1)}{\sqrt{\ln n}} + \pi\sqrt{\frac{2}{3n}}\log e.$$

This implies that the per-symbol length-n redundancy of $\mathcal{M}_h$ diminishes to 0 as $n \to \infty$. Hence $\mathcal{M}_h$ is strongly compressible.

Consider the probability distribution q on $\mathbb{N}$ defined by $q(i) = 1/i(i+1), i \geq 1$. As observed in Example 24, we have

$$\sup_{p\in\mathcal{M}_h} E_p\left(\lceil \log \frac{1}{q(X)} \rceil\right)^2 < 4(\sqrt{h}+1)^2. \tag{33}$$

We consider a scheme that encodes patterns [22] of symbols (i.e. natural numbers in our case) first, followed by an encoding using $\lceil \log \frac{1}{q(x)} \rceil$ bits to describe every symbol x that appeared in the string, in the order in which they arrived. To clarify, recall that the pattern of a sequence of symbols from $\mathbb{N}$ replaces each symbol by $k \in \mathbb{N}$ if the symbol was the k-th new symbol to appear in the sequence. For example, the pattern of the sequence of natural numbers $(2, 3, 17, 4, 3, 3, 1, 2, 4)$ is $(1, 2, 3, 4, 2, 2, 5, 1, 4)$. If in addition to the pattern of a finite sequence of natural numbers, in which there are l distinct symbols, one knows which symbol was the k-th symbol to appear for each $1 \leq k \leq l$, one learns the sequence of symbols.

The expected (not normalized by n) additional number of bits to encode the pattern of a sequence of symbols of length n from any $p \in \mathcal{M}_h$ is at most $\pi\sqrt{\frac{2}{3}n}\log e$, using the results in [22], while the expected number of bits to describe the symbols of length-n strings using a prefix code based on the probability distribution q on $\mathbb{N}$ is at most

$$\sum_{i\in\mathbb{N}} (1-(1-p(i))^n) \lceil \log \frac{1}{q(i)} \rceil.$$

Note that the distinct symbols appearing the the string will need to be specified in the order in which they arrived. Let M_n denote the number of distinct symbols that appear in a sequence of length n. Then the expected number of extra bits the scheme uses for length-n strings is (without normalizing by

n) at most $\pi\sqrt{\frac{2}{3}n}\log e$ plus at most

$$
\begin{aligned}
&\sum_{i\in\mathbb{N}}(1-(1-p(i))^n)\lceil\log\frac{1}{q(i)}\rceil \\
&\overset{(a)}{\leq} \sqrt{\sum_{i\in\mathbb{N}}(1-(1-p_i)^n)\sum_{j\in\mathbb{N}}(1-(1-p_j)^n)\left(\lceil\log\frac{1}{q(j)}\rceil\right)^2} \\
&\leq \sqrt{\sum_{i\in\mathbb{N}}(1-(1-p_i)^n)\sum_{j\in\mathbb{N}}(np_j)\left(\lceil\log\frac{1}{q(j)}\rceil\right)^2} \\
&\overset{(b)}{\leq} \sqrt{4(\mathbb{E}M_n)n(\sqrt{h}+1)^2} \\
&\overset{(c)}{\leq} \frac{2nh^{1/4}(\sqrt{h}+1)}{\sqrt{\ln n}}.
\end{aligned}
$$

Here (a) follows from the Cauchy-Schwarz inequality, while (b) follows from (33) and the definition of M_n. As for (c), a result similar to (c) can be found in [23], but we justify (c) below for completeness. We observe that for all $i\in\mathbb{N}$ we have

$$
\begin{aligned}
1-(1-p_i)^n &= p_i\sum_{j=0}^{n-1}(1-p_i)^j \\
&\leq p_i\left(\sum_{j=0}^{n-1}(1-p_i)^j\right)\frac{\sum_{k=1}^{n}\frac{1}{k}}{\ln n} \\
&\overset{(a)}{\leq} \frac{np_i}{\ln n}\sum_{j=0}^{n-1}\frac{(1-p_i)^j}{j} \\
&\leq \frac{np_i\log\frac{1}{p_i}}{\ln n}.
\end{aligned}
$$

Combining the above with the fact that the entropy of any $p\in\mathcal{M}_h$ is at most $\sqrt{h}$, which was shown in Example 24, proves (c) in the previous set of equations. In the above set of equations, inequality (a) follows from Minkowski's inequality which says that if x_i and y_i $(0\leq i\leq n-1)$ are both decreasing positive sequences, then $n\sum x_iy_i \geq \sum x_j\sum y_k$. Minkowski's inequality is easily proved by noting $\sum x_j\sum y_k = \sum_m\sum x_iy_{(i+m)\mod n}$. and that $\sum x_iy_i \geq \sum x_iy_{(i+m)\mod n}$ for all $0\leq m\leq n-1$.

The claim about the per-symbol length-n redundancy of $\mathcal{M}_h$ follows after normalization by n.

Appendix IV
Typicality of empirical distributions that are not too spread out

In this section we prove a useful result quantifying how close the empirical distribution of a sample drawn *i.i.d.* from a probability distribution p on $\mathbb{N}$ is to p, when the alphabet of symbols showing up in the sample is not too spread out. There is a lemma that looks somewhat similar in [24]. The difference of the result in Lemma 40 from that in [24] is that the right side of the inequality in (34) does *not* depend on p. The result of Lemma 40 will be used in the sufficiency proof in Appendix VI and this property is crucial for its use.

Lemma 40. Let p be any probability distribution on $\mathbb{N}$. Let $\gamma > 0$ and let $k \geq 2$ be an integer. Let X_1^n be a sequence generated *i.i.d.* with marginals p and let $t(X^n)$ be the empirical distribution of X_1^n. Then

$$p\Big(|t(X^n) - p|_1 > \gamma \text{ and } 2\dot{F}_t^{-1}(1-\gamma/6) \leq k\Big) \leq (2^k - 2)\exp\left(-\frac{n\gamma^2}{18}\right). \tag{34}$$

Proof From [25, Proposition 1] we know that for any probability distribution p' on $\mathbb{N}$ with finite support of size L we have

$$p'(|t(X^n) - p'|_1 \geq \alpha) \leq (2^L - 2)\exp\left(-\frac{n\alpha^2}{2}\right), \tag{35}$$

where $t(X^n)$ is the type of X^n generated *i.i.d.* with marginal distribution p'.

Consider the probability distributions p' and t' on A obtained from p and t respectively via the mapping from $\mathbb{N}$ to $A := \{1, \ldots, k-1\} \cup \{-1\}$ that maps i to i for $0 \leq i \leq k-1$ and maps all the other natural numbers to -1. Thus, we have

$$p'(i) = \begin{cases} p(i), & \text{if } 1 \leq i \leq k-1, \\ \sum_{j=k}^{\infty} p(j), & \text{if } i = -1. \end{cases}$$

Further, sequences of natural numbers generated *i.i.d.* with marginal distribution p and with empirical distribution t are mapped to sequences from A that are *i.i.d.* with probability distribution p' and have empirical distribution t'.

Applying (35) to p', we have

$$p'(|p' - t'|_1 > \gamma/3) \leq (2^k - 2)\exp\left(-\frac{n\gamma^2}{18}\right). \tag{36}$$

We first argue that all sequences generated by p with empirical distributions t satisfying

$$|p - t|_1 > \gamma \text{ and } 2\dot{F}_t^{-1}(1-\gamma/6) \leq k$$

are mapped into sequences generated by p' with empirical t' satisfying

$$|p' - t'|_1 > \gamma/3 \text{ and } t'(-1) \leq \gamma/3.$$

This follows from writing

$$\begin{aligned} |p - t|_1 - \sum_{i=1}^{k-1} |p(i) - t(i)| & \\ &\leq \sum_{j=k}^{\infty} (p(j) - t(j)) + 2\sum_{j=k}^{\infty} t(j) \\ &\leq |p'(-1) - t'(-1)| + \gamma/3, \end{aligned}$$

where the last inequality above follows from the fact that $2\dot{F}_t^{-1}(1-\gamma/6) \leq k$ implies $F_t(k-1) \geq 1-\gamma/6$, i.e. $\sum_{j=k}^{\infty} t(j) \leq \gamma/6$. Hence we have

$$|p' - t'|_1 = \sum_{i=1}^{k-1} |p(i) - t(i)| + |p'(-1) - t'(-1)| \geq |p - t|_1 - \gamma/3 > \gamma/3,$$

because $|p-t|_1 > \gamma$.

Thus, from (36), we will have

$$
\begin{aligned}
& p(|t(X^n) - p|_1 > \gamma \text{ and } 2\dot{F}_t^{-1}(1-\gamma/6) \le k) \\
& \le p'(|t'-p'|_1 > \gamma/3 \text{ and } t'(-1) \le \gamma/3) \\
& \le (2^k - 2)\exp\left(-\frac{n\gamma^2}{18}\right).
\end{aligned}
$$

This completes the proof of the lemma. □

Appendix V
τ enters with probability 1

We reproduce the argument from [2] here for completeness.

Every probability distribution $p \in \mathcal{P}$ is contained in at least one of the elements of the cover $(Q_{p,m} \cap \mathcal{P}, \tilde{p} \in \tilde{\mathcal{P}}_m)$, where $Q_{p,m}$ denotes the zone of $\tilde{p} \in \tilde{\mathcal{P}}_m$. Recall the enumeration of $\tilde{\mathcal{P}}_m$. Let p' be be centroid with the smallest index among all centroids in $\tilde{\mathcal{P}}_m$ whose zones contain p. With probability 1, sequences generated by p will eventually have their type (empirical distribution) entirely within $Q_{p',m}$. (see [26] for a proof).

Next note that for all n sufficiently large the analog of (26), (which makes sense for all $p' \in \tilde{\mathcal{P}}_m$) will hold. This follows since the right hand side of (26) diminishes to zero polynomially with n while the left hand side diminishes to zero exponentially fast in n.

Next, (27) will also hold eventually with probability 1, since, if t denotes the empirical probability of a sequence generated by p, then

$$
\dot{F}_t^{-1}(1-\sqrt{D_{p'}}/6) \to \dot{F}_p^{-1}(1-\sqrt{D_{p'}}/6) \tag{37}
$$

with probability 1 as $n \to \infty$, where we note that the quantity on the left hand side of (37) is actually a random variable and t determines n. Furthermore, we also have

$$
\begin{aligned}
2\dot{F}_p^{-1}(1-\sqrt{D_{p',m}}/6) &< 3\left(\sup_{r \in B(p',\epsilon_{p',m};\mathcal{P})} \dot{F}_r^{-1}(1-\sqrt{D_{p',m}}/6)\right) \\
&= \log C(p',m),
\end{aligned}
$$

where the first inequality follows since p is in the $\frac{1}{m}$−reach of p'.

Therefore, both (26) and (27) will eventually hold with probability 1. Furthermore, long enough sequences generated by p fall into the zone of p' with probability 1. This implies in turn that $\tau_{\eta,m}$ enters with probability 1. Note that it is entirely possible that some other probability measure traps strings before they can be trapped by p', but that does not take away from the fact that $\tau_{\eta,m}$ will enter with probability 1.

Appendix VI
Probability of falling into bad traps

Let t be any length-n empirical distribution trapped by $\hat{p}$, which we recall has $\frac{1}{m}$-reach $\epsilon_{\hat{p},m}$, such that $p \notin B(\hat{p}, \epsilon_{\hat{p},m}; \mathcal{P})$. Then we have

$$
\mathcal{J}(\hat{p}, p) \ge \epsilon_{\hat{p},m},
$$

because $p \notin B(\hat{p}, \epsilon_{\hat{p},m}; \mathcal{P})$, and we have

$$|\hat{p} - t|_1 < \frac{{\epsilon_{\hat{p},m}}^2 (\ln 2)^2}{16},$$

because t has to be in the zone $Q_{\hat{p},m}$ in order to be captured by $\hat{p}$. From Lemma 38, which is a consequence of the pseudo-triangle inequality in Lemma 37, we get

$$\mathcal{J}(p,t) \geq \frac{\epsilon^2_{\hat{p},m} \ln 2}{16}.$$

Hence, for all types t that are trapped by $\hat{p}$, by the first part of Lemma 37 we get

$$|p - t|_1^2 \geq \mathcal{J}^2(p,t)(\ln 2)^2 \geq \frac{\epsilon^4_{\hat{p},m}(\ln 2)^4}{256} = D^2_{\hat{p},m}.$$

This means that for every $p \in \mathcal{P}$, the probability that length-n sequences with empirical distribution t are trapped by a bad $\hat{p}$ can be bounded from above as

$$\begin{aligned}
&\leq p\left(|t-p|_1^2 \geq D_{\hat{p},m} \text{ and } 2\dot{F}_t^{-1}(1 - \frac{\sqrt{D_{\hat{p},m}}}{6}) \leq \log C(\hat{p},m)\right) \\
&= p\left(|t-p|_1 \geq \sqrt{D_{\hat{p},m}} \text{ and } 2\dot{F}_t^{-1}(1 - \sqrt{D_{\hat{p},m}}/6) \leq \log C(\hat{p},m)\right) \\
&\overset{(a)}{\leq} (C(\hat{p},m) - 2)\exp\left(-\frac{nD_{\hat{p},m}}{18}\right) \\
&\overset{(b)}{\leq} \frac{\eta(C(\hat{p},m)-2)}{2C(\hat{p},m)\iota(\hat{p})^2 n(n+1)} \\
&\leq \frac{\eta}{2\iota(\hat{p})^2 n(n+1)},
\end{aligned}$$

where the inequality (a) follows from Lemma 40 and (b) from (26). Therefore, the probability of sequences falling into bad traps is bounded above by

$$\leq \sum_{n\geq 1} \sum_{\tilde{p} \in \tilde{\mathcal{P}}} \frac{\eta}{2\iota(\tilde{p})^2 n(n+1)} \leq \frac{\pi^2}{12}\eta < \eta,$$

since $\sum_{\tilde{p}\in\tilde{\mathcal{P}}} \frac{1}{\iota(\tilde{p})^2} = \frac{\pi^2}{6}$ and $\sum_{n\geq 1} \frac{1}{n(n+1)} = 1$.

Appendix VII
A fake proof

In this section we give a fake proof of the following mistaken claim: *if* $\mathcal{P}_1$ *and* $\mathcal{P}_2$ *are* d.w.c.*, then* $\mathcal{P}_1 \cup \mathcal{P}_2$ *is also* d.w.c.. We then explain why it is wrong. In the concluding remarks in [2] it was stated, in passing, that if $\mathcal{P}_1$ and $\mathcal{P}_2$ are insurable then $\mathcal{P}_1 \cup \mathcal{P}_2$ is also insurable. This statement if false, for the reasons explained in this section. This does not affect any of the results in [2].

The argument proceeds as follows. Since $\mathcal{P}_i$ is *d.w.c.* for each $i = 1,2$, there is a probability measure q_i on $\mathbb{N}^\infty$ for each $i = 1,2$ such that for every $m \geq 1$, $0 < 1-\eta < 1$ and $i = 1,2$ there is a universal stopping time $\tau^{(i)}_{\eta,m}$ such that, for all $p \in \mathcal{P}_i$, we have

$$p\left(\exists n \text{ such that } \frac{1}{n}D_n(p||q_i) > \frac{1}{m} \text{ and } \tau^{(i)}_{\eta,m}(X^n) = 1\right) < \eta.$$

Let $q := (q_1 + q_2)/2$ and, for accuracy $\frac{1}{m} > 0$ and confidence $0 < 1 - \eta < 1$, define

$$\tau_{\eta,m}(\mathbf{x}) := \mathbb{1}(\tau^{(1)}_{\eta,2m}(\mathbf{x}) = 1)\mathbb{1}(\tau^{(2)}_{\eta,2m}(\mathbf{x}) = 1)\mathbb{1}(|\mathbf{x}| > 2m). \tag{38}$$

Now, suppose $p \in \mathcal{P}_1 \cup \mathcal{P}_2$. Without loss of generality, assume that $p \in \mathcal{P}_1$. Now, if $n > 2m$ and we have

$$\frac{1}{n} D_n\left(p || \frac{q_1 + q_2}{2}\right) > \frac{1}{m},$$

then we have

$$\frac{1}{n} D_n(p||q_1) > \frac{1}{m} - \frac{1}{n} > \frac{1}{2m}$$

Further, from (38), if $\tau_{\eta,m}(\mathbf{x}) = 1$, then we have $\tau^{(1)}_{\eta,2m}(\mathbf{x}) = 1$ as well. Therefore

$$\begin{aligned} &p\left(\exists n \text{ such that } \frac{1}{n} D_n\left(p || \frac{q_1 + q_2}{2}\right) > \frac{1}{m} \text{ and } \tau_{\eta,m}(X^n) = 1\right) \\ &\quad \leq p\left(\exists n \text{ such that } n > 2m,\ \frac{1}{n} D_n(p||q_1) > \frac{1}{2m} \text{ and } \tau^{(1)}_{\eta,2m}(X_1^n) = 1\right) < \eta, \end{aligned}$$

where we have used (38) to see that the event whose probability is being evaluated on the left hand side of the preceding equation cannot occur unless $n > 2m$. Since the above holds for all $p \in \mathcal{P}_1$ and we can use a similar argument for all $p \in \mathcal{P}_2$, we are "done".

The flaw in the above "proof" is that $\tau_{\eta,m}$, as defined in (38), does not necessarily eventually equal 1 almost surely for all sources in $\mathcal{P}_1 \cup \mathcal{P}_2$, which would mean that it is not a universal stopping time for the model class $\mathcal{P}_1 \cup \mathcal{P}_2$. To see why this issue might arise, note that $\tau^{(i)}_{\eta,2m}$ is known to eventually equal 1 almost surely only for sources in $\mathcal{P}_i$. Thus, if it happens to be the case that there is some event $A \subsetneq \mathbb{N}^\infty$ and $p_1 \in \mathcal{P}_1$ with $p_1(A) > 0$ for which we have $p_2(A) = 0$ for every source $p_2 \in \mathcal{P}_2$, then $\tau^{(2)}_{\eta,2m}$ might never stop waiting on the sequences in A. This doesn't stop $\mathcal{P}_2$ from being *d.w.c.*. But when we introduce sources from $\mathcal{P}_1$, in particular p_1, we find that $\tau_{\eta,m}$, as defined in (38), will never stop waiting under p_1. The stopping rule $\tau_{\eta,m}$ would then not be a universal stopping rule for the model class $\mathcal{P}_1 \cup \mathcal{P}_2$.