Unsupervised Learning and Universal Communication

Vinith Misra and Tsachy Weissman
Department of Electrical Engineering
Stanford University
Email: {vinith, tsachy}@stanford.edu

***Abstract*—Unsupervised learning may be modeled by an extreme version of the universal channel coding problem. Suppose a channel decoder knows only that a randomly generated block code is being employed at the other end of a discrete memoryless channel. The channel statistics, the codebook, the code distribution, the rate, the blocklength, and even the input alphabet are unknown. Using a novel decoding measure, it is shown that the channel outputs may be correctly clustered with vanishing error probability at all rates under capacity. Results provide theoretical motivation for certain heuristic clustering techniques and, more widely, suggest that there are gains to be had via non-pairwise similarity measures.**

I. Introduction

In Shannon's random coding scheme for the discrete memoryless channel (DMC), the decoder is supplied with considerable context about both the encoder and the channel. The decoding is informed by the channel statistics, the blocklength employed, the rate of the codebook, and the codewords themselves.

Suppose none of this information is available and the decoder's only knowledge is the following:

- D1 An (unknown) block code C of (unknown) blocklength n and (unknown) rate R is generated i.i.d. according to a (unknown) probability distribution p_X over a (unknown) finite input alphabet.
- D2 Message symbols $I_i \in \{1, \ldots, 2^{nR}\}$ are distributed uniformly and independently.
- D3 The codewords corresponding to the message symbols are transmitted through a DMC with (unknown) transition probabilities $p_{Y|X}$.

How well can the message be decoded, if at all?

This problem setting is an extreme example of universal communication with respect to both the channel statistics and the codebook itself, and therefore represents an added dimension of robustness over traditional formulations [1]–[4]. It may also represent situations such as eavesdropping [5] and "communicating with aliens" [6]. Of greater interest to us, however, is that it serves as a simple and effective information theoretic model for unsupervised learning.

A. Unsupervised Learning via Universal Decoding

In seeing this connection, consider the output distribution of a DMC subject to block coding. If the mutual information across the channel exceeds the rate, the output will naturally cluster into conditionally typical sets for each codeword (Fig. 1). The objective in designing a block channel decoder is effectively to identify and label these clusters as accurately as possible. Note however that the clusters' specific structure and appearance depends on both the DMC and the codebook. The notion of "similarity," in other words, can be far more complex than Euclidean closeness. To be more specific: the universal communication model for clustering is a generative learning model wherein each label (codeword) generates data (channel outputs) through an unknown DMC.

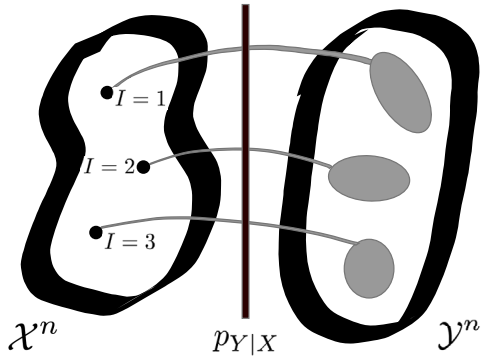

Fig. 1. Channel coding through a DMC is a natural model for clustering.

By leveraging this framework, one may ask two questions: (1) Is there an information-theoretically optimal measure of similarity? (2) What does this say about existing clustering techniques? Our analysis presents the *minimum partition information* as an answer to the first question, and suggests that m-tuple similarity measures could yield considerable gains over the pairwise comparisons commonly used for clustering.

Several attributes of this model (given by **D1–D3**) are worth noting before proceeding further. First, rather than utilizing features, the clustering operation is to be performed on data in its own space, using only the assumption that clusters are generated via DMC. One should expect this universality to come at some cost to performance. Second, observe that codewords are generated randomly and iid. While this is helpful in traditional channel coding both from a proof and implementation standpoint, it is fundamentally necessary in the world of universal decoding [7]. Third, the (unknown) number of clusters is allowed to grow exponentially with data length, and performance is quantified by the maximum decodable rate of growth. Finally, the unknown blocklength n assumption has

978-1-4799-0446-4/13/$31.00 ©2013 IEEE

limited relevance in a model for clustering, where the blocklength can be generally assumed known. It is, however, quite applicable in the original universal communication setting.

The model may also be expanded to incorporate elements of supervised learning [7].

B. Universal Pattern Decoding

The goal of decoding clusters is subtly different from the goal of decoding codewords. To capture this distinction, we introduce the notion of *pattern decoding*.

Consider again the conditions **D1 – D3**. Under this degree of ambiguity no message can be reliably decoded. To see this, note that because the message is uniformly distributed, the output statistics from the channel are unaffected by the choice of bijective map $E : \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n$ between message symbols and codewords. As such, the decoder can neither determine this map nor invert it to obtain the message stream.

However, one could potentially decode into unlabeled clusters corresponding to each message symbol. Equivalently, the *pattern* of the message stream (as introduced by Orlitsky et al. [8]) can be decoded. Rather than asking that the decoder's reconstruction $\widehat{I}^\infty$ resemble the message I^∞, one requests that a symbol-by-symbol relabeling of the reconstruction $(L(\widehat{I}_i))_{i=1}^\infty$ resembles the message. For instance, the patterns of 0012021100 and 2201210022 are identical because there exists a relabeling $\{0, 1, 2\} \to \{2, 0, 1\}$ to transform one into the other. [1]

We describe a universal decoder ϕ_U that pattern-decodes (clusters) the message for any finite source alphabet $\mathcal{X}$, any random codebook distribution p_X, any channel $p_{Y|X}$, and any rate $R < I(X;Y)$, with vanishing error in the limit of large (unknown) blocklength. The decoder is built upon a new similarity measure called the *minimum partition information* (MPI).

Related Work

Previous work into universal channel decoding has focused on varying degrees of uncertainty in the channel. Goppa [1] introduced the maximum mutual information decoder, which was shown to be universal over the class of DMCs. This line of work was later expanded into more general channel models by Ziv [2] and others [4], [9]. See [3] for an overview of these results.

While few information theoretic models have been explicitly designed for statistical learning, certain learning algorithms have certainly taken inspiration from information theoretic quantities or arguments (e.g. [10]). Specifically, mutual information based clustering has been heuristically attempted with applications ranging from congressional votes [11] to gene expression data [12].

This concept is both rigorized and generalized by the elegant Universal Similarity Metric (USM) introduced by Li et al. [13], which has proven successful in contexts as disparate as music clustering and mitochondrial evolutionary history. In this construction, the similarity of two sequences x and y is measured by comparing their marginal and joint Kolmogorov complexities $K(x)$, $K(y)$, and $K(x, y)$. As these quantities are in general noncomputable, complexity is instead approximated by *compressibility*. The resulting (computable) similarity measure is effectively an estimator for the mutual information between x and y.

Unlike pairwise similarity measures such as the mutual information or the USM, the minimum-partition information introduced in our work gauges the similarity of a *group* of m samples. This results in a quantifiable and sizeable asymptotic improvement in performance. Moreover, the formula for minimum partition information immediately suggests a mechanism to similarly boost existing measures — including the USM.

II. Preliminaries

A. Problem Setting

A discrete memoryless channel is specified by transition probabilities $p_{Y|X}(y|x)$ between finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. Let $R \in \mathbb{R}^+$ denote the rate, and $n \in \mathbb{Z}^+$ the blocklength. For every n, 2^{nR} codewords of length n are generated independently and identically (iid) according to a distribution $p_X(x)$ over $\mathcal{X}$. Together, they specify a codebook C_n, which we define as a map $C_n(i)$ from a codeword index $i \in \{1, \ldots, 2^{nR}\}$ to $\mathcal{X}^n$. It is assumed that the codebook itself is asymptotically reliable — i.e. that the rate R is exceeded by the mutual information $I(X;Y)$ across the channel.

The message for transmission is represented as a sequence of indices $I_i, i \in \mathbb{Z}^+$ which are chosen independently and uniformly over $\{1, \ldots, 2^{nR}\}$. The sequence of transmitted codewords is written either as $C_n(I_i)$ or $X_{(i)}^n$, and the corresponding channel outputs as $Y_{(i)}^n$, or in unpartitioned scalar form as Y_j.

Notationally, $I(p(X), p(Y|X))$ is the mutual information from channel $p(Y|X)$ and $X \sim p(X)$. When the distributions corresponding to X and Y are clear, $I(X;Y)$ denotes $I(p_X, p_{Y|X})$. Similarly, $H(p(X)) = H(X)$ is the entropy of distribution $p(X)$. The kth-order empirical entropy of a sequence x^{Nk} is given by $\hat{p}^k[x^{Nk}]\left(\widetilde{x}^k\right) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left(x_{k(i-1)+1}^{ki} = \widetilde{x}^n\right)$ and the corresponding empirical entropy is denoted $\hat{H}^k(x^{Nk}) = H(\hat{p}^k[x^{Nk}])$.

Additionally, we adopt the notation that a vector of vectors can be denoted $\left(x_{(i)}^n\right)_{i=1}^m \equiv \left(x_{(1)}^n, x_{(2)}^n, \ldots, x_{(m)}^n\right)$. When we wish to unwrap with the outer dimension first, we write $T\left(x_{(i)}^n\right)_{i=1}^m \equiv \left(x_{i,(1)}, x_{i,(2)}, \ldots, x_{i,(m)}\right)_{i=1}^n$, where T represents the transpose operation.

B. Performance Evaluation

The error rate of a traditional decoding function $\phi(\cdot) : \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ is a random variable given by

$$e(\phi) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \mathbf{1}(\phi(Y_{(i)}^n) \neq I_i) \quad (1)$$

$$= \lim_{N\to\infty} d_H\left(\left(\phi\left(Y_{(i)}^n\right)\right)_{i=1}^N, I^N\right), \quad (2)$$

[1] Pattern-decoding is essentially a canonical labeling of clusters (by their order of appearance in the channel output) in the absence of the correct labels.

where $d_H(\cdot,\cdot)$ is the normalized Hamming distance between two sequences. Observe that with probability one this limit exists and takes on a single deterministic value.

Any decoding function can also be written as the composition of a *pattern decoder* $\phi(\cdot) : \mathcal{Y}^n \to \{1,\dots,M\}$, which partitions the output space $\mathcal{Y}^n$ into $M \leq 2^{nR}$ arbitrarily labeled regions, and a *label function* $L(\cdot) : \{1,\dots,M\} \to \{1,\dots,2^{nR}\}$, which labels each region with an index from the codebook C_n.

Definition 1: A **universal pattern decoder** ϕ_U implements pattern decoding after training on the channel output Y^∞. Formally, ϕ_U is a collection of several elements:

1) A blocklength estimation function $\widehat{n}(y^\infty)$.
2) An output cardinality function $M(y^\infty) \leq 2^{nR}$.
3) A set of pattern decoders $\{\phi[y^\infty](\cdot) : \mathcal{Y}^{\widehat{n}(y^\infty)} \to \{1,\dots,M(y^\infty)\}\}$, indexed over all possible channel outputs $y^\infty \in \mathcal{Y}^\infty$.

To evaluate the efficacy of a pattern decoder, the *pattern error rate* associated with a universal pattern decoder is a random variable given by

$$e(\phi_U) = \lim_{N\to\infty} d_P\left(\left(\phi[Y^\infty]\left(Y_{(i)}^n\right)\right)_{i=1}^N, I^N\right),$$

where the *pattern distance* $d_P(\cdot,\cdot) : \mathcal{X}^N \times \mathcal{Y}^N \to [0,1]$ between two sequences is defined as smallest Hamming distance possible after an optimized labeling:

$$d_p\left(x^N, y^N\right) = \min_{L:\mathcal{X}\to\mathcal{Y}} d_H\left(\left(L(x_i)\right)_{i=1}^N, y^N\right).$$

Note that $e(\phi_U)$ takes a single deterministic value with probability one (Kolmogorov's 0/1 law).

Definition 2: A universal pattern decoder ϕ_U is said to be *universally reliable* if $e(\phi_U) \to 0$ with probability 1 for any finite source alphabet $\mathcal{X}$, random code distribution p_X, channel $p_{Y|X}$, and rate $R < I(X;Y)$.

C. Minimum Partition Information

The *minimum partition information*, which may be interpreted as a measure of similarity, is at heart of the clustering/pattern-decoding operation. To precisely define this term, several other constructions are first necessary.

A partition P of size $|P|$ for the set $\{1,\dots,m\}$ is a collection of nonintersecting sets $P_i, i \in \{1,\dots,|P|\}$ such that $\cup_{i=1}^{|P|} P_i = \{1,\dots,m\}$.

The set of all size-L partitions is denoted $\mathcal{P}_L^m$, and the set of all partitions is $\mathcal{P}^m$. The *partition function* $P(I_1,\dots,I_m) : (\mathbb{Z}^+)^m \to \mathcal{P}$ partitions the indices $\{1,\dots,m\}$ into equivalence classes according to the equivalence relation $\{i \equiv j \text{ if } I_i = I_j\}$.

Definition 3: The **partition information** for $(y_{(i)}^n)_{i=1}^m \in \mathcal{Y}^{nm}$ and partition $P \in \mathcal{P}^m$ is defined as

$$i_P\left((y_{(i)}^n)_{i=1}^m\right) = \left(\sum_{j=1}^{|P|} \widehat{H}_{P_j}\left(T\left(y_{(i)}^n\right)_{i\in P_j}\right)\right) - \widehat{H}_m\left(T\left(y_{(i)}^n\right)_{i=1}^m\right),$$

where $\widehat{H}_k(T(x_{(i)}^n)_{i=1}^k)$ effectively computes the entropy of the joint type of the sequences $x_{(1)}^n, x_{(2)}^n, \dots, x_{(k)}^n$.

Definition 4: The **minimum partition information (MPI)** for $(y_{(i)}^n)_{i=1}^m$ is

$$\underline{i}\left(\left(y_{(i)}^n\right)_{i=1}^m\right) = \min_{P\in\mathcal{P}^m} \frac{1}{|P|-1} i_P\left(\left(y_{(i)}^n\right)_{i=1}^m\right). \quad (3)$$

MPI may be seen as a generalization of the max mutual information decoding measure (which it reduces to for $m = 2$). If $(\widetilde{Y}_i^n)_i^m$ are each generated from the same source codeword $\widetilde{X}^n$, which itself is generated randomly, then one may demonstrate that with high probability $\underline{i}\left((Y_i^n)_{i=1}^m\right) \approx I(X;Y)$. More importantly, if the source is an impure mix of codewords, large deviations techniques can guarantee that $\underline{i}\left((Y_i^n)_{i=1}^m\right)$ is small.

D. Statement of Result

Theorem 1: There exists a universal pattern decoder ϕ_U that is universally reliable.

The remainder of this paper constructs ϕ_U and bounds its performance. Full proofs of lemmas are relegated to [7] for space considerations.

III. Scheme Description and Performance

Rather than directly specifying $\phi[Y^\infty](\cdot)$ as a function of Y^∞, the construction is performed in stages.

First, define the *tupling parameter* $m(n) = \left\lfloor \log\left(\frac{n}{\log n}\right)\right\rfloor$. This will be frequently be referred to simply as m, with the dependency on n implicit.

A blocklength estimator $\hat{n}(Y^\infty)$ is defined in Sec. III-A, and a rate estimator $\hat{R}_{\hat{n}}(Y^\infty)$ is defined in Sec III-B. In Sections III-C and III-D a decoding function $\phi[\hat{n}, \widehat{R}, Y^{\hat{n}2^{\hat{n}\hat{R}(m(\hat{n})-1)}}](\cdot)$ is described. The universal pattern decoder is then specified as

$$\phi[Y^\infty](\cdot) = \phi[\hat{n}, \widehat{R}_{\hat{n}}, Y^{\hat{n}2^{\hat{n}\hat{R}_{\hat{n}}(m(\hat{n})-1)}}](\cdot).$$

Three performance guarantees are provided:

1) In Sec. III-A, it is shown that $\mathrm{P}\left(\hat{n}(Y^\infty) = n\right)$ goes to 1 with n.
2) Sec. III-B establishes that with probability approaching 1, the rate estimate $\widehat{R}_n(Y^\infty)$ falls into an acceptable range.
3) Secs. III-C and III-D prove that for a suitably accurate rate estimate $\widehat{R}_n$, the pattern error goes to zero with probability 1. i.e.:

$$d_p\left(\left(\phi[n, \widehat{R}_n, Y^{\hat{n}2^{\hat{n}\hat{R}_{\hat{n}}(m(\hat{n})-1)}}](Y_{(i)}^n)\right)_{i=1}^\infty, I^\infty\right) \underset{n\to\infty}{\to} 0,$$

with probability 1.

These three results prove Theorem 1.

A. Blocklength Estimation

Intuitively, the blocklength estimator finds the minimum blocklength at which the blocks appear iid. Let $\tilde{n}$ be a hypothesized blocklength. The empirical mutual information between

adjacent $\tilde{n}$-blocks is denoted by $\widehat{I}_{\tilde{n}}[y^\infty] = I(\hat{p}^{2\tilde{n}}[y^\infty], p_{Y|X})$. The blocklength estimator is then given by

$$\widehat{n}(y^\infty) = \min_{\tilde{n}:\widehat{I}_{\tilde{n}k}=0, \forall k \in \mathbb{Z}^+} \tilde{n}.$$

It is not difficult to show that $\hat{n}(y^\infty)$ is equal to n with high probability.

Lemma 2: $\mathrm{P}\left(\widehat{n}(y^\infty) = n\right)$ goes to 1 as n grows.

B. Rate Estimation

An overview of the rate estimation operation is presented, along with a sketch of certain elements in the proof.

Conditioned on a successful estimation of the blocklength n, the nm-th order empirical distribution of the channel output $\hat{p}^{nm}(Y^\infty)[(y^n_{(i)})^m_{i=1}]$ may be computed. This is helpful in estimating the rate. In particular, one may compute the empirical cumulative distribution of the MPI of neighboring outputs:

$$\hat{P}^m_{\underline{i}}[Y^\infty](\theta) = \hat{p}^{nm}[Y^\infty]\left(\left\{\left(y^n_{(i)}\right)^m_{i=1} : \underline{i}\left(\left(y^n_{(i)}\right)^m_{i=1}\right) \geq \theta\right\}\right).$$

This is the empirical probability that m randomly chosen output blocks have MPI exceeding θ. The following lemma demonstrates that the limiting exponent $\alpha(\theta)$ of this quantity undergoes a transition at $\theta = R$:

Lemma 3: With probability one,

$$\alpha(\theta) = \lim_{n\to\infty} \frac{-\log \hat{P}^m_{\underline{i}}[Y^\infty](\theta)}{n(m-1)} \geq \theta \text{ for } 0 < \theta < R$$
$$\leq R \text{ for } R < \theta < I(X;Y).$$

From observing a graphical depiction of the transition (Fig. 2), it is tempting to simply set the rate estimate $\hat{R} = \min\{\theta : \theta > \frac{-1}{n(m-1)} \log \hat{P}^m_{\underline{i}}[Y^\infty](\theta)\}$ and assume that this yields the corner point. However, the transition in $\frac{-1}{n(m-1)} \log \hat{P}^m_{\underline{i}}[Y^\infty](\theta)$ is only predicted to occur in the limit with n, and as there is no guarantee that this limit is approached uniformly for all $\theta \in [0, I(X;Y)]$, there is no way to bound $|\hat{R} - R|$.

As a means around this, consider quantizing the range of θ into a finite set:

$$\widehat{R}_n = \Delta \min\left\{k \in \mathbb{Z}^+ : k\Delta > \frac{-\log \hat{P}^m_{\underline{i}}[Y^\infty](k\Delta)}{n(m-1)}\right\}. \quad (4)$$

The finite set of threshold values $\left\{\frac{-\log \hat{P}^m_{\underline{i}}[Y^\infty](k\Delta_n)}{n(m-1)}\right\}$ now converges uniformly. Furthermore, if $\Delta < R - I(X;Y)$, then there is guaranteed to be at least one quantization point $k\Delta$ between R and $I(X;Y)$.

To determine a step size Δ that satisfies these two criteria, an estimate of the gap $I(X;Y) - R$ is first computed:

$$\overline{(I(X;Y) - R)}_n = \widehat{H}^1(Y^n_{(1)}) + \frac{1}{n}\log\left(\hat{p}^n[Y^\infty](Y^n_{(1)})\right) \quad (5)$$

The step size is then estimated as

$$\Delta_n = \left\lfloor \left[\frac{1}{4}\overline{(I(X;Y) - R)}_n\right]^{-1} \right\rfloor^{-1}. \quad (6)$$

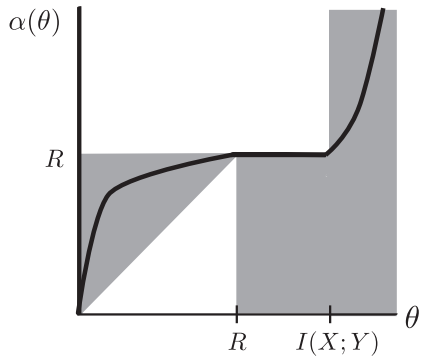

Fig. 2. The limiting MPI exponent, $\alpha(\theta)$, is confined to the grey regions by Lemma 3. Although the function drawn here is continuous, there is no such requirement on $\alpha(\theta)$.

Remarks:

1) The intuition behind the estimate $\overline{(I(X;Y) - R)}_n$ lies in that $(X^n_{(1)}, Y^n_{(1)})$ are likely to be strongly typical. By evaluating the empirical entropy of a strongly typical output, the first term estimates $H(Y)$. By evaluating the probability mass at a strongly typical output, the second term estimates $R + H(Y|X)$.
2) If $\overline{(I(X;Y) - R)}_n$ is a good estimate, then the bin size ensures that quantizer points lie between R and $I(X;Y)$.
3) The floor operation, along with the convergence of $\overline{(I(X;Y) - R)}_n$, guarantees that Δ_n is a deterministic value Δ for sufficiently large n. As a result the threshold values in (4) converge uniformly with n to $\{\alpha(k\Delta)\}$.

The convergence of the rate estimate is captured by the following lemma:

Lemma 4: With probability going to 1 as $n \to \infty$,

$$0 \leq \hat{R}_n - R \leq \frac{1}{4}(I(X;Y) - R). \quad (7)$$

C. Decoding Rule

Traditional channel decoding is performed with respect to a codebook of input codewords. In the case of pattern decoding, the input codebook is not available, but one may still construct a codebook of carefully chosen collections of *outputs* $\{\{Y^n_{(i)}\}_{i\in DC(j)}\}_j$.

Definition 5: A *dirty codebook* $DC(\cdot)$ of size $|DC|$ is a map from $\{1, \ldots, |DC|\}$ into $\mathcal{Y}^n_{(1)} \times \ldots \times \mathcal{Y}^n_{(m-1)}$. A *dirty codeword* is a particular output $DC(i)$, and the $m-1$ sequences $(y^n_{(1)}, \ldots, y^n_{(m-1)})$ that comprise $DC(i)$ are called *dirty entries*.

The canonical unfiltered dirty codebook, for instance, is defined as the first outputs from the channel:

$$UDC_n(i) = \left(Y^n_{(i-1)(m-1)+1}, \ldots, Y^n_{i(m-1)}\right), \quad (8)$$

where $i \in \{1, \ldots, n2^{n\widehat{R}_n(m-1)}\}$.

The universal decoding rule $\phi_{DC}(\cdot) : \mathcal{Y}^n \to 2^{\{1,\ldots,|DC|\}}$ for a dirty codebook DC is given by

$$\phi_{DC}(y^n) = \left\{i : \underline{i}\left((DC(i), y^n)\right) \geq \widehat{R}_n + \frac{\overline{(I(X;Y) - R)}_n}{2}\right\}$$

Remarks:

1) The universal decoding rule does not necessarily decode uniquely. For UDC_n, for instance, one expects a typical output cardinality $|\phi_{UDC_n}(\cdot)|$ to be of the order $n2^{n(\widehat{R}-R)}$.
2) Intuitively, if y^n and the dirty entries of $DC(i)$ are all generated by the same source codeword $C_n(j)$, one expects $\underline{i}\,((DC_i, y^n))$ to approach $I(X;Y)$, which exceeds the threshold.
3) Conversely, if either y^n or one of the dirty entries of $DC(i)$ does not share a source codeword with the others, $\underline{i}((DC_i, y^n))$ is likely to be small.

To formalize some of these notions, consider the application of the universal decoding rule to UDC_n. Define the no-error decoding event $F(I, y^n)$ as occurring when every dirty entry of every dirty codeword in $\phi_{UDC_n}(y^n)$ has index I; i.e.

$$F(I, y^n) = \prod_{i \in \phi_{UDC_n}(y^n)} \prod_{j=(i-1)(m-1)+1}^{i(m-1)} \mathbf{1}(I = I_j).$$

The no-error rate is then given by

$$e_F = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} F(I_i, Y_{(i)}^n).$$

Lemma 5: With probability 1, $e_F \to 1$ as $n \to \infty$.

D. Decoding Codebook

The decoding scheme of the previous section has a vanishing probability of false positives, but there is no guarantee of a true positive and the output space $\{1, \ldots, |UDC_n|\}$ has far more clusters than the 2^{nR} maximum cardinality for a pattern decoder. To fix these issues, a new dirty codebook is constructed.

Notationally, we say a dirty codebook is contained by another $DC^{(1)} \subset DC^{(2)}$ if the dirty codewords of one are a subset of the other: $DC^{(1)}(\{1, \ldots, |DC^{(1)}|\}) \subset DC^{(2)}(\{1, \ldots, |DC^{(2)}|\})$. Additionally, define the uniqueness rate as

$$e_U(DC) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(|\phi_{DC}(Y_{(i)}^n)| = 1\right).$$

Definition 6: The filtered dirty codebook FDC_n is the smallest subset of UDC_n that decodes sufficiently uniquely:

$$FDC_n = \operatorname*{argmin}_{DC} \{|DC| : DC \subset UDC_n, e_U(DC) \geq 1 - \frac{1}{n}\}.$$

Lemma 6: With probability going to 1 as $n \to \infty$, $\{|DC| : DC \subset UDC_n, e_U(DC) \geq 1 - \frac{1}{n}\}$ is nonempty and $|FDC_n| \leq 2^{nR}$.

A decoding function $\phi_n^*(\mathcal{Y}^n)$ can therefore be constructed from the universal decoding rule ϕ_{FDC_n}:

$$\phi_n^*(y^n) = \begin{cases} \phi_{FDC_n}(y^n) & \text{if } |\phi_{FDC_n}(y^n)| = 1 \\ \text{error} & \text{otherwise.} \end{cases}$$

Observe that $\phi_n^*(y^n)$ is a function mapping from $\mathcal{Y}^n$ into $\{1, \ldots, |FDC|_n, \text{ error}\}$. If $|FDC|_n \leq 2^{nR}$ — which by Lemma 6 holds with probability approaching 1 — then it is a valid pattern decoding function. We may therefore complete specification of ϕ_U by setting

$$\phi[n, \widehat{R}_n, Y^{n2^{n\hat{R}_n(m(n)-1)}}](\cdot) = \phi_n^*(\cdot),$$

IV. Remarks

An information theoretic model has been presented for both universal communication and unsupervised learning. By constructing a universal decoder for this setting, we demonstrate both robustness to encoder actions and the ability to reliably cluster at optimal rates; that is, anything up to the information $I(X;Y)$ across the channel. This is made possible by a new m-tuple similarity measure called the MPI, whose existence both theoretically validates certain existing clustering techniques and suggests an avenue for improvement.

Specifically, to compare the performance of MPI to (pairwise) mutual information, the tupling parameter $m(n)$ may be fixed to a constant value of 2 (for which MPI reduces to mutual information). By propagating this change through the analysis of Sec. III-C, one finds that this pairwise clustering operation can only reliably cluster at rates up to the information of a repeated transmission across the channel $I(Y_1, Y_2)$ [7]. [2] This strongly suggests that there may be benefits to clustering with an m-tuple similarity measure.

References

[1] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Cont. Information Theory*, vol. 4, pp. 97–102, 1975.

[2] J. Ziv, "Universal decoding for finite-state channels," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 453–460, 1985.

[3] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2148 –2177, oct 1998.

[4] A. Lapidoth and J. Ziv, "On the universality of the lz-based decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 44, no. 5, pp. 1746 –1755, sep 1998.

[5] A. Wyner, "The wire-tap channel," *Bell Sys. Tech. J.*, vol. 54, pp. 1355–1387, 1974.

[6] B. Juba and M. Sudan, "Universal semantic communication i," in *Proc. of the 40th Annual ACM Symp. on Theory of Computing*, 2008, pp. 123–132.

[7] V. Misra and T. Weissman, "Unsupervised learning and universal commmunication," http://www.stanford.edu/~vinith/PatternDecoding.pdf, in preparation.

[8] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *Information Theory, IEEE Transactions on*, vol. 50, no. 7, pp. 1469 – 1481, july 2004.

[9] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *Information Theory, IEEE Transactions on*, vol. 44, no. 5, pp. 1726 –1745, sep 1998.

[10] D. Erdogmus and J. Principe, "Information theoretic learning," in *Encyclopedia of Artificial Intelligence*, Dopico, Dorado, and Pazos, Eds. IGI Global, 2008.

[11] S. Deng, Z. He, and X. Xu, "G-anmi: A mutual information based genetic clustering algorithm for categorical data," *Knowl.-Based Syst.*, vol. 23, no. 2, pp. 144–149, 2010.

[12] X. Zhou, X. Wang, E. R. Dougherty, D. Russ, and E. Suh, "Gene clustering based on clusterwide mutual information," *Journal of Computational Biology*, vol. 11, no. 1, pp. 147–161, 2004.

[13] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The similarity metric," *Information Theory, IEEE Transactions on*, vol. 50, no. 12, pp. 3250–3264, 2004.

[2] Both operational and theoretic arguments demonstrate this is strictly less than $I(X;Y)$ for all nontrivial channels.