Universal Decoding for Arbitrary Channels Relative to a Given Class of Decoding Metrics

Author(s):
Merhav, Neri

Publication date:
2014

Permanent link:
https://doi.org/10.3929/ethz-a-010094632

International Zurich Seminar on Communications (IZS), February 26 – 28, 2014

Universal Decoding for Arbitrary Channels Relative to a Given Class of Decoding Metrics

Neri Merhav
Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
Email: merhav@ee.technion.ac.il

***Abstract*—We consider the problem of universal decoding for arbitrary, finite–alphabet unknown channels in the random coding regime. For a given random coding distribution and a given class of metric decoders, we propose a generic universal decoder whose average error probability is, within a sub–exponential multiplicative factor, no larger than that of the best decoder within this class of decoders. Since the optimal, maximum likelihood (ML) decoder of the underlying channel is not necessarily assumed to belong to the given class of decoders, this setting suggests a common generalized framework for: (i) mismatched decoding, (ii) universal decoding for a given family of channels, and (iii) universal coding and decoding for deterministic channels using the individual–sequence approach. The proof of our universality result is fairly simple, and it is demonstrated how some earlier results on universal decoding are obtained as special cases. We also demonstrate how our method extends to more complicated scenarios, like incorporation of noiseless feedback, the multiple access channel, and continuous alphabet channels.**

I. Introduction

The topic of universal coding and decoding under channel uncertainty has received very much attention in the last four decades. In [5], Goppa offered the *maximum mutual information* (MMI) decoder, which decides in favor of the codeword having the maximum empirical mutual information with the channel output sequence. Goppa showed that for discrete memoryless channels (DMC's), MMI decoding achieves capacity. Csiszár and Körner [2] have shown that the random coding error exponent of the MMI decoder achieves the optimum random coding error exponent. Csiszár [1] proved that for any modulo–additive DMC and the uniform random coding distribution over linear codes, the optimum random coding error exponent is universally achieved by a decoder that minimizes the empirical entropy of the difference between the output sequence and the input sequence. In [10] an analogous result was derived for a certain parametric class of memoryless Gaussian channels with an unknown interference signal.

For channels with memory, Ziv [16] studied universal decoding problem for finite–alphabe unifilar finite–state channels. For uniform random coding over a given set, he proved that a decoder based on the Lempel–Ziv algorithm asymptotically achieves the error exponent associated with ML decoding. In [6], Lapidoth and Ziv have extended this result to non–unifilar finite–state channels. In [3], Feder and Lapidoth furnished sufficient conditions for families of channels with memory to have universal decoders in the random coding error exponent sense. In [4], Feder and Merhav proposed a competitive minimax criterion, according to which, an optimum decoder is sought in the quest for minimizing the worst–case regret in the error exponent sense.

More recently, interesting attempts (see, e.g., [8], [9], [12], [14]) were made to devise coding and decoding strategies that avoid any probabilistic assumptions concerning the channel. In [8], the notion of *empirical rate functions* has been established and investigated for a given input distribution and a given posterior probability function of the channel input sequence given the output sequence. In [12], porosity–achieving universal encoders and decoders were devised for modulo additive channels with deterministic noise sequences and noiseless feedback.

In this paper, we take a different approach. We consider universal decoding for *arbitrary* unknown channels in the random coding regime. For a given random coding distribution and a given class of metric decoders, we propose a generic universal decoder whose average error probability is exponentially no larger than that of the best decoder in this class. The proof of our universality result is fairly simple and general, and it is demonstrated how some earlier mentioned results on universal decoding are obtained as special cases.

Finally, we demonstrate how our method extends to more complicated scenarios. The first extension corresponds to incorporation of noiseless feedback. This extension is fairly straightforward, but its main importance is in allowing adaptation of the random coding distribution to the channel statistical characteristics. The second extension is to the problem of universal decoding for multiple access channels (MAC's) with respect to a given class of decoding metrics. This extension is not trivial since the universal decoding metric has to confront three different types of error events. In particular, it turns out that the resulting universal decoding metric is surprisingly different from those of earlier works on universal decoding for the MAC [7], [3, Section VIII], [13], mostly because the problem setting here is different from those of these earlier works (in the sense that the universality here is relative to a given class of decoders while the underlying channel is arbitrary, and not relative to a given class of channels). A third possible extension, that refers to the continuous alphabet case, is discussed briefly along with an example.

II. Notation Conventions

Scalar random variables (RV's) will be denoted by capital letters, their sample values are denoted by the respective lower case letters, and their alphabets are denoted by the respective

calligraphic letters. A similar convention applies to random vectors of dimension n and their sample values, which will be denoted with the same symbols in the bold face font. The set of all n–vectors with components taking values in a certain alphabet, will be denoted as the same alphabet superscripted by n. Channels and sources will be denoted generically by the letter P and Q, respectively. For example, the channel input probability distribution function will be denoted by $Q(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^n$, and the conditional probability distribution of the channel output vector $\boldsymbol{y} \in \mathcal{Y}^n$ given the input vector $\boldsymbol{x} \in \mathcal{X}^n$, will be denoted by $P(\boldsymbol{y}|\boldsymbol{x})$. Information theoretic quantities like entropies and conditional entropies, will be denoted following the standard conventions of the information theory literature, e.g., $H(\boldsymbol{X})$, $H(\boldsymbol{X}|\boldsymbol{Y})$, etc. The expectation operator will be denoted by $\boldsymbol{E}\{\cdot\}$ and the cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

For a given sequence $\boldsymbol{x} \in \mathcal{X}^n$, $\mathcal{X}$ being a finite alphabet, $\hat{P}_{\boldsymbol{x}}$ denotes the empirical distribution on $\mathcal{X}$ extracted from $\boldsymbol{x}$, in other words, $\hat{P}_{\boldsymbol{x}}$ is the vector $\{\hat{P}_{\boldsymbol{x}}(x),\ x \in \mathcal{X}\}$, where $\hat{P}_{\boldsymbol{x}}(x)$ is the relative frequency of the letter x in the vector $\boldsymbol{x}$. The type class of $\boldsymbol{x}$, denoted $T_{\boldsymbol{x}}$, is the set of all sequences $\boldsymbol{x}' \in \mathcal{X}^n$ with $\hat{P}_{\boldsymbol{x}'} = \hat{P}_{\boldsymbol{x}}$. Similarly, for a pair of sequences $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$, the empirical distribution $\hat{P}_{\boldsymbol{xy}}$ is the matrix of relative frequencies $\{\hat{P}_{\boldsymbol{xy}}(x,y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ and the type class $T_{\boldsymbol{xy}}$ is the set of pairs $(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{X}^n \times \mathcal{Y}^n$ with $\hat{P}_{\boldsymbol{x}'\boldsymbol{y}'} = \hat{P}_{\boldsymbol{xy}}$. For a given $\boldsymbol{y}$, $T_{\boldsymbol{x}|\boldsymbol{y}}$ denotes the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, which is the set of vectors $\{\boldsymbol{x}'\}$ such that $(\boldsymbol{x}', \boldsymbol{y}) \in T_{\boldsymbol{xy}}$. Information measures induced by empirical distributions, i.e., empirical information measures, will be denoted with a hat and a subscript that indicates the sequence(s) from which they are induced. For example, $\hat{H}_{\boldsymbol{x}}(X)$ is the empirical entropy extracted from $\boldsymbol{x} \in \mathcal{X}^n$, namely, the entropy of a random variable X whose distribution is $\hat{P}_{\boldsymbol{x}}$. Similarly, $\hat{H}_{\boldsymbol{xy}}(X|Y)$ and $\hat{I}_{\boldsymbol{xy}}(X;Y)$ are, respectively, the empirical conditional entropy of X given Y, and the empirical mutual information between X and Y, extracted from $(\boldsymbol{x}, \boldsymbol{y})$, and so on. For two sequences of positive numbers, $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Similarly, $a_n \stackrel{\cdot}{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n}\log\frac{a_n}{b_n} \leq 0$, and so on. The functions $\log(\cdot)$ and $\exp(\cdot)$, throughout this paper, will be defined to the base 2, unless otherwise indicated. The operation $[\cdot]_+$ will mean positive clipping, that is $[x]_+ = \max\{0, x\}$.

III. Problem Formulation

Consider a random selection of a codebook $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\} \subseteq \mathcal{X}^n$, where $M = 2^{nR}$, R being the coding rate in bits per channel use. The marginal probability distribution function of each codeword $\boldsymbol{x}_i$ is denoted by $Q(\boldsymbol{x}_i)$. It will be assumed that the various codewords are *conditionally pairwise independent*.[1] Let $P(\boldsymbol{y}|\boldsymbol{x})$ be the conditional probability distribution of the channel output vector $\boldsymbol{y} \in \mathcal{Y}^n$ given the channel input vector $\boldsymbol{x} \in \mathcal{X}^n$. We make no assumptions at all concerning the channel. We will assume that both the channel input alphabet $\mathcal{X}$ and the channel output alphabet $\mathcal{Y}$ are finite. Finally, we define a class of *decoding metrics*, as a class of real functions, $\mathcal{M} = \{m_\theta(\boldsymbol{x}, \boldsymbol{y}),\ \theta \in \Theta,\ \boldsymbol{x} \in \mathcal{X}^n,\ \boldsymbol{y} \in \mathcal{Y}^n\}$, where Θ is an index set. The decoder associated with m_θ, which will be denoted by $\mathcal{D}_\theta$, decides in favor of the message $i \in \{1, \ldots, M\}$ which maximizes $m_\theta(\boldsymbol{x}_i, \boldsymbol{y})$. The message i is assumed to be uniformly distributed over $\{1, 2, \ldots, M\}$. It should be emphasized that the ML decoding metric for the underlying channel $P(\boldsymbol{y}|\boldsymbol{x})$, may not necessarily belong to $\mathcal{M}$. The average error probability associated with $\mathcal{D}_\theta$, is denoted by $\bar{P}_{e,\theta}(R, n)$. While the decoder $\mathcal{D}_\theta$, that minimizes $\bar{P}_{e,\theta}(R, n)$ within $\mathcal{M}$, depends on the unknown underlying channel, our goal is to devise a universal decoder $\mathcal{U}$, with a decoding metric $U(\boldsymbol{x}, \boldsymbol{y})$, independent of the underlying channel $P(\boldsymbol{y}|\boldsymbol{x})$, whose average error probability would be essentially as small as $\min_\theta \bar{P}_{e,\theta}(R, n)$, whatever the underlying channel may be. By "essentially as small", we mean that the average error probability associated with the universal decoder, denoted $\bar{P}_{e,u}(R, n)$, would not exceed $\min_\theta \bar{P}_{e,\theta}(R, n)$ in the exponential sense.

[1] By "conditionally pairwise independent", we mean that for every three randomly chosen codewords, we have that any two of them are conditionally independent given the third one.

IV. Main Result

Let us define

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) \stackrel{\Delta}{=} \{\boldsymbol{x}' :\ \ \forall\theta \in \Theta\ \ m_\theta(\boldsymbol{x}', \boldsymbol{y}) = m_\theta(\boldsymbol{x}, \boldsymbol{y})\}\,. \quad (1)$$

Our universal decoding metric is defined as

$$U(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\Delta}{=} -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]. \quad (2)$$

Clearly, $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ are equivalence classes for every $\boldsymbol{y} \in \mathcal{Y}^n$, and so $\mathcal{X}^n$ can be partitioned into a disjoint union of them. Let $K_n(\boldsymbol{y})$ denote the number of equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ for a given $\boldsymbol{y}$. Also define $K_n \stackrel{\Delta}{=} \max_{\boldsymbol{y}\in\mathcal{Y}^n} K_n(\boldsymbol{y})$ and $\Delta_n \stackrel{\Delta}{=} \frac{\log K_n}{n}$. Our main result is the following theorem (the proof appears in the full version on the paper [11]):

Theorem 1: Under the above assumptions, the universal decoding metric defined in eq. (2) satisfies:

$$\bar{P}_{e,u}(R, n) \leq 2 \cdot 2^{n\Delta_n} \cdot \min_{\theta\in\Theta} \bar{P}_{e,\theta}(R, n). \quad (3)$$

The theorem is meaningful when $\Delta_n \to 0$ as $n \to \infty$, which means that the number of various equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ grows sub-exponentially, uniformly in $\boldsymbol{y}$. In this case, whenever $\min_{\theta\in\Theta} \bar{P}_{e,\theta}(R, n)$ decays exponentially with n, then so does $\bar{P}_{e,u}(R, n)$, and at least as fast. Therefore, a sufficient condition for the existence of a universal decoder is $\lim_{n\to\infty} \Delta_n = 0$. The behavior of Δ_n for large n is a measure of the richness of $\mathcal{M}$. The larger is $\mathcal{M}$, the smaller are the equivalence classes, and then their total number becomes larger, and so does Δ_n. Universality is enabled, using this method, as long as the set Θ is not too rich, so that Δ_n still vanishes as n grows. When Q is invariant within $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ (i.e., $\boldsymbol{x}' \in \mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ implies $Q(\boldsymbol{x}') = Q(\boldsymbol{x})$), we have $U(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}) + \log|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|]$. The choice of Q that is invariant within $T(\boldsymbol{x}|\boldsymbol{y})$ is convenient, because it is easier to evaluate the log–cardinality of $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ than to evaluate its probability under Q.

Example 1. Let Q be the uniform distribution across a single type class, $T_{\boldsymbol{x}}$, and let $\mathcal{M}$ be the class of additive decoding metrics $m_\theta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n \theta(x_i, y_i)$, where $\{\theta(x, y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ are arbitrary real–valued matrices. Here, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) =$

$T_{\boldsymbol{x}|\boldsymbol{y}}$, the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$. Since the number of conditional type classes is polynomial in n, then $\Delta_n \to 0$. In this case, $U(\boldsymbol{x},\boldsymbol{y}) = \hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) + o(n)$. and so, the proposed universal decoder essentially coincides with the MMI decoder. If, on the other hand, $Q(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$, then $U(\boldsymbol{x},\boldsymbol{y}) = \hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) + D(\hat{P}_{\boldsymbol{x}}\|Q) + o(n)$, where $D(\hat{P}_{\boldsymbol{x}}\|Q)$ is the divergence between $\hat{P}_{\boldsymbol{x}}$ and Q. This concludes Example 1.

One of the elegant points in [16] is that the universality of the proposed decoding metric is proved without recourse to an explicit derivation of the random coding error exponent of the optimum decoder. The proof of Theorem 1 has the same feature. However, thanks to Shulman's lower bound [15] on the probability of a union of events, this proof is both simpler and more general in several respects: (i) it allows a general Q, not just the uniform distribution, (ii) it requires only conditionally pairwise independence between codewords, not mutual independences, and (iii) it assumes nothing concerning the underlying channel. Indeed, it will be seen shortly how Ziv's universal decoding metric is obtained as a special case.

In some situations, it may not be a trivial task to evaluate $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]$. Suppose, however, that one can uniformly lower bound $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})] = \exp\{-nU(\boldsymbol{x},\boldsymbol{y})\}$ by $\exp\{-nU'(\boldsymbol{x},\boldsymbol{y})\}$, for some function $U'(\boldsymbol{x},\boldsymbol{y})$ which is computable and suppose that $U'(\cdot,\cdot)$ satisfies

$$\max_{\boldsymbol{y}\in\mathcal{Y}^n} \sum_{\boldsymbol{x}\in\mathcal{X}^n} Q(\boldsymbol{x}) 2^{nU'(\boldsymbol{x},\boldsymbol{y})} \le 2^{n\Delta'_n} \tag{4}$$

where $\Delta'_n \to 0$. We argue (see [11] for a proof) that in such a case, $U'(\cdot,\cdot)$ can replace $U(\cdot,\cdot)$ as a universal decoding metric and Theorem 1 remains valid. The price of passing from U to U' might be in a slowdown of the convergence of Δ'_n vs. Δ_n. For example, U' might correspond to more refined equivalence classes $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$.

Example 2. As an example of the usefulness of this observation, let us refer to Ziv's universal decoding metric [16]. In particular, let $\mathcal{M}$ be the class of decoding metrics defined as follows: For a given $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$, let $\boldsymbol{s} = (s_1,\ldots,s_n) \in \mathcal{S}^n$ ($\mathcal{S}$ being a finite set), be a sequence generated recursively according to $s_{i+1} = g(x_i,y_i,s_i)$, $i = 1,\ldots,n-1$, where s_1 is a fixed initial state and $g : \mathcal{X}\times\mathcal{Y}\times\mathcal{S} \to \mathcal{S}$ is a given next–state function. Now define $m_\theta(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^n \theta(x_i,y_i,s_i)$. Suppose that Q is the uniform distribution over $\mathcal{X}^n$. Then $Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})]$ is proportional to $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$, but the problem is that here, there is no apparent single–letter expression for the exponential growth rate of $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$ in general. Fortunately enough, however, $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|$, in this case, can be lower bounded [16, Lemma 1] by $|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})| \ge 2^{LZ(\boldsymbol{x}|\boldsymbol{y})-no(n)}$, where $LZ(\boldsymbol{x}|\boldsymbol{y})$ denotes the length of the conditional Lempel–Ziv code of $\boldsymbol{x}$ when $\boldsymbol{y}$ is given as side information at both encoder and decoder. Consequently, one can upper bound $U(\boldsymbol{x},\boldsymbol{y})$ by $U'(\boldsymbol{x},\boldsymbol{y}) = \log|\mathcal{X}| - \frac{LZ(\boldsymbol{x}|\boldsymbol{y})}{n} + o(n)$ as our decoding metric. Indeed, eq. (4) is satisfied by this choice of U'. This explains why Ziv's decoder, which selects the message i with the minimum of $LZ(\boldsymbol{x}_i|\boldsymbol{y})$, is universally asymptotically optimum in the random coding exponent sense. Note that the assumption that Q is uniform is not really essential here. In fact, Q can also be any exchangeable probability distribution. Moreover, if s_i includes a component, say, σ_i, that is fed merely by $\{x_i\}$ (but not $\{y_i\}$), then it is enough that Q would be invariant within conditional types of $\boldsymbol{x}$ given $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_n)$. In such a case, we would have $U'(\boldsymbol{x}.\boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}) + LZ(\boldsymbol{x}|\boldsymbol{y})]$.

V. Extensions

We now demonstrate how our method extends to more involved scenarios.

A. Feedback

In the paradigm of random coding in the presence of feedback, it is convenient to think of an independent random selection of symbols of $\mathcal{X}$ along a tree whose branches are labeled by $\{y_1\},\{y_1,y_2\},\ldots,\{y_1,\ldots,y_{n-1}\}$, for all possible outcomes of these vectors. Accordingly, the random coding distribution $Q(\boldsymbol{x})$ is replaced by $Q(\boldsymbol{x}|\boldsymbol{y}) \stackrel{\Delta}{=} \prod_{i=1}^n Q(x_i|x^{i-1},y^{i-1})$. Each message $i \in \{1,2,\ldots,M\}$ is represented by a complete tree of depth n and $|\mathcal{Y}|^{n-1}$ leaves. Theorem 1 and its proof remain intact with $Q(\cdot)$ being replaced by $Q(\cdot|\boldsymbol{y})$ in all places. Thus, the universal decoding metric is redefined as $U(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{n}\log Q[\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|\boldsymbol{y}]$, the relevant expectations are redefined w.r.t. $P(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^n [Q(x_i|x^{i-1},y^{i-1})P(y_i|x^i,y^{i-1})]$, and in condition (4), $Q(\boldsymbol{x})$ is replaced by $Q(\boldsymbol{x}|\boldsymbol{y})$. One might limit the structure of the feedback, for example, by letting each $Q(\cdot|x^{i-1},y^{i-1})$ depend on (x^{i-1},y^{i-1}) only via a state variable t_i fed by these two sequences, i.e., $t_i = g(t_{i-1},x_{i-1},y_{i-1})$, that is

$$Q(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^n Q(x_i|x^{i-1},y^{i-1}) = \prod_{i=1}^n Q(x_i|t_i). \tag{5}$$

In the above example of decoding metrics corresponding to finite–state channels, one can refine the equivalence classes to include the information about t_i and then Q would be invariant within a type class $T_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s},\boldsymbol{t}}$, where $\boldsymbol{t} = (t_1,\ldots,t_n)$. In this case, the decoding metric U' would become $U'(\boldsymbol{x},\boldsymbol{y}) = -\frac{1}{n}[\log Q(\boldsymbol{x}|\boldsymbol{y}) + LZ(\boldsymbol{x}|\boldsymbol{y})]$, where $Q(\boldsymbol{x}|\boldsymbol{y})$ is understood to be defined by (5).

B. The Multiple Access Channel

Consider a MAC with two inputs, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and one output $\boldsymbol{y}$. User no. i generates $M_i = 2^{nR_i}$ mutually independent codewords, $\boldsymbol{x}_i(1),\ldots,\boldsymbol{x}_i(M_i)$, using a random coding distribution Q_i, $i = 1,2$. We define a class $\mathcal{M} = \{m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}),\ \theta\in\Theta\}$. Decoder $\mathcal{D}_\theta$ picks the pair of messages $(\boldsymbol{x}_1(i),\boldsymbol{x}_2(j))$, maximizes $m_\theta(\boldsymbol{x}_1(i),\boldsymbol{x}_2(j),\boldsymbol{y})$. We assume that the random coding ensemble and the class of decoders are such that for every given $\boldsymbol{X}_1(i) = \boldsymbol{x}_1(i), \boldsymbol{X}_2(j) = \boldsymbol{x}_2(j)$ and $\boldsymbol{Y} = \boldsymbol{y}$, $m_\theta(\boldsymbol{X}_1(i'),\boldsymbol{X}_2(j'),\boldsymbol{y})$ and $m_\theta(\boldsymbol{X}_1(i''),\boldsymbol{X}_2(j''),\boldsymbol{y})$ are conditionally independent whenever $(i',j') \ne (i,j)$, $(i'',j'') \ne (i,j)$ and $(i'',j'') \ne (i',j')$. While this requirement is easily satisfied when $i' \ne i$, $i'' \ne i$, $i'' \ne i'$, $j' \ne j$, $j'' \ne j$, and $j'' \ne j'$ all hold (as all codewords are assumed to be drawn by independent random selection), it is less obvious when some of these indices concide. Still, this requirement is satisfied, for example, if $\mathcal{X}_1 = \mathcal{X}_2 = \{0,1,\ldots,K-1\}$ Q_1 and Q_2 are both uniform across the alphabet, and $m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ depends on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ only via $\boldsymbol{x}_1 \oplus \boldsymbol{x}_2$, where $\oplus$ denotes addition modulo K. Decoding metrics with this property are motivated by classes of multiple access channels, $P(\boldsymbol{y}|\boldsymbol{x}_1,\boldsymbol{x}_2)$,

in which the users interfere with each other additively, i.e., $P(\boldsymbol{y}|\boldsymbol{x}_1,\boldsymbol{x}_2) = W(\boldsymbol{y}|\boldsymbol{x}_1 \oplus \boldsymbol{x}_2)$. Still, the dependence of $\boldsymbol{y}$ on $\boldsymbol{x}_1 \oplus \boldsymbol{x}_2$ can be arbitrary.

We now define three kinds of equivalence classes: $\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})$ is the set of $(\boldsymbol{x}_1',\boldsymbol{x}_2')$ such that $\forall\theta \in \Theta$, $m_\theta(\boldsymbol{x}_1',\boldsymbol{x}_2',\boldsymbol{y}) = m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$, $\mathcal{T}(\boldsymbol{x}_1|\boldsymbol{x}_2,\boldsymbol{y})$ is the set of $\boldsymbol{x}_1'$ such that $\forall\theta \in \Theta$, $m_\theta(\boldsymbol{x}_1',\boldsymbol{x}_2,\boldsymbol{y}) == m_\theta(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ and $\mathcal{T}(\boldsymbol{x}_2|\boldsymbol{x}_1,\boldsymbol{y})$ is defined similarly with '1' and '2' swapped. We also assume, as before, that for every $\boldsymbol{y}$, the number of different type classes $\{\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})\}$ is upper bounded by $2^{n\Delta_n}$. Next, define the following functions:

$$U_0(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log\{(Q_1\times Q_2)[\mathcal{T}(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{y})]\} \quad (6)$$

$$U_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log Q_1[\mathcal{T}(\boldsymbol{x}_1|\boldsymbol{x}_2,\boldsymbol{y})] \quad (7)$$

$$U_2(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y}) = -\frac{1}{n}\log Q_2[\mathcal{T}(\boldsymbol{x}_2|\boldsymbol{x}_1,\boldsymbol{y})]. \quad (8)$$

Define the universal decoding metric $U(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ as the minimum among the following three expressions: $U_0(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_1-R_2$, $U_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_1$, and $U_2(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})-R_2$. We argue that $U(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})$ competes favorably with the best m_θ in a sense analogous to that asserted in Theorem 1. This decoding metric is different from the universal decoding metrics used for the MAC, for example, in [13] and [7], which were based on the MMI decoder and the minimum empirical conditional entropy (minimum equivocation) rule, respectively. Similarly as before, suppose that U_0, U_1 and U_2 can be uniformly upper bounded by U_0', U_1' and U_2', respectively, and assume that:

$$\max_{\boldsymbol{y}}\sum_{\boldsymbol{x}_1,\boldsymbol{x}_2} Q_1(\boldsymbol{x}_1)Q_2(\boldsymbol{x}_2)2^{nU_0'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \stackrel{\cdot}{\leq} 1 \quad (9)$$

$$\max_{\boldsymbol{x}_2,\boldsymbol{y}}\sum_{\boldsymbol{x}_1} Q_1(\boldsymbol{x}_1)2^{nU_1'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \stackrel{\cdot}{\leq} 1 \quad (10)$$

$$\max_{\boldsymbol{x}_1,\boldsymbol{y}}\sum_{\boldsymbol{x}_2} Q_2(\boldsymbol{x}_1)2^{nU_2'(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{y})} \stackrel{\cdot}{\leq} 1. \quad (11)$$

Then, U_0', U_1' and U_2' can replace U_0, U_1 and U_2, respectively, in the universal decoding metric, denoted in turn by U', and the upper and lower bounds continue to hold with U' replacing U. The application of this to the LZ decoding metric is a straightforward extension to the one exercised above in the single–user case (see [11] for details).

C. Comments on the Continuous Alphabet Case

It is possible to extend Theorem 1 to the case of continuous alphabets, but this requires more caution. For one thing, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ should be redefined by allowing some small tolerance, i.e., the requirement $m_\theta(\boldsymbol{x},\boldsymbol{y}) = m_\theta(\boldsymbol{x},\boldsymbol{y})$ should be replaced by $|m_\theta(\boldsymbol{x},\boldsymbol{y}) - m_\theta(\boldsymbol{x},\boldsymbol{y})| \leq \epsilon$, where $\epsilon > 0$ tends to zero after $n \to \infty$. This is to guarantee that $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ captures a positive volume and that $K_n(\boldsymbol{y})$ (now, redefined as the number of $\{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\}$ required to cover the set of channel input vectors, possibly obeying an input constraint) is finite. We will not delve into the technical details of this extension any further[2] Instead, we will merely demonstrate the universal decoding metric in a certain special case, where the class of decoding metrics depend on $\boldsymbol{x}$ and $\boldsymbol{y}$ only via second order empirical statistics extracted from these sequences.

Example 3. Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and let Q be an i.i.d. zero–mean Gaussian density with variance σ^2. Let $\theta = (\theta_1,\theta_2)$ and let $\mathcal{M}$ be the class of decoding metrics $m_\theta(\boldsymbol{x},\boldsymbol{y}) = \theta_1\sum_{i=1}^n x_iy_i + \theta_2\sum_{i=1}^n x_i^2$. Denoting $C(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n x_iy_i$ and $S(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n x_i^2$, then $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ should be redefined as the set of $\boldsymbol{x}'$, where $C(\boldsymbol{x}',\boldsymbol{y})$ and $S(\boldsymbol{x}')$ are within ϵ close to $C(\boldsymbol{x},\boldsymbol{y})$ and $S(\boldsymbol{x})$, respectively. Using the methods of [10], it is easy to show that $U(\boldsymbol{x},\boldsymbol{y}) = \frac{S(\boldsymbol{x})}{2\sigma^2} - \frac{1}{2}\ln[S(\boldsymbol{x})(1-\rho_{\boldsymbol{xy}}^2)]$, where $\rho_{\boldsymbol{xy}} = C(\boldsymbol{x},\boldsymbol{y})/\sqrt{S(\boldsymbol{x})S(\boldsymbol{y})}$ is the empirical correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$, and where we have used natural logarithms instead of base 2 logarithms.

[2] See [10] where these details have been fully worked out in the context of universal decoding for the Gaussian channel with a deterministic interference.

REFERENCES

[1] I. Csiszár, "Linear codes for sources and source networks: error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 4, pp. 585–592, July 1982.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[3] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1726–1745, September 1998.

[4] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, vol. 48, no. 6, pp. 1504–1517, June 2002.

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

[6] A. Lapidoth and J. Ziv, "On the universality of the LZ–based noisy channels decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1746–1755, September 1998.

[7] Y.-S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 376–386, March 1996.

[8] Y. Lomnitz and M. Feder, "Communication over individual channels – a general framework," arXiv:1023.1406v1 [cs.IT] 7 Mar 2012.

[9] Y. Lomnitz and M. Feder, "Universal communication over modulo–additive channels with an individual noise sequence," arXiv:1012.2751v2 [cs.IT] 7 May 2012.

[10] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

[11] N. Merhav, "Universal decoding for arbitrary channels relative to a given family of decoding metrics," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5566–5576, September 2013.

[12] V. Misra and T. Weissman, "The porosity of additive noise sequences," arXiv:1025.6974v1 [cs.IT] 31 May 2012.

[13] J. Pokorny and H. M. Wallmeier, "Random coding bound and codes produced by permutations for the multiple access channel," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 6, pp. 741–750, November 1985.

[14] O. Shayevitz and M. Feder, "Communicating using feedback over a binary channel with arbitrary noise sequence," *Proc. ISIT 2005*, pp. 1516–1520, Adelaide, Australia, September 2005.

[15] N. Shulman, *Communication over an Unknown Channel via Common Broadcasting*, Ph.D. dissertation, Department of Electrical Engineering – Systems, Tel Aviv University, July 2003. http://www.eng.tau.ac.il/∼shulman/papers/Nadav_PhD.pdf

[16] J. Ziv, "Universal decoding for finite–state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.