Linear Complexity Universal Decoding with Exponential Error Probability Decay

Todd P. Coleman, Muriel Médard, and Michelle Effros
{colemant, medard}@mit.edu, effros@caltech.edu
MIT, Caltech

***Abstract*— In this manuscript we consider linear complexity binary linear block encoders and decoders that operate *universally* with exponential error probability decay. Such scenarios may be relevant in wireless scenarios where probability distributions may not be fully characterized due to the dynamic nature of wireless environments. More specifically, we consider the setting of fixed length-to-fixed length near-lossless data compression of a memoryless binary source of unknown probability distribution as well as the dual setting of communicating on a binary symmetric channel (BSC) with unknown crossover probability. We introduce a new 'min-max distance' metric, analogous to minimum distance, that addresses the universal binary setting and has the same properties as that of minimum distance on BSCs with known crossover probability. The code construction and decoding algorithm are universal extensions of the 'Expander Codes' framework of Barg and Zémor and have identical complexity and exponential error probability performance.**

I. Introduction

In this discussion we consider code constructions for fixed block length universal coding for the two dual settings of data compression and channel coding. The compression scenario mentioned could be relevant, for instance, in a wireless sensor network where the following two points apply:

1) Due to the time-varying nature of the field being sensed, the probability distribution on the data is not completely accurately modeled,
2) Complexity, memory, and energy constraints make a universal fixed-to-fixed length algebraic compression approach more viable than a universal fixed-to-variable length compression approach (such as Lempel-Ziv [1], [2] or Burrows-Wheeler [3]) that requires dictionaries and table-lookups.

Similarly, due to the time-varying and multipath effects of the wireless channel, the universal channel coding scenario could be relevant where phase information cannot be accurately tracked.

More specifically, we take interest in universal decoding for binary memoryless settings, where the decoder does not have knowledge of the probability distribution to aid in decoding. We consider the case where a linear mapping $H : \{0,1\}^N \to \{0,1\}^M$ is used to map $\underline{u} \in \{0,1\}^N$ to $\underline{s} \in \{0,1\}^M$ via

$$\underline{s} = H\underline{u} \tag{1}$$

where $M < N$ and U is memoryless with $Pr(U_i = 1) = p$. The decoder knows that $\underline{u}$ must be consistent with $\underline{s}$, in other words it must lie in the coset

$$\mathrm{Co}\,(H, \underline{s}) = \{\underline{u} \mid H\underline{u} = \underline{s}\}, \tag{2}$$

and selects $\hat{\underline{u}}$ as the 'best' coset member (in a universal sense). This encompasses two settings:

a) Fixed-to-fixed length near-lossless data compression, where $\underline{u}$ is identified as the sourceword and $\underline{s}$ is the syndrome, the output of the compression operation.
b) A binary symmetric channel $\underline{y} = \underline{x} \oplus \underline{u}$. By using a linear code $\mathcal{C}$ for $\underline{x}$, and identifying the parity check matrix H with $\mathcal{C}$ as

$$\mathcal{C} = \{\underline{x} : H\underline{x} = \underline{0}\}, \tag{3}$$

then we have that a sufficient statistic for decoding is

$$H\underline{y} = H\underline{u} = \underline{s}.$$

Successfully decoding for the noise vector $\underline{u}$ is equivalent to successfully decoding for the transmitted codeword $\underline{x}$:

$$\hat{\underline{x}} = \hat{\underline{u}} \oplus \underline{y}.$$

It is the job of the decoder to universally (without knowledge of p - in particular, the sign of $p - \frac{1}{2}$) find the best estimate of $\underline{u}$. We assume that the rate R is achievable (i.e. for compression, $R > h\,(U)$ and for the BSC, $R < 1 - h\,(U)$).

We note that if we knew that $p < \frac{1}{2}$ then the optimal decoding rule would be to find the coset leader,

$$\hat{\underline{u}} \in \arg \min_{\underline{u} \in \mathrm{Co}(H, \underline{s})} w_h\,(\underline{u}), \tag{4}$$

where $w_h\,(\cdot)$ is the Hamming weight. In such a setting, a figure of merit for good codes is the minimum distance of H:

$$d_{min} \triangleq \min_{\underline{x} \neq 0 \in \mathcal{C}} w_h\,(\underline{x}),$$

and the larger the minimum distance, the better the guaranteed performance.

Likewise, if $p > \frac{1}{2}$, the optimal rule would be

$$\hat{\underline{u}} \in \arg \min_{\underline{u} \in \mathrm{Co}(H, \underline{s})} w_h\,(\underline{u} \oplus \underline{1}) \Leftrightarrow \hat{\underline{z}} \in \arg \min_{\underline{z} \in \mathrm{Co}(H, \underline{s} \oplus \underline{s}^1)} w_h\,(\underline{z})$$

where $\underline{s}^1 \triangleq H\underline{1}$ and $\underline{1}$ is the all one vector. Since this is also a coset leader decoding problem, d_{min} is again a figure of merit for performance.

In our setting, however, we do not know if $p < \frac{1}{2}$ or $p > \frac{1}{2}$. It has been known in the information theory literature for quite a while [4], [5] that in the *universal* setting, linear codes

0-7803-9305-8/05/$20.00 ©2005 IEEE

still suffice to attain all achievable rates and the same error exponent as the random coding exponent. A universal decoder must select the 'best' source vector consistent with the observation $\underline{s}$. Csiszár's 'minimum-entropy' decoder [5] selects as the source reconstruction the coset's entropy minimizer

$$\hat{\underline{u}} = \arg \min_{\underline{u} \in \mathrm{Co}(H,\underline{s})} h\left(P_{\underline{u}}\right), \tag{5}$$

where:

- For any $\underline{u} \in \{0,1\}^n$, $P_{\underline{u}}$ is the empirical distribution of $\underline{u}$ over $\{0,1\}$ where

$$P_{\underline{u}}(1) = \frac{1}{n} w_h\left(\underline{u}\right), \quad P_{\underline{u}}(0) = 1 - P_{\underline{u}}(1). \tag{6}$$

- For any probability distribution $P = (P_0, P_1)$ over $\{0,1\}$, $h(P)$ is its entropy:

$$h(P) = -\sum_{i=0}^{1} P_i \log_2 P_i \tag{7}$$

Note that since we may express $h\left(P_{\underline{u}}\right)$ in (5) as

$$h_b\left(\frac{1}{n} w_h\left(\underline{u}\right)\right), \tag{8}$$

where $h_b(p) \triangleq -p\log_2 p - (1-p)\log_2 p$. Csiszár illustrated in [5] that linear codes suffice for all achievable rates and moreover that there exist linear codes that attain the random coding exponent under the universal minimum-entropy decoder.

In section II we discuss a measure of good codes - the 'min-max distance', that has the same property as minimum distance: the larger the min-max distance, the better the guarantee we have on successful *universal* minimum-entropy decoding. We also point, out by exploiting well-known results on typical linear codes from the coding literature [6], that the min-max distance of the typical linear code is the same as the minimum-distance of the typical linear code: the Gilbert-Varshamov distance.

Section III considers code construction and decoding based on expander graphs. The algorithm is analogous to the 'Expander Codes' work of Barg and Zémor [7] (originally formulated by Sipser and Spielman [8]) - the difference is that in each iteration, we replace minimum-distance decoding with minimum-entropy decoding, so that we are operating in the universal setting. Here we also illustrate how selecting the component codes to have good min-max distance allows for the decoding algorithm to have the same complexity as well as exponential error probability decay.

II. A Minimum-Distance Style Figure of Merit for Universal Decoding in the Binary Setting

We now discuss the 'min-max distance' of a binary linear code $\mathcal{C}$ with parity check matrix H, given by

$$d_{\text{min.max}} \triangleq \min_{\underline{u} \in \mathcal{C}, \underline{u} \neq 0} \min\left(w_h\left(\underline{u}\right), w_h\left(\underline{1} \oplus \underline{u}\right)\right). \tag{9}$$

We illustrate the motivation for using the min-max distance in code design using the following example. Consider any linear code with parity check matrix H for which $\underline{1}$ is a member of $\mathcal{C} = \mathrm{Co}(H, \underline{0})$. Then the minimum-entropy decoder has probability of error equal to $\frac{1}{2}$ by the following argument. For any $\underline{u} \in \mathrm{Co}(H, \underline{s})$, $\underline{u} \oplus \underline{1} \in \mathrm{Co}(H, \underline{s})$. Further, $h\left(P_{\underline{u}}\right) = h\left(P_{\underline{u} \oplus \underline{1}}\right)$. Thus $\underline{u}$ and $\underline{u} \oplus \underline{1}$ are indistinguishable to a minimum-entropy decoder. Note that the 'min-max' distance of any such linear code H is 0, which captures this undesirable effect.

A. An Analogue to the Half-Minimum Distance Criterion

We know that under minimum-distance decoding, if the error sequence has hamming weight less than half the minimum distance, then we can guarantee success. It is natural to ask if there is an analogous statement regarding min-max distance and minimum-entropy decoding. The answer is yes:

Lemma 2.1: Consider any M by N binary matrix H and its associated $d_{\text{min.max}} = N\delta_{\text{min.max}}$, given by (9). If $w_h\left(\underline{u}\right) < \frac{1}{2} d_{\text{min.max}}$ **or** $w_h\left(\underline{u} \oplus \underline{1}\right) < \frac{1}{2} d_{\text{min.max}}$, then $\underline{u}$ is the unique solution to

$$\min_{\hat{\underline{u}} \in \mathrm{Co}(H,\underline{s})} h\left(P_{\hat{\underline{u}}}\right).$$

Stated alternatively, if $h\left(P_{\underline{u}}\right) < h\left(\frac{1}{2}\delta_{\text{min.max}}\right)$ then $\underline{u}$ is the unique solution to

$$\min_{\hat{\underline{u}} \in \mathrm{Co}(H,\underline{s})} h\left(P_{\hat{\underline{u}}}\right).$$

Proof: see Appendix.

B. Distance Spectrum of the Typical Random Linear Code

We know from traditional coding theory that there exist linear codes with minimum distance lying on the Gilbert-Varshamov bound:

$$\delta_{GV}(R) \triangleq \text{ the root } \delta \leq \frac{1}{2} \text{ of } h_b(\delta) = 1 - R.$$

This allows us to guarantee that using such codes on a BSC with known crossover probability and minimum-distance decoding will result in attaining the random coding error exponent. Considering the decoding success guarantee of the previous subsection and its similarity to minimum distance, it is natural to ask the analogous question the min-max distance of linear codes. The answer to this question can be found from traditional coding theory:

Lemma 2.2 (Barg-Forney [6]): The typical length-N linear code from the random binary linear code ensemble has with probability $1 - 2^{\Omega(N)}$ a distance distribution given by $\mathcal{N}_{TLC}(d)$ for $d \triangleq N\delta \in \{1, 2, \ldots, N\}$:

$$\mathcal{N}_{TLC}(d) = 0, \quad \text{if} \left|\frac{1}{2} - \delta\right| \geq \frac{1}{2} - \delta_{GV}(R) + \epsilon$$

From the symmetry of $\mathcal{N}_{TLC}(d)$ it follows that with exponentially high probability, $d_{\text{min.max}} \geq N\delta_{GV}(R) - \epsilon$.

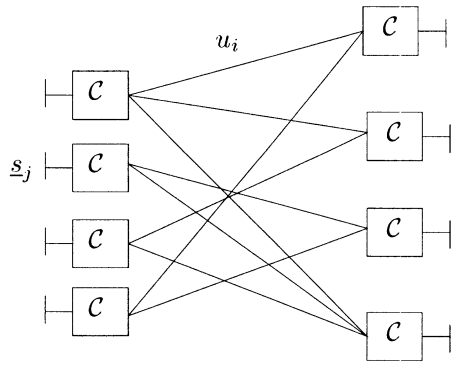

Fig. 1. graphical representation of the expander code

III. Codes Based on Expander Graphs

The motivation for this section follows very closely the work on expander codes by Barg and Zémor [7]. Here we will consider a syndrome-former [9, Sec VIII.B] representation of such codes.

We consider a Δ-regular bipartite graph $G = (V = A \cup B, E)$. The set of edges E will be associated with bits that must satisfy constraints. Each node $j \in V$ is adjacent to Δ edges and corresponds to a $(\Delta, \Delta - m_j)$ code C_j that we can associate with a $m_j \times \Delta$ parity check matrix H_j. Also adjacent to node j is a set of m_j half-edges or 'dongles' [9, Sec VIII.B] connected to m_j bits, given by $\underline{s}_j$. The code C_j, for $j \in V$, enforces the constraint that

$$H_j \underline{u}_{|N(j)} = \underline{s}_j,$$

where $\underline{u}_{|N(j)}$ is $\underline{u}$ projected onto the edge indices that are adjacent to vertex j. Figure 1 provides an illustration.

A. Encoding

Encoding for the compression situation is done quite simply. $\underline{u}$ is mapped to $\underline{s}$ setting the edges on the graph G to $\underline{u}$, and applying $\underline{s}_j = H_j \underline{u}_{|N(j)}$ for all $j \in V$. We note that there are n nodes and each node has degree Δ, and since there are $N = n\Delta$ edges, this is done with linear complexity. For the channel coding scenario, the encoding done is the same as discussed in [8], [7].

B. Decoding

Decoding will be done by applying the syndrome-former equivalent iterative algorithm of Barg and Zémor [7]. However, because we are in the universal setting, we cannot simply perform coset-leader decoding at each code $C_j, j \in V$. Instead, we must perform a universal decoding algorithm, which corresponds to minimum-entropy decoding. What we show in the appendix is that if source sequence projected onto the indices of any subcode behaves 'typically', then we can guarantee that the subcode will decode using the universal minimum-entropy decoder to the true realization. From here we apply the graph expansion arguments of Barg and Zémor to arrive at the same result as in their setting.

Let $\underline{u} \in \{0,1\}^N$ be the true sequence that has been mapped to $\underline{s} \in \{0,1\}^M$ according to (1). The first iteration, which we call a left-decoding step, applies in parallel, for every left vertex $j \in A$, minimum-entropy decoding according to $\underline{s}_j$ to construct a $\underline{u}_{|N(j)}$. In other words, a left-decoding step is a function $L : \{0,1\}^{m_j} \to \{0,1\}^{\Delta}$ where

$$L(\underline{s}_j) \in \arg \min_{\underline{\tilde{u}} \in \mathrm{Co}\left(H_j, \underline{s}_j\right)} h\left(P_{\underline{\tilde{u}}}\right).$$

So $\{L(\underline{s}_j)\}_{j \in A}$ produces a vector $\underline{\hat{u}} \in \{0,1\}^N$. After applying $\{z_j = H_j \underline{\hat{u}}_{|N(j)}\}_{j \in B}$, we then apply the function R in the same manner that L operates. We alternately apply repeat left-decoding and right-decoding steps. The procedure stops if it encounters a fixed point or after having operated for $O(\log N)$ steps.

We will now identify the vectors of $\{0,1\}^N$ with their indices that have entries different from the original sourceword $\underline{u}$. For any left vertex $j \in A$ we will say that j is a left-survivor if $\hat{u}_i \neq u_i$ for some $i \in N(j)$. We likewise define a right-survivor. We note that for the universal case, we need to operate not on the minimum distance of the code corresponding to H_j, but rather the min-max distance given in the appendix A. From the appendix C we have that if $w_h\left(\underline{u}_{|N(j)}\right) < \frac{1}{2} d_{\text{min,max}}$ or $w_h\left(\underline{u}_{|N(j)} \oplus \underline{1}\right) < \frac{1}{2} d_{\text{min,max}}$ then no error will result in the universal decoder corresponding to node j with matrix H_j. Note that if a vector $\underline{\hat{u}} \in \{0,1\}^N$ has no survivors then we will arrive at a fixed point. By applying properties of the expansion graph as discussed in [7], we know that if the number of left-survivors s is small enough, then the number of right-survivors s' is strictly smaller and satisfies $s' \leq \beta s$ where $\beta < 1$. Thus it will follow that the algorithm will converge to a fixed point in a number of iterations logarithmic in N. That the overall decoding complexity is $O(N)$ follows from using a circuit of size $O(N \log N)$ and depth $O(\log N)$, as discussed in [8], [7].

C. Error Probability

We note that the error probability analysis in [7, Sec. III] essentially relies on the following:

a) Using a graph with good expansion properties
b) Using a constituent linear code $\mathcal{C}_j$ of length Δ, associated with each vertex of the graph, that has following two properties:
   1. $\mathcal{C}_j$ attains the BSC random coding error exponent: $P_e^j \leq 2^{-\Delta(E_r(R)-\epsilon)}$ for some small $\epsilon > 0$.
   2. $\mathcal{C}_j$ has minimum distance near the Gilbert-Varshamov bound: $d_{\min}(\mathcal{C}_j) = \Delta \delta_{GV}(R) - \epsilon$ for some small $\epsilon > 0$.

In our setting we may use the same expander graph as in [7] to address a). We may address b.1) by employing the type of code, shown to exist by Csiszár [5, sec. III], that attains the random coding exponent under minimum-entropy decoding. Finally, we address b.2) by noting that from Lemma 2.2 there exist, with exponentially high probability if chosen randomly, codes with $d_{\text{min,max}}$ lying on the Gilbert-Varshamov bound. Consequently, we may replace all the arguments regarding d_{min} in the error probability analysis of [7, Sec. III] with arguments regarding $d_{\text{min,max}}$ and the details carry through

in a straightforward manner. Thus we attain the same error exponent as discussed in the analysis in [7, Sec. III], which is positive for all achievable rates.

REFERENCES

[1] A. Lempel and J. Ziv, "A universal algorithm for sequential data compression," IEEE Transactions on Information Theory, pp. 337–343, 1977.

[2] A. Lempel and J. Ziv, "Compression of individual sequences via variable-rate coding," IEEE Transactions on Information Theory, pp. 530–536, 1978.

[3] M. Effros, K. Visweswariah, S. R. Kulkarni, and S. Verdu, "Universal lossless source coding with the Burrows Wheeler transform," IEEE Transactions on Information Theory, vol. 48, no. 5, pp. 1061–1081, May 2002.

[4] V. D. Goppa, "Universal decoding for symmetric channels," Probl. Peredachi Inform., vol. 11, no. 1, pp. 15–22, 1975, (In Russian).

[5] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," IEEE Transactions on Information Theory, vol. 28, no. 4, pp. 585–592, 1982.

[6] A. Barg and G. D. Forney, "Random codes: Minimum distances and error exponents," IEEE Transactions on Information Theory, pp. 2568–2573, 2002.

[7] A. Barg and G. Zémor, "Error exponents of expander codes," IEEE Transactions on Information Theory, vol. 48, no. 6, pp. 1725–1729, 2002.

[8] M. Sipser and D. Spielman, "Expander codes," IEEE Transactions on Information Theory, vol. 42, no. 6, pp. 1710–1722, 1996.

[9] G. D. Forney, "Codes on graphs: Normal realizations," IEEE Transactions on Information Theory, pp. 101–112, 2001.

APPENDIX

Proof of Lemma 2.1:

Suppose we have that a parity check matrix H has $d_{\text{min.max}} = d = N\delta$. From the definition of $d_{\text{min.max}}$ in (9) it follows that for any nonzero $\underline{\tilde{u}} \in \text{Co}(H, \underline{0})$, the following holds:

$$w_h(\underline{\tilde{u}}) \geq d \quad \Leftrightarrow \quad w_h(\underline{\tilde{u}} \oplus \underline{1}) \leq n - d \qquad (10)$$

$$\text{and } w_h(\underline{\tilde{u}}) \leq n - d \quad \Leftrightarrow \quad w_h(\underline{\tilde{u}} \oplus \underline{1}) \geq d. \qquad (11)$$

Then if

$$w_h(\underline{u}) < \frac{1}{2}d \Leftrightarrow w_h(\underline{u} \oplus \underline{1}) > n - \frac{1}{2}d \qquad (12)$$

is satisfied, we have

1)

$$\begin{aligned} w_h(\underline{u} \oplus \underline{\tilde{u}}) &\leq w_h(\underline{u}) + w_h(\underline{\tilde{u}}) \\ &< \frac{1}{2}d + n - d \text{ owing to (12),(11)} \\ &= n - \frac{1}{2}d \end{aligned}$$

2)

$$\begin{aligned} w_h(\underline{u} \oplus \underline{\tilde{u}}) &= n - w_h(\underline{u} \oplus \underline{\tilde{u}} \oplus \underline{1}) \\ &\geq n - [w_h(\underline{u}) + w_h(\underline{\tilde{u}} \oplus \underline{1})] \\ &> n - \left[\frac{1}{2}d + n - d\right] \text{ owing to (12),(10)} \\ &= \frac{1}{2}d. \end{aligned}$$

Likewise, if

$$w_h(\underline{u} \oplus \underline{1}) < \frac{1}{2}d \Leftrightarrow w_h(\underline{u}) > n - \frac{1}{2}d \qquad (13)$$

then

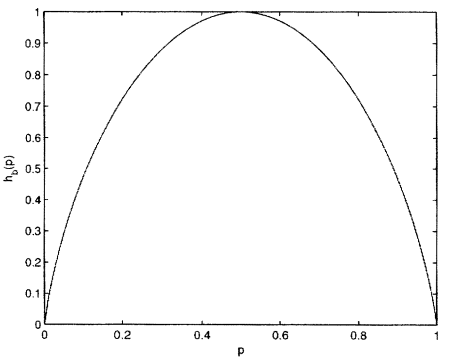

Fig. 2. The Binary Entropy Function

1)

$$\begin{aligned} w_h(\underline{u} \oplus \underline{1} \oplus \underline{\tilde{u}}) &\leq w_h(\underline{u} \oplus \underline{1}) + w_h(\underline{\tilde{u}}) \\ &< \frac{1}{2}d + n - d \text{ owing to (13),(11)} \\ &= n - \frac{1}{2}d \end{aligned}$$

2)

$$\begin{aligned} w_h(\underline{u} \oplus \underline{1} \oplus \underline{\tilde{u}}) &= n - w_h(\underline{u} \oplus \underline{1} \oplus \underline{\tilde{u}} \oplus \underline{1}) \\ &\geq n - [w_h(\underline{u} \oplus \underline{1}) + w_h(\underline{\tilde{u}} \oplus \underline{1})] \\ &> n - \left[\frac{1}{2}d + n - d\right] \text{ owing to (13),(10)} \\ &= \frac{1}{2}d. \end{aligned}$$

Thus in either case, because of the following properties:

i. $\delta \leq \frac{1}{2}$ (this follows from the definition (9) of $d_{\text{min.max}}$),

ii. The binary entropy function $h_b(\cdot)$ is monotonically increasing on $[0, \frac{1}{2})$ (see Figure 2),

iii. The binary entropy function $h_b(\cdot)$ is symmetric around $\frac{1}{2}$ (see Figure 2),

we have that $h(P_{\underline{\tilde{u}} \oplus \underline{u}}) > h(P_{\underline{u}})$. Thus if we define $\underline{s} = H\underline{u}$ then we have that $\underline{u}$ is the unique solution to

$$\min_{\underline{\hat{u}} \in \text{Co}(H, \underline{s})} h(P_{\underline{\hat{u}}}).$$

The alternative statement in the lemma holds because the two statements are equivalent:

- $w_h(\underline{u}) < \frac{1}{2}d_{\text{min.max}}$ **or** $w_h(\underline{u} \oplus \underline{1}) < \frac{1}{2}d_{\text{min.max}}$,
- $h(P_{\underline{u}}) < h(\frac{1}{2}\delta)$.

This also follows from properties i.–iii. above. ∎