Efficiently Decodable Low-Rate Codes Meeting Gilbert-Varshamov Bound

VENKATESAN GURUSWAMI
Department of Computer Science and Engineering
University of Washington
Seattle, WA
Email: venkat@cs.washington.edu

PIOTR INDYK
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA
Email: indyk@theory.lcs.mit.edu

Abstract

We demonstrate a probabilistic construction of binary linear codes meeting the Gilbert-Varshamov bound (with overwhelming probability) for rates up to about 10^{-4}, together with *polynomial* time algorithms to perform encoding and *decoding up to half the distance*. This is the first such result (for some positive rate) with polynomial decoding complexity; previously a similar result (up to rate about 0.02) was known with sub-exponential time decoding (Zyablov and Pinsker, 1981).

1 Introduction

One of the central challenges in coding theory is to construct codes with optimal rate vs. distance trade-off together with effi cient encoding and decoding algorithms. Despite decades of research, however, even the best trade-off between rate and distance is unknown for binary codes. A random binary linear code, although not known to achieve optimal rate vs. distance tradeoff, provides very good bounds with overwhelming probability. In fact, they achieve the *best known* rate vs. distance trade-off called the Gilbert-Varshamov (GV) bound. At the same time, since there is no general way to certify that a random code is good or to actually decode a good code, the random code construction is only of theoretical interest.

While a random linear code has little structure, Thommesen [7] proved that one can meet the GV bound by picking a random code from a more structured ensemble of binary linear codes. Specifi cally, he proved that a concatenated code, with an outer Reed-Solomon code and binary

linear inner codes of suitable parameters picked *independently at random* for the various outer codeword positions, meets the GV bound with high probability. Still, unique decoding the resulting codes up to the optimal radius (i.e., up to half the GV bound) was posed as an open problem by Thommesen, and it has remained unresolved for over 20 years now.

In this paper, we solve this question for low rates (or large distance). Specifically, for rates up to about 10^{-4}, we present a decoding algorithm that can decode a randomly chosen code from Thommesen's ensemble up to half the distance with overwhelming probability. A similar result was shown by Zyablov and Pinsker [8] for rates up to 0.02, but their decoding algorithm ran in time $2^{O(\sqrt{N})}$ where N is the block length, where as we achieve polynomial decoding complexity. Ours is the first result to achieve polynomial decoding complexity up to half the GV bound for some positive rate. The question of extending Zyablov and Pinsker's result to higher rates was mentioned as Open Problem 6.13 in [1], and the question of doing so with *polynomial* decoding complexity is an important question left open by our work.

We remark that though we do not know how to certify that the distance of the overall code will meet the GV bound (but we do know it will do so with high probability), we can certify the decoding property deterministically in the following sense: the decoding algorithm is guaranteed to (list) decode the code up to a fraction $1/4$ of errors, regardless of whether the distance of the code meets the GV bound or not. This certification property gives us the desirable feature that a failure of the algorithm to uniquely decode the closest codeword from the received word (due to there being multiple close-by codewords) is in fact a "proof" that the distance fell short of the GV bound, and till we detect this failure, all decodings produced by the algorithm are indeed the correct closest codewords.[1]

We note that the classical GMD approach to decoding concatenated codes can correct only up to half the product of the outer and inner distances, or half the so-called Zyablov bound, which is in general much smaller than half the GV bound. Instead, we use the list decoding algorithms known for Reed-Solomon codes to accomplish this task. To get the best bound on rate up to which we can decode up to half the GV bound, we use a recent algorithm from [5] that uses soft information passed from the inner decodings to the Reed-Solomon list decoder, though we also discuss "more straightforward" ways of using list decoding to decode the concatenated code.

It should be said that our result is technically simple to achieve given Thommesen's result [7] and the recent progress on list decoding concatenated codes [5]. But we hope that the statement of the result is inspiring and that it motivates further investigation of structured sample spaces of codes that meet the GV bound as well as improved ways to exploit the list decoding algorithms in decoding concatenated codes. Given that we currently seem quite far from the target of explicit constructions of codes meeting the GV bound, augmenting existing probabilistic constructions with efficient decoding algorithms is a worthwhile pursuit. In this regard, improving the rate up to which we can decode is a central open question arising out of our work, and we end the paper with a discussion of why doing this using the current technology for list decoding concatenated codes is likely to be quite challenging.

[1] If the algorithm fails to output any codeword, then that is "proof" that more than a fraction $1/4$ of errors occurred, and in such a case we are not required to be able to decode anyway.

2 Background on the Thommesen Construction

A random code in the ensemble of codes considered by Thommesen is sampled as follows. There is a fixed Reed-Solomon code of block length N (say) over $\mathrm{GF}(2^t)$ that is used as the outer code, and it is concatenated with N binary linear inner codes of block length b (say) picked independently at random. The argument that this process leads to a linear code that meets the GV bound with high probability proceeds roughly as follows. The weight distribution of the Reed-Solomon code, which is known explicitly, tells us the number of outer codewords of a specific weight. Now, an outer codeword of weight, say u, is mapped into a specific bit vector which is 0's outside the u blocks corresponding to the u non-zero symbols, with probability exactly 2^{-ub} (note that this probability depends only on the weight of the outer codeword). By performing a union bound over all low-weight vectors and all outer codewords (whose weight distribution is known), gives an upper bound on the probability that the resulting code fails to have a certain minimum distance. A careful choice of parameters and calculations then shows that the bound on minimum distance achieved indeed matches the GV bound.

To give some intuition as to why the GV bound is met, we now describe a quick back-of-the-envelope calculation for the case of relative distance very close to $1/2$.

Proposition 1 *Let C_0 be a linear code of dimension k over $\mathrm{GF}(2^t)$ that has relative distance $(1-\varepsilon)$ and rate r_0, for some very small $\varepsilon > 0$. Consider the ensemble of codes similar to the Thommesen scheme where independently picked binary linear inner codes of rate ε and dimension t are used as inner codes with C_0 as outer code. A random code from this ensemble fails to have relative distance at least $(1/2 - O(\varepsilon))$ with exponentially small probability.*

Proof: Consider a non-zero codeword of the resulting concatenated code. There are $(1-\varepsilon)k/r_0$ outer symbols that are non-zero, and for each such symbol, the probability that any single bit of its encoding by the associated inner code equals 0 is exactly $1/2$. Overall, the weight of the codeword is a Bernoulli random variable with success probability $1/2$ and $M = (1-\varepsilon)\cdot(k/r_0)\cdot b/\varepsilon = (1-\varepsilon)\frac{kb}{r_0\varepsilon}$ trials. The probability that we exceed the expectation by ε is at most $\exp(-O(1)\varepsilon^2 M) = \exp(-O(kb\varepsilon/r_0)) = \exp(-O(kb))$, where the last step follows since $r_0 \leq \varepsilon$ by the Singleton bound. Since the total number of codewords is 2^{kb}, a union bound shows that except with exponentially small probability, the code has relative distance at least $(1-\varepsilon)(1/2-\varepsilon) \geq 1/2 - 2\varepsilon$. $\square$

Applying the above with a code like the Reed-Solomon code which achieves the Singleton bound (so $r_0 = \varepsilon$), we get codes of relative distance $1/2 - O(\varepsilon)$ and rate ε^2. Up to constant factors, this matches the GV bound, since for small ε, $H(1/2-\varepsilon) = 1 - O(\varepsilon^2)$. The computation in Thommesen's paper [7] is more complicated than the above calculation as it shows that the GV bound can be met exactly.

We refer to the paper by Thommesen [7] for further details of the construction and proof, and below state the main result from that paper in a form which will be useful to us. For $0 \leq x \leq 1$, define the function $\alpha(x) = 1 - H(1 - 2^{x-1})$ (where as usual $H(y)$ denotes the binary entropy function of y).

Theorem 1 *[7] Let r_0 and R_0 be given such that $0 < r_0 \leq 1$ and $0 < R_0 \leq \alpha(r_0)/r_0$. For a large enough integer t, let $\mathrm{RS}_t(R_0)$ be the Reed-Solomon code over $\mathrm{GF}(2^t)$ of block length $N = 2^t$ and*

rate R_0. Consider an ensemble of concatenated codes with $\mathrm{RS}_t(R_0)$ as outer code with varying inner codes, that is codes where the codewords are of the form

$$uG = [u_1G_1, u_2G_2, \ldots, u_NG_N], \quad u = \langle u_1, u_2, \ldots, u_N \rangle \in \mathrm{RS}_t(R_0)$$

where G_i, $1 \leq i \leq N$, are binary $t \times t/r_0$ matrices, each picked at random and independently of the others. Then, for every $\varepsilon > 0$, the probability that a random concatenated code from this ensemble (defined by a random choice of the G_i's) has relative distance at most $H^{-1}(1-r_0R_0)-\varepsilon$ is exponentially small in the block length. This also implies that such a code has rate r_0R_0 with high probability.[2] *In other words, a random code from this ensemble meets the Gilbert-Varshamov bound with high probability.*

3 Background on Decoding Concatenated Codes

We now review some approaches to decode concatenated codes that exploit the power of list decoding algorithms when the outer code is an algebraic code like the Reed-Solomon code. The basic idea is to first decode the inner codes, and then use information from this stage to (list) decode the outer code. However, there are several choices available in implementation of this idea when it comes to how much and what kind of information is passed to the outer decoder.

One natural approach is the following. Given a received word $\mathbf{r} = r_1r_2 \ldots r_N$ where r_i corresponds to the encoding of the i'th outer codeword symbol, decode each r_i up to a certain radius producing a list $\mathcal{L}_i$ of up to L inner codewords, or equivalently up to L candidate symbols for the i'th location of the outer codeword. (One way to bound the radius up to which such decoding can be done while producing list size at most L is via the Johnson bound, cf. [6].) If the total fraction of errors is bounded (say by $1/4$ which will be the target in our work here), then a good fraction of the lists $\mathcal{L}_i$ must contain the correct symbol. The list decoding algorithm for the outer Reed-Solomon code, on input the lists $\mathcal{L}_i$, can then determine all such codewords, and we can check to see whether there is a unique codeword within half-the-GV bound from $\mathbf{r}$, and if so output it. If there is no such codeword, then too many errors have occurred; if there is more than one such codeword, then that serves as proof that the minimum distance of the code fell short of the GV bound. Either way, the algorithm's failure gives useful "proven" information.

When concatenating an outer Reed-Solomon code of rate R_0 with (binary) inner codes of relative distance δ, it is an easy computation using the Johnson bound and Reed-Solomon list decoding algorithm of [4] to show that the above strategy can decode up to a fraction

$$\frac{1}{2}(1 - \sqrt{1 - 2\delta + 2\delta/L})(1 - \sqrt{LR_0}) \tag{1}$$

of errors. The above will suffice to give a qualitative result similar to the one we are after, but the rate up to which we can meet the GV bound *and* decode up to half-the-distance leaves room for improvement (even in the order of magnitude; more on this in the discussion at the end of this paper).

[2] Note that the rate is not argued by showing that each of the G_i's will have full rank with high probability. In fact, such a claim is not true. However, if the minimum distance of the overall code is greater than 0, then there must be 2^{tNR_0} distinct codewords.

Therefore, we use a slightly more sophisticated decoding strategy which gives bounds as stated in Theorem 2 below, which is the result of Theorem 3 of [5]. (There is a slight difference in that the result below is stated with possibly different inner codes at the various positions, and also allows a small fraction of the inner codes to not have any guarantee on relative distance. However, these differences are minor and the proof from [5] can be trivially adapted to this setting.) We refer the reader to [5] for actual details of the proof, but just mention the basic idea here. The approach is similar to the above-mentioned one, except that when a list $\mathcal{L}_i$ of possible symbols are returned for the i'th position, the inner decoder also passes for each symbol in the list an associated "weight" or "confidence information" that represents, qualitatively, the likelihood of that symbol being the actual one. The list decoding algorithm for Reed-Solomon codes can handle such weights on the inputs (such weights are usually referred to as "soft information" in coding theory, and the associated decoders that can take advantage of such information are referred to as soft decoding algorithms). The specific weights used in [5] are linear functions that decrease with the distance of the symbol's encoding from the received block r_i. This seems like a reasonable choice but by no means the only conceivable one. Whether there are better ways to set weights that can be exploited to improve the decoding algorithms for concatenated codes is a central question in this area, whose importance is further underscored by the application presented here.

Theorem 2 *[5] Consider a family of binary linear codes where each member of the family is a concatenated code with the outer code a Reed-Solomon code of rate R_0 with block length equal to the size of the underlying field, say Q,[3] and each inner code a (possibly different) binary linear code of dimension $\lg Q$ such that a $(1-\gamma)$ fraction of the inner codes have relative distance at least δ. Then, codes from such a family can be encoded in polynomial time, as well as list decoded in deterministic polynomial time up to a fractional radius of*

$$\frac{1}{2} \cdot \left(1 - \sqrt{1-2\delta}\right) - \sqrt{\delta R_0} - \gamma \, . \tag{2}$$

4 Our Main Result

In this section, we describe how we obtain probabilistic constructions of codes that lie on the GV bound and which can be decoded up to half-the-distance in polynomial time. Our approach is to pick a concatenated code with outer Reed-Solomon code and varying inner codes picked at random. Such a code will lie on the Gilbert-Varshamov bound with high probability as per Theorem 1. We will also pick the outer and inner distances Δ and δ so that the error fraction that can be corrected as per Equation (2) will equal $1/4$ (which is the largest fraction of errors that can be unique decoded for binary codes). Our overall rate will be very close to $(1-\Delta)(1-H(\delta))$. Optimizing over the choice of δ, Δ to maximize the rate will give us our final bound. Note that our result actually guarantees list decoding up to a fraction $1/4$ of errors, while we only need to guarantee unique decoding up to half the distance for our claim. However, since our final codes will be in the low-rate or large-distance regime, $1/4$ is a very good approximation to half the relative distance, and so we don't expect to improve the rate significantly by being careful about this part. Therefore,

[3] This restriction on the block length is not necessary and the claim holds for any field size which is a polynomially growing function of the block length.

we simply set the desired decoding radius to be a fraction $1/4$ of errors. We now state the formal result that follows by combining the statements of Theorems 1 and 2.

Theorem 3 *For every $\varepsilon > 0$ and $0 < \delta < 1/2$, let $r_0 = 1 - H(\delta) - \varepsilon$ and let R_0 satisfy $0 < R_0 \leq \alpha(r_0)/r_0$. For a large enough integer t, let $\mathrm{RS}_t(R_0)$ be the Reed-Solomon code over $\mathrm{GF}(2^t)$ of block length $N = 2^t$, rate R_0, and relative distance $(1 - R_0)$. Consider an ensemble of concatenated codes with $\mathrm{RS}_t(R_0)$ as outer code with varying inner codes of dimension t and block length t/r_0 picked as in Theorem 1. Then:*

1. *A random concatenated code from this ensemble has rate $r_0 R_0$ and relative distance at least $H^{-1}(1-r_0R_0)-\varepsilon$, and thus meets the GV bound, except with probability that is exponentially small in N.*
2. *There is a deterministic polynomial time decoding algorithm that can, except with exponentially small probability, decode a randomly drawn code from such an ensemble up to a fraction of errors equal to*

$$\frac{1}{2} \cdot \left(1 - \sqrt{1-2\delta}\right) - \sqrt{\delta R_0} - \varepsilon \, . \tag{3}$$

Moreover, given a particular choice of the concatenated code, it can be certified in deterministic polynomial time whether or not the above decoding guarantee will be met.

Proof Sketch: Follows from Theorems 1 and 2 using the fact that a binary linear code defined by picking a random $t \times t/r_0$ matrix as its generator matrix where $r_0 = 1 - H(\delta) - \varepsilon$ has relative distance at least δ with overwhelming probability. In particular, except with exponentially small probability (in N), more than a fraction $(1-\varepsilon)$ of the inner codes will have relative distance at least δ. The claim on the decoding radius (3) follows from Theorem 2, and the "certification" property follows since we can check in $\mathrm{poly}(N)$ time exactly how many inner codes have relative distance at least δ. □

Plugging in specific choice of the parameters, specifically δ, R_0, we conclude our main result:

Theorem 4 (Main) *There is a probabilistic polynomial time procedure to construct codes whose rate vs. distance trade-off meets the Gilbert-Varshamov bound with high probability for all rates less than 10^{-4}. Furthermore, these codes can be decoded in polynomial time up to half the relative distance, and in fact this latter decoding property can be "certified", i.e., one can verify in deterministic polynomial time that such decoding will indeed be possible for the constructed code.*

Proof: This follows by plugging into Theorem 3 the following choice of parameters (found by search using a simple program): $\delta = 0.421$, $r_0 = 0.01808$, and $R_0 \leq 0.00623$ (the exact value of R_0 will depend on the desired overall rate which is $r_0 R_0$). The choice of δ and R_0 can be seen to ensure that the decoding radius from Equation (3) is at least $1/4$, so that we are guaranteed to decode up to radius $1/4$ and in particular up to half the relative distance. It can also be checked that the choice of r_0, R_0 satisfy $r_0 R_0 < \alpha(r_0)$, so that the condition for Thommesen's result is met and the codes will meet the GV bound w.h.p. Thus, we have efficiently decodable codes meeting the GV bound for rates up to $0.01808 \times 0.00623 \approx 1.126 \times 10^{-4}$. □

We should mention that we did not attempt to optimize the rate up to which our construction works beyond the order of magnitude (i.e., 10^{-4}). We just remark that optimization similar to

Theorem 4, when performed for the decoding guarantee of Equation (1), yielded an overall rate of about 1.4×10^{-5} (achieved by setting inner rate $r_0 = 0.008179$, the list size for inner decodings $L = 18$, and outer rate $R_0 = 0.001673$). A different approach to improving the rate would be to use a similar algorithm to the one that leads to the bound of Equation (1), but instead of using the Johnson bound to relate the list-decodability of the inner codes to their minimum distance, directly use the trade-off between rate and list-decodability of binary linear codes. Unfortunately, this trade-off is not well-understood, and the bounds for random linear codes (that hold with high probability for most linear codes) are much weaker than those for general random codes (see [2, 3] for a detailed discussion about this and the exact trade-offs). For general random codes, the trade-off that holds with high probability is $r_0 = 1 - H(p)(1 + 1/L)$ where r_0 is the rate, L is the list size, and p is the fraction of errors. However, even assuming such a trade-off can be shown to hold for most linear codes, a similar optimization to that of Theorem 4 still yields an overall rate that is of order of magnitude only 10^{-4}. Therefore, it seems that to improve the rate substantially, say to about 10^{-2}, will need a somewhat sophisticated improvement to our approach.

References

[1] Ilya Dumer. Concatenated codes and their multilevel generalizations. In V. S. Pless and W. C. Huffman, editors, *Handbook of Coding Theory*, Volume II, pages 1911–1988. North Holland, 1998.

[2] Peter Elias. Error-correcting codes for list decoding. *IEEE Transactions on Information Theory*, 37:5–12, 1991.

[3] Venkatesan Guruswami, Johan Håstad, Madhu Sudan, and David Zuckerman. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory*, 48:1021–1034, May 2002. Preliminary version in *Proceedings of the 38th Annual Allerton Conference on Communication, Control and Computing*, pages 603–612, October 2000.

[4] Venkatesan Guruswami and Madhu Sudan. Improved decoding of Reed-Solomon and algebraic-geometric codes. *IEEE Transactions on Information Theory*, 45:1757–1767, 1999.

[5] Venkatesan Guruswami and Madhu Sudan. Decoding concatenated codes using soft information. In *Proceedings of the 17th IEEE Conference of Computational Complexity*, pages 148-157, May 2002.

[6] Venkatesan Guruswami and Madhu Sudan. Extensions to the Johnson bound. *Manuscript*, February 2001. Available at `http://theory.lcs.mit.edu/~madhu/papers.html`.

[7] Christian Thommesen. The existence of binary linear concatenated codes with Reed-Solomon outer codes which asymptotically meet the Gilbert-Varshamov bound. *IEEE Transactions on Information Theory*, 29(6):850–853, 1983.

[8] Victor V. Zyablov and Mark S. Pinsker. List cascading decoding, *Problems of Information Transmission*, 17(4):29–34, 1981 (in Russian), pages 236–240 (in English), 1982.