List Decoding Algorithms for certain Concatenated Codes

Venkatesan Guruswami* Madhu Sudan†

Abstract

We give efficient (polynomial-time) list-decoding algorithms for certain families of error-correcting codes obtained by "concatenation". Specifically, we give list-decoding algorithms for codes where the "outer code" is a Reed-Solomon or Algebraic-geometric code and the "inner code" is a Hadamard code. Codes obtained by such concatenation are the best known constructions of error-correcting codes with very large minimum distance. Our decoding algorithms enhance their nice combinatorial properties with algorithmic ones, by decoding these codes up to the currently known bound on their list-decoding "capacity". In particular, the number of errors that we can correct matches (exactly) the number of errors for which it is known that the list size is bounded by a polynomial in the length of the codewords.

1 Introduction

In this paper we consider constructions of linear codes over fixed size alphabets with very high list-decoding capabilities, i.e., one can efficiently recover a small list of possible codewords when a very large fraction of the symbols are either erased or are in error. We consider a number of errors or erasures which is essentially the best one can hope to recover from. Over the finite field $\mathrm{GF}(q)$ (also denoted as $\mathcal{F}_q$), we consider problems which are of the following form. Suppose one needs to transmit k symbols (over $\mathcal{F}_q$) and one wishes to recover from a fraction $(1 - 1/q - \gamma)$ of errors (this is essentially the best one can hope to recover from, as a random string agrees with any given codeword in $1/q$ fraction of the places). We want to encode the k bits into n symbols over $\mathrm{GF}(q)$ and then transmit the encoded string so that this can be achieved, and our goal is to keep the value of n as small as possible so that the redundancy in the encoding is small (or, equivalently, the rate of the code is high). Note that we also allow the parameter $\gamma = \gamma(n)$ to be a $o(1)$ function which depends on n, and one can present our results as constructing a family of codes $\mathcal{C}_n$ such that a fraction $(1 - 1/q - \gamma_n)$ of errors can be corrected in $\mathcal{C}_n$.

We are also interested in the corresponding question in presence of only erasures, and also in the presence of both errors and erasures. For instance, we can ask the same question above when all but γn of the transmitted symbols are erased, and similarly when e errors and s erasures occur which satisfy $\frac{q}{q-1}e + s \leq (1-\gamma)n$.

Constructions of codes with this kind of performance typically achieve $n = \mathrm{poly}(k/\gamma)$; we are interested in the explicit specification of this polynomial. We also show special interest in the case when the dependence of n is linear in k, so that the resulting code is asymptotically good.

Previous and Related Work. These problems do not seem to have been considered explicitly in the literature. However there certainly are constructions which imply something in our context. Let us first consider the case of erasures. It is shown in [12] that a Reed-Solomon code concatenated with a Hadamard code together with the outer (list) decoder for Reed-Solomon codes of [18, 9] implies that a blocklength $n = \Omega(k^2/\gamma^4)$ suffices to tolerate up to $(1-\gamma)$ fraction of erasures[1]. We will prove a stronger result in this paper using a much simpler approach. We do this by showing that any q-ary code of distance $(1 - 1/q)(1 - \delta)$ can be list decoded from a fraction $(1 - \delta)$ of erasures, and hence it suffices to construct rate k q-ary codes with minimum distance $\frac{(q-1)}{q}(1-\delta)n$ for as small a value of n as possible. Reed-Solomon concatenated with Hadamard code achieves a distance of $(1 - 1/q) \cdot (1 - k/\sqrt{n})$, and this gives $n = O(k^2/\delta^2)$. For the case of asymptotically good codes, by a brute force search for the best inner code along with a Reed-Solomon code as an outer code gives roughly $n = O(k/\delta^3)$, and this was shown by [5]. This construction is sometimes not "explicit" enough, as the construction complexity is not a fixed polynomial independent of δ, and also the entries of the generator matrix cannot be computed in polylogarithmic (in n) time. The first explicit asymptotically good construction achieving this was given by [23] following the work of Justesen [10], and related constructions appear in [20, 21]. Constructions of algebraic-geometric codes imply such codes with $n = O(k/\delta^3)$, but these codes have very high construction complexity. The best constructions with reasonable complexity is due to [3], which achieves $n = O(k/\delta^3)$.

For the case of errors, no such simple connection between the distance of the code and its *efficient* list decodability exists, and constructions are much harder to come by. It is known that a concatenation of Reed-Solomon codes with Hadamard codes can achieve $n = \mathrm{poly}(k/\gamma)$ to decode up to a fraction $(1 - 1/q - \gamma)$ of er-

*MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139. Email: venkat@theory.lcs.mit.edu. Research supported by an IBM Research Fellowship.

†MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139. Email: madhu@mit.edu. Supported in part by an MIT-NEC Research Initiation Award, a Sloan Foundation Fellowship and NSF Career Award CCR-9875511.

[1]The reader unfamiliar with the definitions of Reed-Solomon codes, Hadamard codes or concatenation, can find a brief description in Section 2.

rors [19] (this is also implicit in [12]), but even for this code the dependence of n on γ was not optimized. The situation is worse for asymptotically good codes. In Justesen's original paper [10], he also gives an algorithm to decode his code construction to half the minimum distance. Note that unique codeword decoding immediately implies that we cannot hope to recover from more than $\frac{1}{2}(1-1/q)$ fraction of errors as any q-ary code (with exponentially many codewords) has distance less than $(1-1/q)$. Moreover, the original binary codes due to Justesen [10] have a distance only about 0.11, implying decoding up to about 0.055 fraction of errors. It turns out, however, that the binary code construction of Weldon [23] has distance $1/2-\varepsilon$, and can be decoded up to half the minimum distance using similar ideas. This implies decoding up to a fraction $1/4-\gamma$ of errors for the binary case, and a similar result can be shown for the q-ary case. Beyond this error radius, one needs the power of list decoding, and no list decoders to handle such high noise seem to be known for asymptotically good codes. We are able to give two constructions, one resorting to algebraic-geometric codes, and the other a simpler one by concatenating a Reed-Solomon code with *any* asymptotically good inner code with good distance properties. Moreover, our decoding algorithms can handle both errors and erasures.

Our Results.

[Recovering from erasures:] We show that, using a Reed-Solomon code concatenated with the Hadamard code, we can reconstruct, in polynomial time, a list of all candidate codewords when only a fraction $\gamma > 0$ of the transmitted symbols are received, provided the blocklength of the code is $n \geq (k/\gamma)^2$. We also show, following the approach in [6], that such codes which are efficiently list decodable from a $(1-\gamma)$ fraction of erasures exist provided $n \geq k^2/\gamma$ (in fact a random code has such a property), though we know of no explicit construction achieving this. Our result implies, in the terminology of [6, 12], an explicit construction of a binary code, for any $\varepsilon > 0$, with exponentially (2^{N^β} for some $\beta > 0$) many codewords such that given *any* $N^{1/2+\varepsilon}$ of the N bits of the codeword, the list of codewords consistent with these bits can be efficiently recovered.

[Recovering from errors and erasures:] We first establish a combinatorial result proving an upper bound on the list size possible when decoding from a certain number of errors and erasures. Our result is analogous to the Johnson bound (see also [8]), and reduces to the one in [8] in presence of only errors; our proof is in fact simpler than the one in [8] even though our result is more general. This places limits on the radius to which one can (currently) hope to list decode in polynomial time, and restricts $qe/(q-1)+s$, where e, s are the number of errors and erasures respectively, to be at most some quantity that is a function of the minimum distance of the code. We then give polynomial time list decoding algorithms for certain concatenated codes to recover from e errors and s erasures provided $qe/(q-1)+s \leq (1-\gamma)n$, and express the blocklength n required to achieve this as a function of k and γ. The specific results we obtain are the following:

(a) A decoding algorithm for Reed-Solomon concatenated with Hadamard code whose blocklength is $n = O(k^2/\gamma^4)$. Our decoding algorithm is novel and uses "soft information" that is passed by the inner decoder in order to decode the outer Reed-Solomon code. Underlying this procedure is a powerful weighted polynomial time reconstruction algorithm due to [9], which will probably find other applications as well (for example, recent work in [13] uses this to give decoding algorithms for Reed-Solomon codes that appear to be superior to Forney's GMD algorithm [5]). We also use the Linear Programming based bounds for codes to provide evidence that this (quartic) dependence of n on γ is probably the best possible, given current (combinatorial) techniques to bound the list size of candidate codewords by a polynomial.

(b) For asymptotically good codes, we give an algorithm for list decoding algebraic-geometric codes concatenated with Hadamard code, and the required blocklength is $n = O(k/(\gamma^6 \log 1/\gamma))$.

(c) The code from (b) above relies on algebraic-geometric codes which have quite a high construction complexity and moreover the decoding algorithm has to assume complicated subroutines over function fields, so we also give a polynomial time decoding algorithm for the alternative simpler code obtained by concatenating a Reed-Solomon code with *any* inner code with large enough minimum distance. This code is constructible in polynomial time and satisfies the required error-erasure correction property with a blocklength $n = O(k/\gamma^8)$.

The above concatenated codes have the best known blocklength (up to constant factors) for codes with a given rate and minimum distance, and our results are therefore qualitatively significant in that no constructions of codes with better information-theoretic list decoding capabilities are known. More precisely, consider the task of constructing codes where decoding $(1-\frac{1}{q}-\gamma)$ fraction of errors leads to polynomial size lists. The smallest block length, given k and γ, for which we know of constructions of codes with such properties, is at least a constant fraction of the blocklength that we can demonstrate; however we do so with a polynomial time decoding algorithm! Another fact of interest is that for the cases (a) and (b) above, the radius to which we are able to list decode in polynomial time matches (exactly) the bound on the radius for which it is known that the list size is bounded by a polynomial in the blocklength.

Organization. In Section 2 we describe the codes we shall use and the high level idea behind our decoding algorithms. Section 3 proves an upper bound on the list size when decoding q-ary codes from both errors and erasures. Our code constructions and decoding algorithms for erasures as well for the errors and erasures case are described in detail in Section 5. Section 6 describes a couple of applications of our codes to problems in complexity theory.

2 Basic Outline

We start by identifying some standard parameters of codes. A linear error-correcting code $\mathcal{C}$ over a q-ary alphabet of block length n is a linear subspace of $\mathrm{GF}(q)^n$. Its dimension, typically denoted by k, is the information content of the code. The minimum distance of the code, is the minimum Hamming distance between any two distinct members of the code. Some of the standard constructions of codes we deal with are described below.

- (Generalized) Reed Solomon Codes: These are obtained by viewing the message as specifying a $k-1$-dimensional polynomial; and evaluating it at n distinct points of $\mathrm{GF}(q)$. It needs $n \leq q$, and often we will use $n = q$. The minimum distance of this code is $n-k+1$.
- Hadamard Codes: These are obtained by viewing the message as the coefficients of a homogeneous degree 1 polynomial in k variables, and evaluating it at all inputs: Thus it gives a code of block length q^k with minimum distance $q^k - q^{k-1}$.
- Algebraic Geometry codes: We will not be able to describe the codes here; however we can describe their parameters. They are constructible for any q that is an even power of a prime, and can achieve a distance of at least $n-k+1-n/(\sqrt{q}-1)$.
- Concatenated codes [5]: These are codes obtained by combining an "outer" code over a q^k-ary alphabet with an "inner"

code of dimension k over a q-ary alphabet. The combined codeword corresponding to a given message is obtained by first encoding the message using the outer code, and then encoding each symbol of the resulting string by the inner code. The resulting code has block length that is a product of the two block lengths and distance that is the product of the two distances.

All codes for which we give efficient list decoding algorithms from high noise are based on the idea of code concatenation. The *outer codes* we use will be algebraic codes like Reed-Solomon or Algebraic-geometric codes. The specific concatenated codes we give decoding algorithms for are:

(a) Reed-Solomon code concatenated with Hadamard code

(b) Reed-Solomon code concatenated with any q-ary inner code which has relative distance very close to $(1 - 1/q)$

(c) Algebraic-geometric code concatenated with Hadamard code

These codes are by no means new to our paper and have been (especially the Reed-Solomon concatenated with a Hadamard code) often considered in the past (for instance in [1, 12, 19]). The novel aspect of our work is in the (list) decoding algorithms we give to decode these codes in the presence of a very large number of errors and erasures. Our decoding algorithms for these concatenated codes begin by a decoding of the inner code which can be accomplished by brute-force since the total number of inner codewords is of polynomial complexity. Information from the inner decoding is then passed onto the outer decoder which then completes the decoding. The information passed from the inner decoder can be of many kinds: one possibility could be to just pass the most likely inner codeword for each of the outer codeword positions. This typically is too weak as the inner decoder is forced to make a single hard decision on its codeword which may lead the outer decoder astray. Our inner decoding algorithms work in one of the following two ways:

1. Returns a (small) list of possible codewords; this list is typically the list of all codewords within a certain distance of the inner received word. Thus the outer decoder has for each codeword position, a list of possibilities, and needs to list decode from this information. This is possible for Reed-Solomon and algebraic-geometric codes [18, 9]. One important aspect here is that the list size must be small, and we need combinatorial bounds on the maximum possible codewords with a certain number of errors and erasures from a received (inner) word. Such a bound, which is of independent interest, is stated and proved in the next section.
2. Returns weights for each possible inner codeword; the weight corresponding to an inner codeword is a measure of the confidence which the inner decoder has in the fact that it was that codeword which was transmitted. It is reasonable that the weight a certain inner codeword receives should in some sense decrease with its distance from the received inner word, and as we shall this will be indeed be the case. The outer decoder then uses this “soft decision” reliability/confidence information in its decoding. An algorithm that can use soft decision information was given for Reed-Solomon codes in [9]; a similar algorithm actually exists for algebraic-geometric codes as well; these algorithms are detailed in Appendix B.

3 A bound on list size in presence of errors and erasures

The aim of this section is to state and prove an upper bound on the number of codewords possible when list decoding from e errors and s erasures provided e, s satisfy some condition with respect to the distance d of the code. This bound will place limits on the number of errors and erasures for which one is guaranteed a polynomial list size, and can therefore hope for efficient list decoding algorithms. The bound will also be important to one of our decoding algorithms (specifically the one which decodes a Reed-Solomon code concatenated with any inner code, see Theorem 10), where we need an upper bound on the number of inner codewords that can exist with a certain number of errors and erasures. The result below generalizes a similar bound in [8] and specializes to that bound for the errors only case, although our proof is different and simpler.

Theorem 1 *For a q-ary code of blocklength n and distance $d = (1-1/q)(1-\delta)n$, and for any received word with $s = \sigma n$ erasures, the number of codewords differing from the received word in at most e places, where $qe/(q-1) + s = (1-\gamma)n$, is at most $\frac{(1-\sigma)(1-\delta)}{\gamma^2-(1-\sigma)\delta}$, provided $\gamma > \sqrt{(1-\sigma)\delta}$.*

Proof: Let $y \in \mathcal{F}_q^{n-s}$ be a received word with s erasures, say the last $s = \sigma n$ positions are erasures. Also assume without loss of generality that y is the symbol q repeated $(n-s)$ times (we let the field elements to be in one-one correspondence with the integers $1, 2, \ldots, q$). Let $C_1, C_2, \ldots, C_m$ be all the codewords which differ from y in at most e places, where $qe/(q-1) + s = (1-\gamma)n$. Our goal is to get an upper bound on m provided γ is large enough.

We associate with the received word y and each codeword C_i an nq-dimensional real vector. The vector is to be viewed as having n blocks each having q components (the n blocks correspond to the n codeword positions). For $1 \le l \le q$, denote by $\hat{e}_l$ the q-dimensional unit vector with 1 in the lth position and 0 elsewhere. For $1 \le i \le m$, the vector $\vec{v_i}$ associated with the codeword C_i has in its jth block the components of the vector $\hat{e}_{C_i[j]}$ ($C_i[j]$ is the jth symbol of C_i, treated as an integer between 1 and q). The vector $\vec{r}$ associated with the received word y is defined similarly for the first $(n-s)$ blocks, and the last s blocks of $\vec{r}$ (which correspond to the erased positions) will have $1/q$ in every position. (The intuition behind this is the following: the vector $(1/q, 1/q, \ldots, 1/q) \in \mathcal{R}^q$ is the centroid of the q points corresponding to the q field elements, and hence associating this vector with a position amounts to saying that we have absolutely no idea about the value at this position, or in other words this position was erased.)

The key quantity we will estimate now is the sum

$$S = \sum_{1 \le j,k \le m} \langle (\vec{v_j} - \vec{r}), (\vec{v_k} - \vec{r}) \rangle .$$

Let us first give a lower bound on S. The dot product above can be written as the sum of the dot products over the n blocks. We ignore the contribution from the s erased positions (which is clearly non-negative); for blocks p, $1 \le p \le (n-s)$, let N_p denote the number of vectors $\vec{v_j} - \vec{r}$ which are non-zero, and let $N_{p\beta}$, for $1 \le \beta \le (q-1)$, denote the number of those vectors which are of the form $0^{\beta-1}10^{q-\beta-1}(-1)$; clearly $\sum_\beta N_{p\beta} = N_p$. The contribution to S from the q columns in block p is

$$N_p^2 + \sum_{\beta=1}^{q-1} N_{p\beta}^2 \ge \left(\frac{q}{q-1}\right) N_p^2 .$$

Now, $\sum_{p=1}^{n-s} N_p = \sum_{j=1}^{m} e_j = m\bar{e}$ where e_j is the number of places C_j differs from y, and $\bar{e}$ is the average number of errors over the codewords $C_1, C_2, \ldots, C_m$. Hence $\sum_p N_p^2 \ge (m\bar{e})^2/(n-s)$, and thus we get

$$S \ge \frac{q}{q-1}\left(\frac{m^2\bar{e}^2}{n-s}\right). \tag{1}$$

Now for the upper bound on S. Let us consider a fixed pair of vectors $(\vec{v_j} - \vec{r})$ and $(\vec{v_k} - \vec{r})$. If $j = k$, then one easily computes

$$\langle(\vec{v_j}-\vec{r}),(\vec{v_j}-\vec{r})\rangle = 2e_j + \frac{(q-1)}{q}\cdot s. \qquad (2)$$

When $j \neq k$, if d_{jk} is the distance between the codewords C_j, C_k (note $d_{jk} \geq d$), one can show that

$$\begin{aligned}\langle(\vec{v_j}-\vec{r}),(\vec{v_k}-\vec{r})\rangle &= \langle\vec{v_j},\vec{v_k}\rangle + \langle\vec{r},\vec{r}\rangle - \langle\vec{v_j},\vec{r}\rangle - \langle\vec{v_k},\vec{r}\rangle \\ &= \left(1-\frac{1}{q}\right)s + e_j + e_k - d_{jk} \\ &\leq \frac{q-1}{q}s + e_j + e_k - d. \qquad (3)\end{aligned}$$

From Equations (2) and (3), we get

$$S \leq 2m^2\bar{e} + \frac{q-1}{q}m^2 s - m(m-1)d \qquad (4)$$

The proposition follows (after some algebraic manipulation) from (1) and (4). □

4 Distance properties of some concatenated codes

From Theorem 1, it is clear that in order to list decode from a large amount of noise, we would like the underlying code to have large minimum distance, so that we will, to begin with, at least have the combinatorial guarantee that size of the list to be output will be small. We now quantify the distance properties of the main concatenated codes we will use; namely we express the blocklength n in terms of the rate k and the distance parameter $\delta = 1-qd/(q-1)$.

Proposition 2 *For every $k, \delta > 0$, there is an explicitly specified q-ary code, denoted $\mathcal{C}_{\mathrm{RS-Had}}$, obtained by concatenating a Reed-Solomon code with a Hadamard code, which has rate k, relative distance $d = (1-1/q)(1-\delta)$, and blocklength $n = O(\frac{k^2}{\delta^2 \log^2(1/\delta)})$.*[2]

Proof: Let the code $\mathcal{C}_{\mathrm{RS-Had}}$ be obtained by concatenating a rate k/m Reed-Solomon code over $GF(q^m)$ with the Hadamard code associated with $GF(q^m)$. The combined rate is clearly k and the blocklength is $n \triangleq (q^m)^2$. The relative minimum distance of the code is $(1-1/q)(1-\frac{k/m-1}{q^m})$, and thus $n = O((\frac{k}{\delta \log 1/\delta})^2)$ as desired. □

The above construction has good (in fact the best possible) dependence of the blocklength on δ, but is, however, not asymptotically good (i.e n is not linear in k). We next describe two asymptotically good constructions.

Proposition 3 *For every $k, \delta > 0$, there exists an explicitly specified q-ary code, denoted $\mathcal{C}_{\mathrm{AG-Had}}$, that is obtained by concatenating an (appropriate) algebraic-geometric code with a Hadamard code, and which has rate k, relative distance $d = (1-1/q)(1-\delta)$, and blocklength $n = O(\frac{k}{\delta^3 \log 1/\delta})$.*

Proof: Let the code $\mathcal{C}_{\mathrm{AG-Had}}$ be obtained by concatenating a rate k/m, blocklength n_0, algebraic-geometric code over $GF(q^m)$ with the Hadamard code associated with $GF(q^m)$. The distance of the outer code is $n_0 - k/m - g + 1$ where g is the genus of the underlying function field; we will use function fields with $g = n_0/(q^{m/2}-1)$ (such constructions are given, for instance, in [22, 15, 7]). The rate of the concatenated code is clearly k and the blocklength is $n \triangleq n_0 q^m$. The relative minimum distance of the code is at least $(1-1/q)\big(1-\frac{k/m-1}{n_0}-1/(q^{m/2}-1)\big)$, and this gives $n = O(\frac{k}{\delta^3 \log 1/\delta})$. □

Proposition 4 *For every $k, \delta, \rho > 0$, there is a q-ary code, denoted $\mathcal{C}_{\mathrm{RS-GoodInner}}$, that is obtained by concatenating a Reed-Solomon code with any relative distance $(1-1/q)(1-\rho)$ q-ary inner code, and which has rate k, distance $d = (1-1/q)(1-\delta)$, and blocklength $n = O(\frac{k}{\delta\rho^2})$. Moreover such a code can be constructed in polynomial (in n) time.*[3]

Proof: We use the same construction as in Proposition 2 except we use, instead of the Hadamard code, an $[n_1, m, d_1]_q$ inner code where $d_1 = (1-1/q)(1-O(\delta))$ with $m = \Omega(\delta^2 n_1)$ (such a code exists by the Gilbert-Varshamov bound and can be found in $2^{O(n_1)} = \mathrm{poly}(n)$ time by searching in the Wozencraft's ensemble of randomly shifted codes [23]). The blocklength is $n = n_1 q^m = O(n_1 \cdot \frac{k/m}{\delta}) = O(k/\delta^3)$. □

Remark: The above codes have, up to constant factors, the smallest blocklength possible for a given value of k and δ. In addition they have this concatenated structure with a nice algebraic outer code, and hence we will be able to design list decoding algorithms for these that can handle a large amount of noise. In particular, for the codes $C_{\mathrm{RS-Had}}$ and $C_{\mathrm{AG-Had}}$ we will be able to decode up to exactly the radius specified in the combinatorial bound of Theorem 1. This makes our results qualitatively significant in that no constructions of codes with better information-theoretic list decoding capabilities are known than the ones we are able to achieve algorithmically.

5 Performance of the Decoding algorithms

This section formally describes our code constructions and decoding algorithms and quantifies their error-erasure correction performance.

5.1 Erasure Codes

The following simple but useful fact also follows from Theorem 1, but we include an easier proof below.

Lemma 1 *If $\mathcal{C}$ is a q-ary code of blocklength n and minimum distance d, then for any received word with at most $\frac{q}{q-1}d$ erasures, the list of possible codewords consistent with the received word is of size at most $\frac{q^2}{q-1}d$.*

Proof: Let y be a received word with $s \leq qd/(q-1)$ erasures. Suppose $C_1, C_2, \ldots, C_M$ are the distinct codewords of $\mathcal{C}$ consistent with y, i.e they agree with y in all the non-erased positions. By the distance property of $\mathcal{C}$, these codewords must differ from each other in at least d places in the s erased positions. Projecting the codewords $C_1, C_2, \ldots, C_M$ to the erased positions, we get a code of blocklength $s \leq qd/(q-1)$ with distance d. By a standard coding theory bound, any such code can have at most qs codewords, implying $M \leq \frac{q^2}{q-1}d$. □

Corollary 1 *Any q-ary code of relative minimum distance $(1-1/q)(1-\gamma)$ can be efficiently list decoded from erasures as long as the fraction of erasures is at most $(1-\gamma)$.*

Proof: By Lemma 1, we know that for any received word with less than a fraction $(1-\gamma)$ of erasures, the list of possible codewords is of polynomial (in fact linear) size. Now to recover from erasures, a small list size implies efficient decodability, since recovering from erasures only involves finding all possible solutions to a linear system of equations, and if the number of solutions is guaranteed to be

[2] Here and by a bound like $n = O(k^a/\delta^b)$ we mean the following: there exists a constant c such that for every k and every $\delta > 0$, there is a distance $(1-1/q)(1-\delta)$ q-ary code of rate k and blocklength at most ck^a/δ^b.

[3] Actually we can even explicitly specify such a code by using a different code from the Wozencraft's ensemble of codes for the various outer codeword positions (see, for instance, [10, 23]).

polynomial in number, then they can certainly be found and output in polynomial time. □

Hence one way to construct codes that handle a large number of erasures is to construct codes with very large minimum distance.

Theorem 5 *For any finite field $\mathcal{F}_q$ and for any integer k and $\gamma > 0$, there exists an explicitly specified code which encodes k symbols over $\mathcal{F}_q$ into n symbols over $\mathcal{F}_q$ where $n = O(k/\gamma^3)$, such that the list of possible codewords can be recovered in polynomial time when up to a fraction $(1-\gamma)$ of the symbols in the received word are erased.*[4]

Proof: By Corollary 1, we only need to construct a code with rate k, blocklength n and minimum distance $(1-\frac{1}{q})(1-\gamma)n$. As shown by Alon *et. al.* [3], such a code can be constructed in polynomial in n time (with the exponent being independent of γ) provided $n = \Omega(k/\gamma^3)$. We could have similarly used the code $\mathcal{C}_{\mathrm{AG-Had}}$ from Proposition 3. □

While the construction of the previous theorem has linear dependence of n on k (and hence the code family is asymptotically good), we would also like constructions with better dependence on γ. The Gilbert-Varshamov bound shows the **existence** of codes with $n = O(k/\gamma^2)$ (in fact a random code with such a value of n has the required property), but we know of no explicit way of constructing such codes. We now present an explicit construction with better than a γ^{-3} dependence at the expense of worse dependence of n on k (hence this construction is not asymptotically good).

Theorem 6 *For any prime power q, and any k, γ, the statement of Theorem 5 holds with $n = O(k^2/\gamma^2 \log^2(1/\gamma))$.*

Proof: By Proposition 2 and Corollary 1, it follows that the code $\mathcal{C}_{\mathrm{RS-Had}}$ has this property. □

The quadratic dependence of n on γ is unavoidable using just the "distance based" approach of Corollary 1. This is because the McEliece-Rodemich-Rumsey-Welch upper bound [16] on the rate of codes implies that a q-ary code relative minimum distance $(1-1/q)(1-\delta)$ can have relative rate k/n at most $O(\delta^2 \log(1/\delta))$. We now prove an existential result for codes that achieves a better dependence of n on γ; we have no idea, however, on how to construct such codes deterministically in polynomial time. In fact, Alon [2] has pointed out that an explicit construction of erasure codes which beat the quadratic dependence of n on γ is probably difficult as it would imply improvements on the bipartite Ramsey problem; specifically it would give an explicit construction of an $N \times N$ matrix over $\mathrm{GF}(2)$ with no monochromatic $p \times p$ submatrix for p much smaller than $N^{1/2}$, and it is currently not known how to achieve this.

Theorem 7 *For any finite field $\mathcal{F}_q$ and for any integer k and $\gamma > 0$, there is a linear code over $\mathcal{F}_q$ with rate k and blocklength n where $n = O(k^2/\gamma)$, such that whenever up to a fraction $(1-\gamma)$ of the symbols in the received word are erased, the list of possible codewords can be recovered in polynomial time.*

Proof: The proof is by the probabilistic method and follows the approach in [6]. We show that a random linear code (picked using a random $n \times k$ generator matrix G with $n = \Theta(k^2/\gamma)$) has the required property with high probability. Note that we only need to prove that the list of candidate codewords is of polynomial size whenever at least a fraction γ of the codeword is specified, and then efficient list decoding follows simply by finding all solutions of a linear system of equations.

Suppose γn positions of a codewords are specified as a vector $y \in \mathcal{F}_q^{\gamma n}$. This gives a linear system $G'x = y$ where G' is the $\gamma n \times k$ matrix obtained by picking rows of G corresponding to the γn non-erased positions. The list size of codewords consistent with y is precisely the number m of solutions $x \in \mathcal{F}_q^k$ of the this linear system. Clearly, $m = q^{k-\mathsf{rank}(G')}$, and hence we need to argue that, with high probability, *every* $\gamma n \times k$ sub-matrix of G has "high" rank, say a rank at least $k - c\log_q n$ for some constant $c > 0$. The number of such sub-matrices is $\binom{n}{\gamma n}$, and the number of subspaces Γ of $\mathcal{F}_q^k$ of rank less than $k - c\log_q n$ is at most q^{k^2}. The probability that a specific $\gamma n \times k$ sub-matrix G' has row span contained in a specific subspace Γ of rank less than $k - c\log_q n$, is at most $(q^{-c\log_q n})^{\gamma n}$. Hence the overall probability of the code not having the required property is, by union bound, at most

$$\begin{aligned}\binom{n}{\gamma n} \cdot q^{k^2} \cdot n^{-c\gamma n} &\leq (\frac{e}{\gamma})^{\gamma n} \cdot q^{k^2} \cdot n^{-c\gamma n} \\ &\leq (\frac{qe}{\gamma n^c})^{\gamma n} \text{ (provided } k^2 \leq \gamma n) \\ &< 1.\end{aligned}$$

Such a code therefore exists and the proof is complete. □

5.2 Decoding from errors and erasures

Theorem 8 *For an $[n,k,d]$ code $\mathcal{C}_{\mathrm{RS-Had}}$ over $\mathrm{GF}(q)$ with $d = (1-1/q)(1-\delta)$, there is a polynomial time list decoding algorithm for e errors and $s = \sigma n$ erasures as long as*

$$\frac{qe}{q-1} + s \leq n\big(1 - \sqrt{(1-\sigma)\delta}\big) - O(1)\ .$$

Corollary 2 *For any finite field $\mathcal{F}_q$ and for any integer k and $\gamma > 0$, there exists an* explicitly specified *linear code over $\mathcal{F}_q$ of rate k and blocklength n where $n = O(k^2/\gamma^4)$, such that for any received word y with s erasures, the list of all codewords differing from y in at most e places can be found in polynomial time, provided $q/(q-1)e + s \leq (1-\gamma)n$.*

Proof: Follows from Proposition 2 and Theorem 8. □

Proof of Theorem 8: Before we begin proving the Theorem note that this matches the combinatorial bound for list decoding proved in Theorem 1. Recall that $\mathcal{C}_{\mathrm{RS-Had}}$ is constructed by concatenating an outer $[n' = q^m, k/m, n' - k/m + 1]_{q^m}$ Reed-Solomon code with the Hadamard code associated with $\mathrm{GF}(q^m)$, and it has blocklength $n = n'^2$ and distance $d = (1-1/q)(1-\delta)$ with $\delta = \frac{k/m-1}{n'}$. Now let y be a received word with $s = \sigma n$ erasures in all, we would like to obtain a list of all codewords in $\mathcal{C}_{\mathrm{RS-Had}}$ that differ from y in at most e places. For $1 \leq i \leq n'$, denote by y_i the portion of y in block i of the codeword (i.e the portion corresponding to the encoding of the i^{th} symbol of the outer code), and let s_i be the number of erasures in y_i (where $\sum_{i=1}^{n'} s_i = s$). The n' codewords of the inner Hadamard code are in one-to-one correspondence with the n' elements $\alpha_1, \alpha_2, \ldots, \alpha_{n'}$ of the field $\mathrm{GF}(q^m)$ (which are viewed as m-tuples over $\mathrm{GF}(q)$). For $1 \leq i, j \leq n'$, let e_{ij} be the number of positions where y_i differs from α_j, and define the *weight* w_{ij} as:

$$w_{ij} \triangleq (1 - \frac{s_i}{n'} - \frac{q}{q-1} \cdot \frac{e_{ij}}{n'}).$$

One key property of these weights, proved in Corollary 7 in Appendix A is that, for each i,

$$\sum_j w_{ij}^2 \leq (1 - \frac{s_i}{n'}). \tag{5}$$

[4] Both the construction of the code and the list decoding can be performed in time which strongly polynomial in both n and $1/\gamma$.

These weights will be the "soft information" passed to the outer decoder for Reed-Solomon codes. In order to exploit this information, we will use as outer decoder, a weighted polynomial reconstruction algorithm presented in [9] (see also Proposition 16 of Appendix B). More precisely, the decoder is given weights w_{ij} on pairs (α_i, α_j) of field elements, and the algorithm can find, in $\mathrm{poly}(n', 1/\varepsilon)$ time, a list of all outer codewords $\mathrm{RS}(p)$, which correspond to degree $(k/m-1)$ polynomials p over $\mathrm{GF}(q^m)$, that satisfy

$$\sum_{i=1}^{n'} w_{i,\tilde{p}(i)} > \sqrt{(k/m-1)\sum_{1\le i,j\le n'} w_{ij}^2} + \varepsilon \max_{ij} w_{ij}$$

where $\tilde{p} : [n'] \to [n']$ is defined by $\tilde{p}(i) = j$ iff $p(\alpha_i) = \alpha_j$.

For our definition of weights, using Equation (5), the decoding algorithm can thus retrieve all codewords corresponding to polynomials p for which

$$\sum_{i=1}^{n'}(1 - \frac{s_i}{n'} - \frac{q}{q-1}\cdot\frac{e_{i,\tilde{p}(i)}}{n'}) > \sqrt{(k/m-1)\sum_{i=1}^{n'}(1-\frac{s_i}{n'})} + \varepsilon,$$

or, equivalently, one can find all codewords at a distance e from the received word y provided

$$\begin{aligned} n' - \frac{s}{n'} - \frac{qe}{(q-1)n'} &> \sqrt{(\frac{k}{m}-1)(n'-\frac{s}{n'})} + \varepsilon \text{ or} \\ \frac{qe}{q-1} + s &< n\Big(1 - \sqrt{\frac{k/m-1}{n'}(1-\frac{s}{n})} - \frac{\varepsilon}{\sqrt{n}}\Big) \\ \Longleftarrow \quad \frac{q}{q-1}e + s &\le n\Big(1 - \sqrt{(1-\sigma)\delta}\Big) - O(1) \end{aligned}$$

provided we pick $\varepsilon \le 1/\sqrt{n}$. □

Remark: The (quartic) dependence of n on γ in Corollary 2 seems to be the best one can currently hope for. Current combinatorial bounds guarantee a small list size as long as γ in Corollary 2 is of the order of $\sqrt{\delta}$ where $(1-1/q)(1-\delta)$ is the relative minimum distance of the code. Beyond this radius it is unknown if the number of codewords can always be bound by a polynomial. By the McEliece-Rodemich-Rumsey-Welch upper bound [16], for codes with such high minimum distance, we must have $k/n = O(\gamma^4 \log(1/\gamma))$. For completeness sake, we include in Appendix C, a proof that any binary code with minimum distance $d = n/2 - c\sqrt{n}$ for any $c < 1/2$ can have only polynomially many codewords; this implies that when $\gamma = n^{-1/4}$, we must have $k = O(\log n)$, and this shows that the γ^{-4} dependence of n matches the performance possible given the best known combinatorial bounds on list decodability.

The dependence of n on k in $\mathcal{C}_{\mathrm{RS-Had}}$ is quadratic, however, and hence the code family constructed above is not asymptotically good. We next provide a list decoding algorithm for $\mathcal{C}_{\mathrm{AG-Had}}$ that matches the bound of Theorem 1, and then also give a good list decoding algorithm for the simpler construction $\mathcal{C}_{\mathrm{RS-GoodInner}}$. These results will prove that one can indeed construct asymptotically good codes which can correct from such large fractions of errors and erasures as we are interested in, with only a moderate worsening of the dependence of n on γ.

Theorem 9 *For an $[n,k,d]$ code $\mathcal{C}_{\mathrm{AG-Had}}$ over $\mathrm{GF}(q)$ with $d = (1-1/q)(1-\delta)$, there is a polynomial time list decoding algorithm for e errors and $s = \sigma n$ erasures as long as $qe/(q-1) + s \le n\big(1 - \sqrt{(1-\sigma)\delta}\big) - O(1)$.*

Corollary 3 *For any finite field $\mathcal{F}_q$ and for any integer k and $\gamma > 0$, there exists an* explicitly specified *linear code over $\mathcal{F}_q$ of rate k and blocklength n where $n = O(\frac{k}{\gamma^6 \cdot \log(1/\gamma)})$, such that for any received word y with s erasures, the list of all codewords differing from y in at most e places can be found in polynomial time, provided $q/(q-1)e + s \le (1-\gamma)n$.*

Proof of Theorem 9 *(Sketch)*: Recall that $\mathcal{C}_{\mathrm{AG-Had}}$ is constructed by concatenating an outer $[n_0, k/m, d_0]_{q^m}$ algebraic-geometric code with the Hadamard code associated with $\mathrm{GF}(q^m)$, and it has blocklength $n = n_0 q^m$, rate k and distance $d = (1-1/q)(1-\delta)$ where $\delta = 1 - d_0/n_0$.

The decoding algorithm is exactly similar to the one in Theorem 8 except that instead of a weighted polynomial reconstruction routine, one uses a weighted version of the decoding algorithm of [9] for algebraic-geometric codes (stated formally in Appendix B). Using exactly the same definition of weights as earlier and arguing as in Theorem 8, we conclude that, for any $\varepsilon > 0$, we can find all codewords at a distance e from a received word y with s erasures provided

$$\begin{aligned} n_0 - \frac{s}{q^m} - \frac{q}{q-1}\cdot\frac{e}{q^m} &> \sqrt{(n_0-d_0)(n_0-\frac{s}{q^m})} + \varepsilon \\ \Longleftarrow \quad \frac{q}{q-1}e + s &\le n(1 - \sqrt{(1-\sigma)\delta}) - O(1) \end{aligned}$$

provided we choose $\varepsilon \le 1/q^m$. □

Caveat: When we claim polynomial time decodability in the above theorem, this is valid only under some assumptions about the function field underlying the algebraic-geometric code which we use as the outer code (say the Garcia-Stichtenoth codes of [7]). We next present a decoding algorithm for the simpler construction of $\mathcal{C}_{\mathrm{RS-GoodInner}}$ from Proposition 4.

Theorem 10 *For an $[n,k,d]_q$ code $\mathcal{C}_{\mathrm{RS-GoodInner}}$, where $d = (1-1/q)(1-\delta)(1-\rho)$, for every $\varepsilon > 0$, one can list decode in $\mathrm{poly}(n, 1/\varepsilon)$ time from e errors and $s = \sigma n$ erasures as long as*

$$\frac{qe}{q-1} + s \le n\Big(1 - \sqrt{(1+\varepsilon)\rho} - \sqrt{\frac{\delta(1-\sigma)}{\varepsilon\rho}}\Big).$$

Proof: Recall that $\mathcal{C}_{\mathrm{RS-GoodInner}}$ with the specified parameters is obtained by concatenating an outer $[n_0, k/m, d_0]_{q^m}$ Reed-Solomon code where $n_0 = q^m$ and $d_0 = n_0 - k/m + 1 = (1-\delta)n_0$, with any inner code over $\mathrm{GF}(q)$ that has blocklength n_1, rate m and minimum distance $(1-1/q)(1-\rho)n_1$.

The decoding algorithm will work as follows. The received word y (which has $s = \sigma n$ erasures) can be divided into n_0 blocks y_i corresponding to the n_0 outer codeword positions. For $1 \le i \le n_0$, let s_i denote the number of erasures in y_i, with $\sum_i s_i = s$. For each block i, the inner decoder, by going over all inner codewords (since there are $q^m = n_0$ inner codewords, we can do this in polynomial time), outputs a list $\mathcal{L}_i$ of all (inner) codewords that differ from y_i in at most e_i places where e_i is defined so that $qe_i/(q-1) + s_i = (1-\zeta)n_1$ for some $\zeta > \sqrt{\rho}$; by Theorem 1 the size ℓ_i of each $\mathcal{L}_i$ is at most $(1 - s_i/n_1)(1-\rho)/(\zeta^2-\rho)$;

The inner decoder thus passes to the outer decoder a list of at most $L = \sum_{i=1}^{n_0} \ell_i \le \frac{(1-\sigma)(1-\rho)n_0}{(\zeta^2-\rho)}$ points $\{(x_i, z_{ij}) : 1 \le i \le n_0, 1 \le j \le \ell_i\}$ (here $x_1, x_2, \ldots, x_{n_0}$ are the elements of the field $\mathrm{GF}(q^m)$ and $z_{ij} \in \mathrm{GF}(q^m)$). Using the decoding algorithm in [9], we can find all outer codewords, i.e polynomials $p \in \mathrm{GF}(q^m)[X]$ of degree less than k/m, such that $p(x_i) \in \{z_{ij} : 1 \le j \le \ell_i\}$ for more than $t \triangleq \sqrt{(k/m-1)L}$ values of i.

If this algorithm fails to output a codeword corresponding to a polynomial p when receiving y, there must be at least $(n_0 - t)$ blocks i of y for which $p(x_i)$ does not belong to the list $\mathcal{L}_i$ output by the inner decoder. This implies that the number e_i of positions where the encoding of $p(x_i)$ (by the inner code) differs from y_i satisfies $qe_i/(q-1) + s_i > (1-\zeta)n_1$. Hence the algorithm fails to output a codeword which differs from y at e places (recall y has s erasures) only if

$$\begin{aligned}\frac{q}{q-1}\cdot e + s &> (n_0 - t)(1-\zeta)n_1 \\ &= \left(n_0 - \sqrt{(k/m-1)L}\right)(1-\zeta)n_1.\end{aligned}$$

Thus we can decode from e errors and s erasures provided

$$\begin{aligned}\frac{q}{q-1}\cdot e + s &\leq \left(1 - \sqrt{\frac{(k/m-1)}{n_0}\frac{L}{n_0}}\right)(1-\zeta)n_0n_1 \\ \Leftarrow \frac{q}{q-1}\cdot e + s &\leq \left(1 - \sqrt{\delta(1-\sigma)\frac{(1-\rho)}{(\zeta^2-\rho)}}\right)(1-\zeta)n.\end{aligned}$$

Setting $\zeta = \sqrt{(1+\varepsilon)\rho}$, we see that the above condition is met if

$$\frac{qe}{q-1} + s \leq n\left(1 - \sqrt{(1+\varepsilon)\rho} - \sqrt{\frac{\delta(1-\sigma)}{\varepsilon\rho}}\right). \qquad \square$$

Corollary 4 [Simpler Construction than that of Corollary 3]: *For any finite field $\mathcal{F}_q$ and for any integer k and $\gamma > 0$, the statement of Corollary 3 holds with $n = O(k/\gamma^8)$, and the code is actually a Reed-Solomon code concatenated with* any *inner code with large enough distance.*

Proof: Follows from Theorem 10 and Proposition 4 with the following choice of parameters: $\delta = O(\gamma^4)$, $\rho = \sqrt{\delta}$, $\varepsilon = 1$. $\square$

6 Applications

6.1 Computing from partial solutions

We now point out applications of our codes and decoding algorithms to the framework of [6, 12] which concerns checking membership for languages in NP when the only either a small fraction of the witness is given (referred to as a partially publishable proof system in [12]), or when a large portion of the witness is in error. For this application we are only interested in binary codes. Corollaries 5 and 6 below improve the corresponding bounds (of $N^{2/3+\varepsilon}$ and $N/2 + N^{4/5+\varepsilon}$ respectively) proved in [12]. The result of Theorem 6 immediately implies:

Corollary 5 *For any language L in NP and every witness predicate R_L for L, and every $\varepsilon > 0$, there is a witness predicate R'_L and a polynomial time procedure to map witnesses y of R_L into N-bit witnesses z of R'_L such that given any $N^{1/2+\varepsilon}$ bits of a witness z which satisfies $R'_L(x,z)$, one can construct in polynomial time a witness y that satisfies $R_L(x,y)$.*

Using the result of Corollary 2, we can get the errors-analogue of the above statement.

Corollary 6 *For any $\varepsilon > 0$ and any language L in NP, we can construct a polynomial time computable witness predicate R_L, such that given any string which agrees with an N-bit witness y that satisfies $R_L(x,y)$ in at least $N/2 + N^{3/4+\varepsilon}$ positions, one can in polynomial time compute a witness which satisfies the predicate R_L.*

The above results imply that membership checking for languages in NP can be performed even when a partial witness or a highly noisy witness is given. They, however, treat witnesses simply as strings, and even if we start out with witnesses which have a nice semantic property, say, of being satisfying assignments to a SAT formula, in the encoding process they get mapped into arbitrary binary strings. This raises the question, considered by [6] of whether one can map SAT instances ϕ to (longer) SAT instances ϕ' such that a satisfying assignment for the original instance ϕ can be inferred given only very few bits of a satisfying assignment for ϕ'. Combining the techniques in their paper together with the code construction of Theorem 6, we get the following (compare with Theorem 2 of [6]):

Theorem 11 *For any $\varepsilon > 0$, there exist deterministic polynomial time algorithms Enc_{sat} and Rec_{sat} such that*

(i) If ϕ is a CNF formula over n variables, then $\phi' = \mathrm{Enc}_{sat}(\phi)$ is a CNF formula over $N = n^{O(1)}$ variables, with $|\phi'| = |\phi| + n^{O(1)}$.

(ii) If s' is an assignment to any $N^{3/4+\varepsilon}$ of the variables in ϕ' that can be extended to a full satisfying assignment, then $\mathrm{Rec}_{sat}(\phi, \phi', s')$ is a satisfying assignment for ϕ.

Proof: The proof follows the proof of Theorem 2 in [6]. Use the construction of Theorem 6 to get a code $\mathcal{C}$ which encode the n-bit strings x (corresponding to satisfying assignments to ϕ) into n^c-bit strings y such that given any $n^{c(1/2+\varepsilon)}$ bits of y, all strings x consistent with this can be found in polynomial time.

Let $C(x_1, x_2, \ldots, x_n, y_1, \ldots, y_{n^c})$ be a circuit which verifies that y is the encoding of x; using the nice structure of linear codes, it is easy to see that one can write C as a CNF formula using only n^{c+1} temporary variables as $\phi_C(x_1, \ldots, x_n, y_1, \ldots, y_{n^c}, z_1, \ldots, z_{n^{c+1}})$. For each variable y_i, introduce $n^c - 1$ new variables y_i^j $(1 \leq j \leq n^c - 1)$, and encode the equalities $y_i = y_i^1 = y_i^2 = \cdots$ in a CNF formula ϕ_{eq}. Finally set $\phi' = \phi \wedge \phi_C \wedge \phi_{eq}$, thus obtaining a formula over $N = n + n^{2c} + n^{c+1}$ variables.

Now suppose we are given some $N^{3/4+\varepsilon}$ bits of a satisfying assignment b of ϕ'. We have $N^{3/4+\varepsilon} = (n + n^{c+1} + n^{2c})^{3/4+\varepsilon} > n^{3c/2+2\varepsilon c} > n^{3c/2+\varepsilon c} + n^{c+1} + n$ (for large enough n). Since we have only n x's and n^{c+1} z's, this means we are left with at least $n^{3c/2+\varepsilon c}$ y-variables, which means we must have the value of at least $n^{(1/2+\varepsilon)c}$ different y_i's (as each y_i is only duplicated n^c times). By the property of the code $\mathcal{C}$, we can now find a list of all possible x's in polynomial time, and simply check if any of them satisfies ϕ. $\square$

6.2 Decoding from 100% error and Membership Comparable Sets

The results of this paper, specifically that of Corollary 3 or Corollary 4, imply the following interesting result (which is trivial for the binary case $q = 2$, but is far from obvious otherwise):

Theorem 12 *For any prime power q, there is an explicitly constructible asymptotically good family of linear codes $\mathcal{C}_n$ over $\mathcal{F}_q$ such that given a received word $y \in \mathcal{F}_q^n$, a list of all codewords in $\mathcal{C}_n$ which differ from y in every position can be found in polynomial time.*

Proof: Simply use the construction of Corollary 4 with a value of $\gamma < \frac{1}{(q-1)^2}$. Such a code can be efficiently list decoded up to a radius of $\left(1 - 1/q - (q-1)\gamma/q\right) > \left(1 - \frac{1}{q-1}\right)$ from any received word. Now given y which differs from some codeword c in *every* position, form a word z by setting z_i, $1 \leq i \leq n$, to be a random value not equal to y_i. The expected agreement between

z and c is thus $1/(q-1)$, and by the construction of the code all codewords with agreement at least $1/(q-1)$ with z can be found in polynomial time. This gives a randomized procedure to recover a list of all codewords that differ from y in every position. The method can be easily derandomized, completing the proof. $\square$

The above has application to *membership comparable sets* [17]. A set A is said to be $k(n)$ membership comparable if there is a polynomial time computable function that, given $k(n)$ instances of A of length at most n, excludes one of the $2^{k(n)}$ possibilities for memberships of the given strings in A. The above theorem also gives a proof of the following fact; our proof is different from the one in [17].

Theorem 13 ([17]) *If* SAT *is* $O(\log n)$ *membership comparable, then* $\mathrm{UniqueSAT} \in \mathrm{P}$.

Proof: Suppose there exists a constant d such that SAT is $d \log n$ membership comparable. Let ϕ be an instance of UniqueSAT on n boolean variables. Set $p = d \log n$ and $q = 2^p$ and consider the code $\mathcal{C}$ of rate n over $\mathrm{GF}(q)$ as guaranteed by Theorem 12; let the blocklength of $\mathcal{C}$ be m. For each $1 \le i \le m$, construct $p = d \log n$ SAT formulae ϕ_{ij} over n variables for $1 \le j \le p$, such that for $a \in \{0,1\}^n$, $\phi_{ij}(a) = \big(\phi(a) \wedge \text{The } j\text{th bit of } \mathsf{Enc}_{\mathcal{C}}(a) = 1\big)$ (here $\mathsf{Enc}_{\mathcal{C}}(a)$ is the encoding of a in the code $\mathcal{C}$, and elements of $\mathrm{GF}(2^p)$ are viewed as p-bit vectors).

Suppose ϕ were satisfiable (in case it is not, we will never find a witness, so we only worry about the satisfiable case), and let a be the *unique* satisfying assignment to ϕ. We use the polynomial membership comparator function f guaranteed by the hypothesis, to get, for $1 \le i \le m$, vectors $b_i = f(\phi_{i1}, \ldots, \phi_{ip}) \in \{0,1\}^p$ such that $b_i \ne (\chi_{SAT}(\phi_{i1}), \ldots, \chi_{SAT}(\phi_{ip}))$. By the definition of ϕ_{ij} and the fact that a is the unique satisfying assignment to ϕ, we can conclude, for $1 \le i \le m$, that b_i when viewed as an element of $\mathrm{GF}(2^p)$ is not equal to the ith symbol of $\mathsf{Enc}_{\mathcal{C}}(a)$, and thus we have a word $(b_1, b_2, \ldots, b_m) \in \mathrm{GF}(q)^m$ with **all** symbols in disagreement with the codeword $\mathsf{Enc}_{\mathcal{C}}(a)$. Now using the decoding algorithm for $\mathcal{C}$ as in Theorem 12, we can find a list of all such a's in polynomial time, and by simply going over the list also find the unique satisfying assignment to ϕ. $\square$

References

[1] N. ALON. Packings with large minimum kissing numbers. *Discrete Mathematics*, 175 (1997), pp. 249-251.

[2] N. ALON. Personal Communication, October 1999.

[3] N. ALON, J. BRUCK, J. NAOR, M. NAOR AND R. ROTH. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Trans. on Information Theory*, 38 (1992), pp. 509-516.

[4] P. DELSARTE. Bounds for unrestricted codes, by linear programming. *Philips Res. Reports*, 27 (1972), pp. 272-289.

[5] G. D. FORNEY. Generalized Minimum Distance Decoding. *IEEE Trans. Inform. Theory*, Vol. 12, pp. 125-131, 1966.

[6] A. GAL, S. HALEVI, R. J. LIPTON AND E. PETRANK. Computing from partial solutions. *Proc. of 14th Annual IEEE Conference on Computation Complexity*, pp. 34-45, 1999.

[7] A. GARCIA AND H. STICHTENOTH. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound. *Inventiones Mathematicae*, 121 (1995), pp. 211-222.

[8] O. GOLDREICH, R. RUBINFELD AND M. SUDAN. Learning polynomials with queries: The highly noisy case. *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*, pp. 294-303, 1995.

[9] V. GURUSWAMI AND M. SUDAN. Improved decoding of Reed-Solomon and algebraic-geometric codes. *IEEE Transactions on Information Theory*, 45 (1999), pp. 1757-1767. Preliminary version in *Proc. of FOCS'98*.

[10] J. JUSTESEN. A class of constructive asymptotically good algebraic codes. *IEEE Trans. Inform. Theory*, 18 (1972), pp. 652-656.

[11] M. KIWI. Testing and weight distributions of dual codes. *ECCC Technical Report TR-97-010*, 1997.

[12] RAVI KUMAR AND D. SIVAKUMAR. Proofs, codes, and polynomial-time reducibilities. *Proc. of 14th Annual IEEE Conference on Computation Complexity*, 1999.

[13] R. KOTTER AND A. VARDY. Algebraic soft-decoding of Reed-Solomon codes. Manuscript, August 1999.

[14] F. J. MACWILLIAMS AND N. J. A. SLOANE. *The Theory of Error-Correcting Codes*. Amsterdam: North Holland, 1977.

[15] Y. I. MANIN AND S. G. VLADUT. Linear codes and modular curves. *J. Soviet. Math.*, 30 (1985), pp. 2611-2643.

[16] R. J. MCELIECE, E. R. RODEMICH, H. C. RUMSEY JR. AND L. R. WELCH. New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities. *IEEE Trans. on Inform. Theory*, 23 (1977), pp. 157-166.

[17] D. SIVAKUMAR. On membership comparable sets. In *Proc. of 13th Annual IEEE Conference on Computational Complexity*, pp. 2-7, 1998.

[18] M. SUDAN. Decoding of Reed-Solomon codes beyond the error-correction diameter. *Proceedings of the 35th Annual Allerton Conference on Communication, Control and Computing*, 1997. Also appears in *Journal of Complexity*, 13(1):180-193, 1997.

[19] M. SUDAN, L. TREVISAN AND S. VADHAN. Pseudorandom generators without the XOR lemma. In *Proc. of STOC'99*, pp. 537-546.

[20] Y. SUGIYAMA, M. KASAHARA, S. HIRASAWA AND T. NAMEKAWA. A new class of asymptotically good codes beyond the Zyablov bound. *IEEE Trans. Inform. Theory*, 24 (1978), pp. 198-204.

[21] Y. SUGIYAMA, M. KASAHARA, S. HIRASAWA AND T. NAMEKAWA. Superimposed concatenated codes. *IEEE Trans. Inform. Theory*, 26 (1980), pp. 735-736.

[22] M. A. TSFASMAN, S. G. VLĂDUT AND T. ZINK. Modular curves, Shimura curves, and codes better than the Varshamov-Gilbert bound. *Math. Nachrichten*, 109:21-28, 1982.

[23] E. J. WELDON, JR. Justesen's construction – The low-rate case. *IEEE Trans. Inform. Theory*, 19 (1973), pp. 711-713.

A Fourier Transforms over a q-ary alphabet

Proposition 14 *Let* $f : \mathcal{F}_q^m \to \mathcal{F}_q$ *be an arbitrary function, and for every* $\alpha \in \mathcal{F}_q^m$, *let the linear function* $l_\alpha : \mathcal{F}_q^m \to \mathcal{F}_q$ *be defined by:* $l_\alpha(x) = \sum_{i=1}^m \alpha_i x_i$ *(all operations performed over* $\mathcal{F}_q$*). Then*

$$\sum_{\alpha \in \mathcal{F}_q^m} \left(1 - \frac{q}{q-1} \mathsf{Dist}(f, l_\alpha)\right)^2 \le 1.$$

Remark: For the case $q = 2$, $\big(1 - 2\mathsf{Dist}(f, l_\alpha)\big)$ equals the *Fourier coefficient* $\hat{f}_\alpha$ of f with respect to l_α, and the statement of the Proposition holds with equality, and is simply the standard Parseval's identity $\sum_\alpha \hat{f}_\alpha^2 = 1$. The result for the non-binary case appears in [11], and the proof there is based on the MacWilliams-Sloane identities for the weight distribution of dual codes; we give a more elementary proof below.

Proof: The proof works by viewing any $f : \mathcal{F}_q^m \to \mathcal{F}_q$ as a q^m-tuple over $\mathcal{F}_q$, and embedding it as a $q^m \cdot q$-dimensional real unit vector. The vectors associated with l_α and l_β will be orthogonal (in the usual dot product over $\mathcal{R}^{q^m \cdot q}$) whenever $\alpha \neq \beta$. The quantity $(1 - \frac{q}{q-1}\mathsf{Dist}(f,g))$ for any two functions f, g will simply be the dot product of the vectors associated with f, g. The result will then follow since the sum of the squares of the projections of a unit vector along pairwise orthogonal vectors can be at most 1.

Suppose the q elements of $\mathcal{F}_q$ are $x_1, x_2, \ldots, x_q$. Associate a q-dimensional vector e_i with x_i as follows (e_{ij} denotes the jth component of e_i): $e_{ii} = \sqrt{(q-1)/q}$ and $e_{ij} = -1/\sqrt{q(q-1)}$ for $j \neq i$. Note that this definition satisfies $\langle e_i, e_i \rangle = 1$ and $\langle e_i, e_j \rangle = \frac{-1}{q-1}$ for $i \neq j$. Treating a function $f : \mathcal{F}_q^m \to \mathcal{F}_q$ as a string over $\mathcal{F}_q$, we view f as the $q^m \cdot q$-dimensional vector obtained in the obvious way by juxtaposing the q-dimensional vectors for each of the q^m values which f takes on its domain, and then normalizing it to a unit vector. Note that when we take the inner product $\langle f, g \rangle$, we get a contribution of 1 corresponding to the positions where f, g agree, and a contribution of $-1/(q-1)$ corresponding to places where f, g differ. Hence $\langle f, g \rangle = (1 - \mathsf{Dist}(f,g)) \cdot 1 + \mathsf{Dist}(f,g) \cdot -1/(q-1) = 1 - \frac{q}{q-1}\mathsf{Dist}(f,g)$. Hence $\langle l_\alpha, l_\alpha \rangle = 1$. For $\alpha \neq \beta$, $\mathsf{Dist}(l_\alpha, l_\beta) = (q-1)/q$ (two distinct codewords in the Hadamard code corresponding to $\mathcal{F}_q^m$ agree is exactly q^{m-1} places and differ in $q^{m-1}(q-1)$ places), and thus $\langle l_\alpha, l_\beta \rangle = 0$ when $\alpha \neq \beta$. The result now follows since

$$\sum_{\alpha \in \mathcal{F}_q^m} \left(1 - \frac{q}{q-1}\mathsf{Dist}(f, l_\alpha)\right)^2 = \sum_\alpha \langle f, l_\alpha \rangle^2 \leq \langle f, f \rangle = 1. \square$$

Corollary 7 *Suppose $f : \mathcal{F}_q^m \to \mathcal{F}_q$ is a string of q^m symbols over $\mathcal{F}_q$ except that a fraction s of them are erased. Let e_α be the fraction of positions (among the non-erased positions) where f differs from l_α. Then*

$$\sum_{\alpha \in \mathcal{F}_q^m} (1 - s - \frac{q}{q-1} \cdot e_\alpha)^2 \leq (1-s).$$

Proof: As in the proof of Proposition 14, view f as a $q^m \cdot q$-dimensional vector over the reals, except that now the vector has zeroes at the q coordinates corresponding to every erased position. Now

$$\sum_\alpha (1 - s - \frac{q}{q-1} e_\alpha)^2 = \sum_\alpha \langle f, l_\alpha \rangle^2 \leq \langle f, f \rangle = (1-s). \qquad \square$$

B List decoding Reed-Solomon and Algebraic-geometric codes with weights

In this section we present a version of the weighted polynomial reconstruction algorithm due to [9]. The algorithm as presented in [9] handled integer weights and ran in time polynomial in the sum of the weights. Here we note that with an ε degradation in performance, the algorithm can be implemented to run in $\mathrm{poly}(n, 1/\varepsilon)$ time, even when the weights are arbitrary rational numbers.

Let us first formally define the weighted polynomial reconstruction problem.

(Weighted polynomial reconstruction)
INPUT: n distinct points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $F \times F$, F a field, together with n non-negative weights $w_1, \ldots, w_n$, and parameters k and t. (Assume $w_1 \leq w_2 \leq \cdots \leq w_n$.)
OUTPUT: All polynomials p of degree less than k such that

$$\sum_{i : p(x_i) = y_i} w_i \geq t.$$

Note: The x_i's above need **not** be distinct.

Theorem 15 ([9]) *If the weights are non-negative integers the weighted polynomial reconstruction problem can be solved in time polynomial in the sum of w_i's provided $t > \sqrt{(k-1)\sum_{i=1}^n w_i^2}$.*

Proposition 16 *For any tolerance parameter $\varepsilon > 0$, the weighted polynomial reconstruction problem can solved in polynomial (in n and $1/\varepsilon$) time provided*

$$t > \sqrt{(k-1)\sum_{i=1}^n w_i^2 + \varepsilon w_n}.$$

Proof: Pick any large integer $L \geq \frac{n}{\varepsilon}$, and form the integer weights $w_i' = \lfloor L w_i / w_n \rfloor$. Since $w_i \leq w_n$ for all i, the weights w_i' are all at most L, and now the algorithm of [9] can used to in find, in $\mathrm{poly}(nL)$ time, a list of all polynomials p of degree less than k, provided

$$\sum_{i : p(x_i) = y_i} w_i' > \sqrt{(k-1)\sum_{i=1}^n w_i'^2}.$$

But since $L w_i / w_n \geq w_i' \geq L w_i / w_n - 1$, this implies we find in $\mathrm{poly}(nL) = \mathrm{poly}(n, 1/\varepsilon)$ time all polynomials p of degree less than k, provided

$$\begin{aligned} \sum_{i : p(x_i) = y_i} \left(\frac{L w_i}{w_n} - 1\right) &> \sqrt{(k-1)\sum_{i=1}^n \left(\frac{L w_i}{w_n}\right)^2} \\ \Longleftarrow \quad \sum_{i : p(x_i) = y_i} w_i &> \sqrt{(k-1)\sum_{i=1}^n w_i^2 + \frac{n w_n}{L}} \\ \Longleftarrow \quad \sum_{i : p(x_i) = y_i} w_i &> \sqrt{(k-1)\sum_{i=1}^n w_i^2 + \varepsilon w_n} \end{aligned}$$

(the last step follows since $L \geq n/\varepsilon$). $\square$

Weighted list decoding of Algebraic-geometric codes: Though it is not explicitly stated in [9], their techniques also imply a decoding algorithm for algebraic-geometric codes with weights on the codewords positions. Let $x_0, x_1, \ldots, x_n$ be $n+1$ distinct *rational points* in an algebraic function field over $\mathcal{F}_q$. Then an algebraic-geometric code has codewords corresponding to the evaluations of "functions" f which have at most α poles at x_0 (α is a parameter of the code) and no poles elsewhere (this space of functions is denoted by L_{α, x_0}), at the rational points $x_1, x_2, \ldots, x_n$. The results of [9], together with the trick of Proposition 16 above, imply the following.

Proposition 17 *Let $\mathcal{C}$ be an algebraic-geometric code of blocklength n defined over $\mathcal{F}_q$ with rational points $\{x_0, x_1, \ldots, x_n\}$ and the space L_{α, x_0} of functions; the designed distance d of $\mathcal{C}$ is $(n - \alpha)$. Suppose we are given N pairs (p_i, y_i), $1 \leq i \leq N$ with associated weights w_i, where $p_i \in \{x_1, x_2, \ldots, x_n\}$ and $y_i \in \mathcal{F}_q$. Then, for any $\varepsilon > 0$, a list of all $f \in L_{\alpha, x_0}$ such that*

$$\sum_{i : f(p_i) = y_i} w_i > \sqrt{(n-d)\sum_{i=1}^N w_i^2 + \varepsilon \max_i w_i}$$

can be found in $\mathrm{poly}(N, 1/\varepsilon)$ time provided certain assumptions about the algebraic function field underlying $\mathcal{C}$ hold (see [9] for details on these assumptions). $\square$

C The Linear Programming Bound for large distance binary codes

In this section we include a proof that binary linear codes with a minimum distance $n/2 - c\sqrt{n}$ for any $c < 1/2$ can have at most polynomially many codewords (in fact at most $O(n^{3/2})$ codewords). The proof uses the Linear Programming bound for linear codes (see for example [14]). This bound on the distance is *tight*, in the sense that there are linear codes with exponentially many codewords and with minimum distance $n/2 - n^{1/2+\varepsilon}$, for any $\varepsilon > 0$ (Reed-Solomon codes concatenated with Hadamard codes gives one such construction).

Proposition 1 *A binary linear code of blocklength n and minimum distance $d = n/2 - c\sqrt{n}$ can have at most $O(n^{3/2})$ codewords if $c < 1/2$.*

Proof: The proof follows exactly along the lines of the McEliece-Rodemich-Rumsey-Welch upper bound [16] (a description can also be found in [14, Chapter 17]), but is actually easier as we only prove a very specific result.

We use the dual version of the linear programming bound [4] which states the following: if a polynomial $\beta(x)$ of degree at most n with *Krawtchouk expansion* $\beta(x) = \sum_{k=0}^{n} \beta_k P_k(x)$ can be found such that $\beta_0 = 1$,[5] $\beta_k \geq 0$ for $1 \leq k \leq n$, and $\beta(j) \leq 0$ for $d \leq j \leq n$, then the maximum number of codewords in a binary code of blocklength n and distance d, denoted $A(n,d)$, is at most $\beta(0)$.

Let us now focus on the case $d = n/2 - c\sqrt{n}$ with $c < 1/2$. Pick $a = n/2 - p\sqrt{n}$ where $c < p < 1/2$, so that $a < d$. Consider the polynomial

$$\alpha(x) = \frac{1}{a - x}\Big(P_2(x)P_1(a) - P_1(x)P_2(a)\Big)^2.$$

$\alpha(x)$ is a polynomial of degree 3 and can be expanded as $\alpha(x) = \sum_{k=0}^{3} \alpha_k P_k(x)$. We will then choose $\beta(x) = \alpha(x)/\alpha_0$. In order to get a bound on $A(n,d)$, we need to check that (i) $\alpha(i) \leq 0$ for $d \leq i \leq n$; (ii) $\alpha_k \geq 0$ for $1 \leq k \leq n$; and (iii) $\alpha_0 > 0$.

By the definition of $\alpha(x)$, $\alpha(x) \leq 0$ if $x > a$, and therefore also if $x > d$ (since $a < d$). Now $\alpha_k = 2^{-n}\sum_{i=0}^{n} \alpha(i)P_i(k)$, and the claim that $\alpha_k \geq 0$ follows from a few further properties of Krawtchouk polynomials (see [14, Chap. 17] for details). Similarly it can also be shown that

$$\begin{aligned}
\alpha_0 &= -nP_1(a)P_2(a) \\
&= -n(n-2a)\Big(\binom{n}{2} - 2na + 2a^2\Big) > 0 \\
&= -n(2p\sqrt{n})\Big(-2(1/4 - p^2)n\Big) \\
&= 4n^{5/2}p(\frac{1}{4} - p^2) > 0.
\end{aligned}$$

We can now conclude,

$$\begin{aligned}
A(n,d) &\leq \beta(0) = \alpha(0)/\alpha_0 \\
&= \frac{\frac{1}{a}\Big\{\binom{n}{2}P_1(a) - nP_2(a)\Big\}^2}{\alpha_0} \\
&= \frac{\Big\{\binom{n}{2}2p\sqrt{n} + 2n^2(1/4 - p^2)\Big\}^2}{4n^{5/2}p(1/4 - p^2)(n/2 - p\sqrt{n})} \\
&= O(n^5/n^{7/2}) = O(n^{3/2}). \qquad \square
\end{aligned}$$

[5] The Krawtchouk polynomials are a family of orthogonal polynomials and can be defined recursively by: $P_0(x) = 1$, $P_1(x) = n - 2x$ and $(k+1)P_{k+1}(x) = (n-2x)P_k(x) - (n-k+1)P_{k-1}(x)$.