List Decoding Algorithms for Certain Concatenated Codes*

Venkatesan Guruswami† Madhu Sudan‡

November, 2000

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

The *list decoding* problem for an error-correcting code consists of reconstructing a list of *all* codewords within a specified Hamming distance from a received word. List decoding, which was introduced independently by Elias and Wozencraft [6, 33], offers a potential for recovery from errors beyond the traditional "error-correction radius" (i.e. half the minimum distance) of a code.

Though the notion of list decoding is more than 40 years old, until recently no *efficient* list decoding algorithms were known for any (non-trivial) family of codes that could correct asymptotically more errors than the traditional error-correction radius. The first such list decoding algorithm was given by Sudan [27] for Reed-Solomon codes, and this algorithm was later improved and also generalized to the case of algebraic-geometric codes [26, 14]. Reed-Solomon codes suffer from the drawback of large alphabet size, and though AG-codes can be constructed over small alphabets, their list-decodability is limited by their rate vs. distance trade-off, which in turn is limited by certain lower bounds on the genus of function fields.

*An extended abstract of this paper [15] appeared in the *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC)*, Portland, Oregon, May 2000, pp. 181-190.

†MIT Laboratory for Computer Science, 200 Technology Square, Cambridge, MA 02139. Email: venkat@theory.lcs.mit.edu.

‡MIT Laboratory for Computer Science, 200 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.

In this paper we consider list decoding algorithms for certain families of concatenated codes that have very large minimum distance. We are motivated mainly by the quest for linear codes over a small alphabet 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 information-theoretically the best one can hope to recover from, and are interested in codes of good rate that can be list decoded up to so many errors. Specifically, over the finite field $\mathrm{GF}(q)$ (also denoted as $\mathbb{F}_q$), we consider problems which are of the following form. Suppose one needs to transmit k symbols (over $\mathbb{F}_q$) 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 $\mathcal{C}_n$. In addition to their obvious relevance to coding theory, such codes also have surprising and elegant applications to certain areas of computer science, some of which are discussed in an earlier version of this paper [15].

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 $\frac{q}{q-1}e + s \le (1-\gamma)n$.

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. We remark here that *all* (currently known) codes that can be list decoded efficiently from a fraction $(1-1/q-\gamma)$ of errors and that have blocklength $n = \mathrm{poly}(k/\gamma)$, are based on the powerful idea of "code concatenation" introduced by Forney [8]. For example, the maximum fraction of errors for which algebraic-geometric codes over $\mathbb{F}_q$ can be decoded using current techniques is (about) $(1-q^{-1/4})$, and we do not know how to efficiently decode AG-codes from a fraction of errors approaching $(1-1/q)$. Code families with minimum distance approaching $(1-1/q-\varepsilon)$ can in principle be list decoded to a radius of (about) $(1-1/q-O(\sqrt{\varepsilon}))$, and constructions of such codes are known without relying on concatenation [4], but no *efficient* list decoding algorithm is known for these codes.

1.1 Previous Work

Let us first consider the case of erasures. It is shown in [19] that a Reed-Solomon code concatenated with a Hadamard code together with the outer (list) decoder for Reed-Solomon codes of [27, 14] 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 relative distance $(1-1/q)(1-\delta)$ can be list decoded from a fraction $(1-\delta)$ of erasures. 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, one can achieve a blocklength of $n = O(k/\delta^3)$ by concatenating a low-rate outer Reed-Solomon code with an inner code obtained by brute force search for the "best" inner code [9]. This construction is sometimes not "explicit enough", as the construction complexity is not a fixed polynomial independent of δ. The first *explicit* asymptotically good construction of very large minimum distance was given by Weldon [32] following the work of Justesen [17], and related constructions appear in the [29, 30]. One can also use algebraic-geometric codes as outer codes in a concatenation scheme to give explicit construction of codes with $n = O(k/\delta^3)$, but these codes inherit the high construction complexity of algebraic-geometric codes. The best code construction of reasonable complexity that has relative distance $(1-1/q)(1-\delta)$ is due to [3], and this also achieves $n = O(k/\delta^3)$.

For the case of errors, Zyablov and Pinsker [34] (see also [7]) show that binary linear codes of rate $\Omega(\gamma^2)$ *exists* that have a small number of codewords in any Hamming ball of radius $(1/2-\gamma)$, and thus can, *in principle*, be list decoded up to this radius. This result, specifically the bound on the number of codewords, was recently improved in [13]. However, these results are existential in nature and it is completely unclear how to construct these codes in polynomial time, let alone how to *efficiently* list decode such a code. (A recent result [13] shows how to get some constructive results along this direction together with efficient list decoding algorithms; see Section 1.3 on subsequent work below for more details.)

Codes with large minimum distance are known to have a large list decoding radius, but no simple connection between the distance of the code and its *efficient* list decodability exists. Constructions are therefore much harder to come by if one also requires an efficient list decoding algorithm. It is known that a concatenated code with an outer Reed-Solomon code and an inner Hadamard code with blocklength $n = \mathrm{poly}(k/\gamma)$ can be efficiently list decoded up to a fraction $(1-1/q-\gamma)$ of errors [28] (this is also implicit in [19]), 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 [17], he also gives an algorithm to decode his code construction up to half the minimum distance. Note that unambiguous decoding (as opposed to list decoding) 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 [17] 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 [32] 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

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

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.

1.2 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 $n \geq (k/\gamma)^2$. Our result implies, in the terminology of [10, 19], 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 [12]), and reduces to the one in [12] in presence of only errors. 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 reconstruction algorithm due to [14]. The linear programming based bounds for codes to provide evidence that this (quartic) dependence of n on γ is probably the best possible, as long as one uses current minimum-distance based arguments 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 a high construction complexity, and moreover the decoding algorithm has to assume complicated subroutines over function fields. Therefore, 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 $n = O(k/\gamma^8)$.

One 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.

1.3 Recent Related Work

Independently of our work, Nielsen [24] also gave decoding algorithms for concatenated schemes based on Reed-Solomon codes as outer codes, again using the weighted polynomial reconstruction algorithm from [14] for decoding the outer code. His algorithms perform well for inner codes with a small (constant) minimum distance, where as our results focus on inner codes with very large minimum distance (since our motivation is primarily to correct from very high noise in the low-rate regime).

Koetter and Vardy [20] use the polynomial reconstruction algorithm of [14] to design *soft-decision* decoding algorithms for Reed-Solomon codes that appear to be superior to Forney's GMD algorithm [9]. Their techniques can also be used to decode concatenated schemes with an outer Reed-Solomon (or algebraic-geometric) code, though they do not provide a quantitative (theoretical) analysis of the performance of their algorithm for concatenated codes. They do report experiments demonstrating extraordinary coding gains using an outer Reed-Solomon code and an inner convolutional code of appropriate parameters. They employ the BCJR [5] forward-backward algorithm to decode the inner convolutional code and pass reliability information to the outer Reed-Solomon decoder, after which the outer decoding proceeds along the lines of our approach in this paper.

The concatenated codes we consider in this paper 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 were known prior to our work. In subsequent work, Guruswami, Håstad, Sudan and Zuckerman [13] improve our bounds and give constructions of *binary* codes of blocklength $O(k/\gamma^4)$ that can be efficiently list decoded from a fraction $(\frac{1}{2} - \gamma)$ of errors.

1.4 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.

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, denoted usually by d, is the minimum Hamming distance between any two distinct members of the code. A code with these parameters is referred to as an $[n, k, d]_q$ code. It is also convenient to normalize these quantities by dividing them by the blocklength. We use the *rate* ($\stackrel{\text{def}}{=} k/n$) and the *relative (minimum) distance* ($\stackrel{\text{def}}{=} d/n$) to denote the normalized quantities.

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-geometric 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 [9]: 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 often considered in the past. The Reed-Solomon code concatenated with a Hadamard code, in particular, has been a popular code and has been considered for instance in [1, 4, 19, 28]. 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 several 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 for each codeword position, the outer decoder

is given a *list* of possibilities for the corresponding symbol, and it needs to list decode from this information. This is possible for Reed-Solomon and algebraic-geometric codes [27, 14]. One important aspect that is required for the outer decoder to work is that the total number of candidates passed by the inner decoder is not too large. In order to ensure this, we need combinatorial bounds on the maximum number of possible codewords with a certain number of errors and erasures from a received 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 actually transmitted. (This is similar to the approach behind Forney's GMD decoding algorithm [9], except that in GMD decoding, only *one* candidate codeword is passed by the inner decoder for each position, together with an associated confidence parameter. Thus our approach gives the decoding process more flexibility as it allows the inner decoder to pass more "soft information" to the outer decoder.) It is reasonable that the weight a certain inner codeword receives under the inner decoding should be somehow related to and decrease with its distance from the actual inner word that was received. The outer decoder then uses this "reliability/confidence information" in its decoding. An algorithm that can use such reliability information was given for Reed-Solomon codes by the authors in [14]; 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 9), 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 [12] and specializes to that bound for the errors only case, although our proof is different and simpler. In very recent work [16], this result has been improved and its proof further simplified, but we include a full proof here for the sake of completeness.

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 $y \in \mathbb{F}_q^{n-s}$ 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

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

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

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

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

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

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

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

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

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

¿From (1) and (4), we get

$$m \leq d\left(\frac{q}{q-1}\frac{\bar{e}^2}{n-s} - 2\bar{e} - \frac{q-1}{q}s + d\right)^{-1} . \tag{5}$$

Using $\bar{e} \leq e$, and that the function in the parentheses above decreases in the range $(0, e)$, we can get an upper bound on m as long as

$$\begin{aligned}
\left(\frac{q}{q-1}\right)\frac{e^2}{n-s} - 2e - \frac{q-1}{q}s + d &> 0 \\
\iff \quad \left(n - s - \frac{q}{q-1}e\right)^2 &> (n-s)^2 + (n-s)s - d(n-s)\left(\frac{q}{q-1}\right) \\
\iff \quad (\gamma n)^2 &> (n-s)(n - \frac{q}{q-1}d) \\
\iff \quad \gamma^2 &> (1-\sigma)\delta \ .
\end{aligned}$$

(The last step follows since $d = (1 - 1/q)(1-\delta)n$ and $s = \sigma n$.) Hence if $\gamma > \sqrt{(1-\sigma)\delta}$, we get, using Equation (5), that

$$m \leq \frac{(1-\sigma)(1-\delta)}{\gamma^2 - (1-\sigma)\delta} \ . \qquad \square$$

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 is small. We now quantify the distance properties of the main concatenated codes we will use; namely we express the blocklength n of the code in terms of the dimension k and the distance parameter $\delta = 1 - qd/(q-1)$.

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

Proof: Let the code $\mathcal{C}_{\mathrm{RS-Had}}(k,\delta)$ be obtained by concatenating a Reed-Solomon code of dimension k/m over $\mathrm{GF}(q^m)$ with the Hadamard code associated with $\mathrm{GF}(q^m)$. The dimension of the concatenated code is clearly k, and its blocklength is $n \stackrel{\text{def}}{=} (q^m)^2$. 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. $\square$

[2]Here and elsewhere 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 dimension k and blocklength at most ck^a/δ^b.

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 $\mathcal{C}_{\mathrm{AG-Had}}(k,\delta)$, with the following properties:*

(i) $\mathcal{C}_{\mathrm{AG-Had}}(k,\delta)$ has dimension k, blocklength $n = O(\frac{k}{\delta^3 \log 1/\delta})$ and minimum distance at least $d = (1-1/q)(1-\delta)n$.

(ii) It is obtained by concatenating an (appropriate) algebraic-geometric code with a Hadamard code.

Proof: Let the code $\mathcal{C}_{\mathrm{AG-Had}}(k,\delta)$ be obtained by concatenating an algebraic-geometric code over $\mathrm{GF}(q^m)$ of dimension k/m and blocklength n_0, with the Hadamard code that encodes an element of $\mathrm{GF}(q^m)$ by q^m symbols over $\mathrm{GF}(q)$. 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 [31, 22, 11]). The dimension of the concatenated code is clearly k and the blocklength is $n \stackrel{\text{def}}{=} n_0 q^m$. The relative minimum distance of the code is at least $(1-1/q)(1-\frac{k/m-1}{n_0} - 1/(q^{m/2}-1))$, and this gives $n = O(\frac{k}{\delta^3 \log 1/\delta})$. $\square$

Proposition 4 *For every $k, \delta, \rho > 0$ and prime power q, there is a linear code over $\mathrm{GF}(q)$, denoted $\mathcal{C}_{\mathrm{RS-GoodInner}}(k,\delta,\rho)$, with the following properties:*

(i) $\mathcal{C}_{\mathrm{RS-GoodInner}}(k,\delta,\rho)$ has dimension k, blocklength $n = O(\frac{k}{\delta\rho^2})$ and minimum distance at least $d = (1-\frac{1}{q})(1-\delta)(1-\rho)n$.

(ii) It is obtained by concatenating an outer Reed-Solomon code of rate δ with any *q-ary inner code of appropriate dimension that has relative distance at least $(1-\frac{1}{q})(1-\rho)$.*

Moreover such a code can be constructed in time polynomial in n.[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 $d_1 \geq (1-1/q)(1-\rho)n_1$ with $m = \Omega(\rho^2 n_1)$ (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 [32]). The blocklength is $n = n_1 q^m = O(n_1 \cdot \frac{k/m}{\delta}) = O(\frac{k}{\delta\rho^2})$. $\square$

Remark: The above codes have, up to constant factors, the smallest blocklength possible for a given value of rate and relative minimum distance. 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.

[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, [17, 32]).

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 $M \leq \frac{q^2}{q-1}d$. $\square$

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 polynomial in number, then they can certainly be found and output in polynomial time. $\square$

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 $\mathbb{F}_q$ and for any integer k and $\gamma > 0$, there exists an explicitly specified code which encodes k symbols over $\mathbb{F}_q$ into n symbols over $\mathbb{F}_q$ 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 dimension 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 $\mathcal{C}_{\mathrm{AG-Had}}(k,\gamma)$ from Proposition 3. $\square$

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

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 $\mathcal{C}_{\mathrm{RS-Had}}(k, \gamma)$ 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 [23] on the rate of codes implies that a q-ary code relative minimum distance $(1 - 1/q)(1 - \delta)$ can have rate at most $O(\delta^2 \log(1/\delta))$. 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 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.

5.2 Decoding from errors and erasures

Theorem 7 *For an $[n, k, d]_q$ code $\mathcal{C}_{\mathrm{RS-Had}}(k, \delta)$ with $d \geq (1 - 1/q)(1 - \delta)n$, there is a polynomial time list decoding algorithm for e errors and $s = \sigma n$ erasures as long as*

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

Corollary 2 *For any finite field $\mathbb{F}_q$ and for any integer k and $\gamma > 0$, there exists an* explicitly specified *linear code over $\mathbb{F}_q$ of dimension 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 7. □

Proof of Theorem 7: Before we begin proving the theorem note that this matches the combinatorial bound for list decoding proved in Theorem 1.

Recall that $\mathcal{C}_{\mathrm{RS-Had}}(k, \delta)$ 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 $\mathcal{C}_{\mathrm{RS-Had}}(k, \delta)$ 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 $i^{\rm th}$ 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 $1 \le i, j \le n'$, let e_{ij} be the number of positions where y_i differs from α_j, and define the *weight* w_{ij} as:

$$w_{ij} \stackrel{\rm def}{=} (1 - \frac{s_i}{n'} - \frac{q}{q-1} \cdot \frac{e_{ij}}{n'}).$$

One key property of these weights, proved in Corollary 5 in Appendix A is that, for each i,

$$\sum_j w_{ij}^2 \le (1 - \frac{s_i}{n'}). \tag{6}$$

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 [14] (see also Proposition 12 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 (6), 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

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

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 [23], for codes with such high minimum distance, we must have $k/n = O(\gamma^4 \log(1/\gamma))$. This is shown in [23] only for the case of fixed constant $\gamma > 0$, but the proof can be extended to a more general case, by studying the asymptotic behavior of the smallest roots of Krawtchouk polynomials [4, 25]. Thus 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 the blocklength on the dimension in the construction $\mathcal{C}_{\mathrm{RS-Had}}$ is quadratic, and hence the above code family 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 8 *For an $[n,k,d]$ code $\mathcal{C}_{\mathrm{AG-Had}}(k,\delta)$ over $\mathrm{GF}(q)$ with $d \geq (1-1/q)(1-\delta)n$, there is a polynomial time list decoding algorithm for e errors and $s = \sigma n$ erasures as long as*

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

Corollary 3 *For any finite field $\mathbb{F}_q$ and for any integer k and $\gamma > 0$, there exists an* explicitly specified *linear code over $\mathbb{F}_q$ of dimension 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 $q/(q-1)e + s \leq (1-\gamma)n$.*

Proof of Theorem 8 *(Sketch)*: Recall that $\mathcal{C}_{\mathrm{AG-Had}}(k,\delta)$ 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 $n = n_0 q^m$, dimension k and minimum distance at least $d = (1-1/q)(1-\delta)$ where $\delta = 1 - d_0/n_0$.

The decoding algorithm is exactly similar to the one in Theorem 7 except that instead of a weighted polynomial reconstruction routine, one uses a weighted version of the decoding algorithm of [14] for algebraic-geometric codes (stated formally in Appendix B). Using exactly the same definition of weights as earlier and arguing as in Theorem 7, 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

$$\begin{array}{rcl} n_0 - \dfrac{s}{q^m} - \dfrac{q}{q-1} \cdot \dfrac{e}{q^m} & > & \sqrt{(n_0 - d_0)(n_0 - \dfrac{s}{q^m})} + \varepsilon \\ \Longleftarrow \qquad \dfrac{q}{q-1}e + s & \leq & n(1 - \sqrt{(1-\sigma)\delta}) - O(1) \end{array}$$

provided we choose $\varepsilon \leq 1/q^m$. $\square$

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 [11]). We next present a decoding algorithm for the simpler construction of $\mathcal{C}_{\text{RS}-\text{GoodInner}}(k,\delta,\rho)$ from Proposition 4.

Theorem 9 *For an* $[n,k,d]_q$ *code* $\mathcal{C}_{\text{RS}-\text{GoodInner}}(k,\delta,\rho)$, *where* $d \geq (1-1/q)(1-\delta)(1-\rho)n$, *for every* $\varepsilon > 0$, *one can list decode in* $\text{poly}(n, 1/\varepsilon)$ *time from* e *errors and* $s = \sigma n$ *erasures as long as*

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

Proof: Recall that $\mathcal{C}_{\text{RS}-\text{GoodInner}}(k,\delta,\rho)$ 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 $\text{GF}(q)$ that has blocklength n_1, dimension m and minimum distance at least $(1-1/q)(1-\rho)n_1$.

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 $1 \leq i \leq n_0$, 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 $L = \sum_{i=1}^{n_0} \ell_i \leq \frac{(1-\sigma)(1-\rho)n_0}{(\zeta^2-\rho)}$ points $\{(x_i, z_{ij}) : 1 \leq i \leq n_0, 1 \leq j \leq \ell_i\}$ (here $x_1, x_2, \ldots, x_{n_0}$ are the elements of the field $\text{GF}(q^m)$ and $z_{ij} \in \text{GF}(q^m)$). Using the decoding algorithm in [14], we can find all outer codewords, i.e polynomials $p \in \text{GF}(q^m)[X]$ of degree less than k/m, such that $p(x_i) \in \{z_{ij} : 1 \leq j \leq \ell_i\}$ for more than $t \stackrel{\text{def}}{=} \sqrt{(k/m-1)L}$ 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

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

Thus we can decode from e errors and s erasures provided

$$\begin{aligned}
\frac{q}{q-1}\cdot e + s &\leq \Big(1 - \sqrt{\frac{(k/m-1)}{n_0}\frac{L}{n_0}}\Big)(1-\zeta)n_0 n_1 \\
\Leftarrow \frac{q}{q-1}\cdot e + s &\leq \Big(1 - \sqrt{\delta(1-\sigma)\frac{(1-\rho)}{(\zeta^2-\rho)}}\Big)(1-\zeta)n.
\end{aligned}$$

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

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

Corollary 4 [Simpler Construction than that of Corollary 3]: *For any finite field* $\mathbb{F}_q$ *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 9 and Proposition 4 with the following choice of parameters: $\delta = O(\gamma^4)$, $\rho = \sqrt{\delta}$, $\varepsilon = 1$. □

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] N. Alon, O. Goldreich, J. Håstad and R. Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures and Algorithms*, 3 (1992), pp. 289-304.

[5] L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv. Optimal decoding of linear codes minimizing symbol error rate. *IEEE Trans. Information Theory*, Vol. 20, pp. 284-287, 1974.

[6] P. Elias. List decoding for noisy channels. *Wescon Convention Record*, Part 2, Institute of Radio Engineers (now IEEE), pp. 94-104, 1957.

[7] P. Elias. Error-correcting codes for list decoding. *IEEE Trans. Info. Theory*, **37** (1), pp. 5-12, 1991.

[8] G. D. Forney. *Concatenated Codes.* MIT Press, Cambridge, MA, 1966.

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

[10] 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.

[11] 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.

[12] 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.

[13] V. Guruswami, J. Håstad, M. Sudan and D. Zuckerman. Combinatorial Bounds for List Decoding. *Proc. of the 38th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, October 2000.

[14] 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.*

[15] V. Guruswami and M. Sudan. List decoding algorithms for certain concatenated codes. *Proc. of the 32nd Annual ACM Symposium on Theory of Computing (STOC)*, May 2000, pp. 181-190.

[16] V. Guruswami and M. Sudan. *The Johnson Bound: Revisited and Improved.* Manuscript in preparation, December 2000.

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

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

[19] S. R. Kumar and D. Sivakumar. Proofs, codes, and polynomial-time reducibilities. *Proc. of 14th Annual IEEE Conference on Computation Complexity*, 1999.

[20] R. Koetter and A. Vardy. Algebraic soft-decision decoding of Reed-Solomon codes. Manuscript, May 2000. (Preliminary versions appeared in *IEEE International Symposium on Information Theory*, Sorrento, Italy, June 2000, and the *38th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, October 2000.)

[21] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes.* Amsterdam: North Holland, 1977.

[22] Y. I. Manin and S. G. Vladut. Linear codes and modular curves. *J. Soviet. Math.*, 30 (1985), pp. 2611-2643.

[23] 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.

[24] R. R. Nielsen. Decoding concatenated codes using Sudan's algorithm. Manuscript submitted for publication, May 2000.

[25] A. Samorodnitsky. Personal communication, March 2000.

[26] M. A. Shokrollahi and H. Wasserman. List decoding of algebraic-geometric codes. *IEEE Trans. on Information Theory*, Vol. 45, No. 2, March 1999, pp. 432-437.

[27] 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.

[28] M. Sudan, L. Trevisan and S. Vadhan. Pseudorandom generators without the XOR lemma. In *Proc. of STOC'99*, pp. 537-546.

[29] 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.

[30] Y. Sugiyama, M. Kasahara, S. Hirasawa and T. Namekawa. Superimposed concatenated codes. *IEEE Trans. Inform. Theory*, 26 (1980), pp. 735-736.

[31] 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.

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

[33] J. M. Wozencraft. List Decoding. *Quarterly Progress Report*, Research Laboratory of Electronics, MIT, Vol. 48 (1958), pp. 90-95.

[34] V. V. Zyablov and M. S. Pinsker. List cascade decoding. In *Prob. Information Transmission*, **17** (4), pp. 29-34 (in Russian), 1981; pp. 236-240 (in English), 1982.

A Fourier Transforms over a q-ary alphabet

Proposition 10 *Let $f : \mathbb{F}_q^m \to \mathbb{F}_q$ be an arbitrary function, and for every $\alpha \in \mathbb{F}_q^m$, let the linear function $l_\alpha : \mathbb{F}_q^m \to \mathbb{F}_q$ be defined by: $l_\alpha(x) = \sum_{i=1}^m \alpha_i x_i$ (all operations performed over $\mathbb{F}_q$). Then*

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

Remark: For the case $q = 2$, $(1 - 2\mathsf{Dist}(f, l_\alpha))$ 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 [18], 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 $f : \mathbb{F}_q^m \to \mathbb{F}_q$ as a q^m-tuple over $\mathbb{F}_q$, 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 $\mathbb{F}_q$ 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 $f : \mathbb{F}_q^m \to \mathbb{F}_q$ as a string over $\mathbb{F}_q$, 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 $\mathbb{F}_q^m$ 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

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

Corollary 5 *Suppose $f : \mathbb{F}_q^m \to \mathbb{F}_q$ is a string of q^m symbols over $\mathbb{F}_q$ 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*

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

Proof: As in the proof of Proposition 10, 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
$$\sum_\alpha (1-s-\frac{q}{q-1}e_\alpha)^2 = \sum_\alpha \langle f, l_\alpha\rangle^2 \le \langle f, f\rangle = (1-s). \qquad \square$$

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 [14]. The algorithm as presented in [14] 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 $w_1 \le w_2 \le \cdots \le w_n$.)
OUTPUT: All polynomials p of degree less than k such that
$$\sum_{i:p(x_i)=y_i} w_i \ge t.$$

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

Theorem 11 ([14]) *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 12 *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 $L \ge \frac{n}{\varepsilon}$, and form the integer weights $w_i' = \lfloor L w_i / w_n \rfloor$. Since $w_i \le w_n$ for all i, the weights w_i' are all at most L, and now the algorithm of [14] 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 $Lw_i/w_n \ge w_i' \ge Lw_i/w_n - 1$, this implies we find in $\mathrm{poly}(nL) = \mathrm{poly}(n, 1/\varepsilon)$ time all polynomials p of degree less than k, provided
$$\sum_{i:p(x_i)=y_i} \left(\frac{Lw_i}{w_n} - 1\right) \quad > \quad \sqrt{(k-1)\sum_{i=1}^n \left(\frac{Lw_i}{w_n}\right)^2}$$

$$\Longleftarrow \quad \sum_{i:p(x_i)=y_i} w_i \quad > \quad \sqrt{(k-1)\sum_{i=1}^{n} w_i^2 + \frac{nw_n}{L}}$$

$$\Longleftarrow \quad \sum_{i:p(x_i)=y_i} w_i \quad > \quad \sqrt{(k-1)\sum_{i=1}^{n} w_i^2 + \varepsilon w_n}$$

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

As already remarked in the introduction, the weighted polynomial reconstruction routine from [14] is also at the heart of the soft-decision Reed-Solomon decoding algorithm reported recently by Koetter and Vardy [20]. Nielsen [24] has also used it in decoding concatenated schemes with an outer Reed-Solomon code up to half the (designed) minimum distance (and for some specific inner codes even beyond half the designed distance).

Weighted list decoding of Algebraic-geometric codes: Though it is not explicitly stated in [14], 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 $\mathbb{F}_q$. 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 [14], together with the trick of Proposition 12 above, imply the following.

Proposition 13 *Let C be an algebraic-geometric code of blocklength n defined over $\mathbb{F}_q$ with rational points $\{x_0, x_1, \ldots, x_n\}$ and the space L_{α,x_0} of functions; the designed distance d of 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 $y_i \in \mathbb{F}_q$. 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 C hold (see [14] for details on these assumptions).* $\square$