List Decoding from Erasures: Bounds and Code Constructions

Venkatesan Guruswami*

University of California at Berkeley
Computer Science Division
Berkeley, CA 94720.
Email: venkat@theory.lcs.mit.edu

Abstract. We consider the problem of list decoding from erasures. We establish lower and upper bounds on the rate of a (linear) code that can be list decoded with list size L when up to a fraction p of its symbols are adversarially erased. Our results show that in the limit of large L, the rate of such a code approaches the capacity $(1-p)$ of the erasure channel. Such nicely list decodable codes are then used as inner codes in a suitable concatenation scheme to give a uniformly constructive family of asymptotically good binary linear codes of rate $\Omega(\varepsilon^2/\lg(1/\varepsilon))$ that can be efficiently list decoded using lists of size $O(1/\varepsilon)$ from up to a fraction $(1-\varepsilon)$ of erasures. This improves previous results from [14] in this vein, which achieved a rate of $\Omega(\varepsilon^3 \lg(1/\varepsilon))$.

Keywords: Error-correcting codes, Linear codes, Decoding algorithms, List decoding, Erasure channel.

1 Introduction

The list decoding problem for an error-correcting code consists of outputting a list of all codewords that lie within a certain Hamming distance of the received word [6, 19]. The decoding is considered successful as long as the correct codeword is included in the list. If the decoder is restricted to output lists of size L, the decoding is called list-of-L decoding. List decoding with even moderate-sized lists allows one to correct significantly more errors than possible with unambiguous decoding (cf. [7, 13, 11]).

In this paper, we consider the problem of list decoding from *erasures*, i.e. list decoding when a certain fraction of the symbols of the received word are *adversarially* erased. Specifically, we are interested in codes with non-trivial list decoding performance – i.e. codes of "large" rate that are list decodable using "small" lists from a "large" number of erasures. Such a study has already been carried out with some success for the case of errors [7, 20, 4, 11]. Though the case

* Work done while the author was at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

of erasures is algorithmically always easier to handle, it raises several interesting combinatorial questions which this paper addresses. Also, the simpler situation of erasures makes better tradeoffs (between rate and list decodability) possible than for the errors case. For example, for binary codes the maximum fraction of errors one can hope to correct from is $(1/2 - \varepsilon)$, where as decoding up to a fraction $(1-\varepsilon)$ of erasures is possible. The rate (as a function of ε) that one can achieve for the erasures case also turns out be better than what is possible for the errors case.

1.1 Definitions and Notation

We focus on linear codes throughout this paper. For simplicity we also restrict our attention to binary codes though all our results go through for larger alphabets as well. Recall that a binary linear code C of blocklength n, dimension k and minimum distance d, denoted an $[n, k, d]_2$ code, is a k-dimensional subspace of $\mathbb{F}_2^n$ such that the minimum Hamming weight of a non-zero element (called codeword) of C equals d. When we omit the distance, we refer to the code as an $[n, k]_2$ code. The rate of the code is defined to be the ratio k/n and the relative distance is defined to be d/n. The main thrust of this paper is the asymptotic performance of codes, and we therefore focus on infinite families $\mathcal{C} = \{C_i\}_{I \geq 1}$ of $[n_i, k_i, d_i]_2$ codes of increasing blocklength $n_i \to \infty$. The rate $R(\mathcal{C})$ of such a family is defined to be $R(\mathcal{C}) \stackrel{\text{def}}{=} \liminf_i \left\{\frac{k_i}{n_i}\right\}$ and its relative distance is defined to be $\delta(\mathcal{C}) \stackrel{\text{def}}{=} \liminf_i \left\{\frac{d_i}{n_i}\right\}$.

We next define the list decoding radius for recovering from erasures. For $y \in \{0,1\}^n$ and $T \subseteq \{1, 2, \ldots, n\}$, define $[y]_T \in \{0,1\}^{|T|}$ to be the projection of y onto the coordinates in T. A code C of blocklength n is said to be (s, L)-*erasure list-decodable* if for every $r \in \{0,1\}^{n-s}$ and every set $T \subseteq \{1, 2, \ldots, n\}$ of size $(n-s)$, we have $|\{c \in C : [c]_T = r\}| \leq L$. In other words, given any received word with at most s erasures, the number of codewords consistent with the received word is at most L. Note that a code of minimum distance d is $(d-1, 1)$-erasure list-decodable, but is *not* $(d, 1)$-erasure list-decodable.

For linear codes, list decoding from erasures is algorithmically easy as it just amounts to finding the space of solutions to a linear system. Thus, if a linear code is (s, L)-erasure list-decodable for some small L, then a list of size at most L can also be *efficiently output* given any received word with at most s erasures. (This is not true for the case when there are errors, where algorithmic list decodability is potentially a lot more difficult to achieve than combinatorial list decodability – in fact there are known constructions of codes with good combinatorial list decodability, but for which efficient list decoding algorithms are not known.) Since our focus in this paper is on recovery from erasures, we only focus on linear codes with good combinatorial erasure list decodability, and this automatically implies a polynomial time list decoding algorithm from a large number of erasures. But for our main explicit code construction, we do mention how the decoding time can be improved beyond the time required to solve a linear system.

Definition 1 (Erasure List Decoding Radius) *For an integer $L \geq 1$ and a code C, the* list-of-L erasure decoding radius *of C, denoted* $\mathrm{radius}_L(C)$, *is defined to be the maximum value of s for which C is (s, L)-erasure list-decodable.*

Definition 2 (ELDR for Code Families) *For an infinite family $\mathcal{C} = \{C_i\}_{i\geq 1}$ of codes and an integer $L \geq 1$, the* list-of-L erasure decoding radius *of $\mathcal{C}$, denoted $\mathsf{ELDR}_L(\mathcal{C})$, is defined to be*

$$\mathsf{ELDR}_L(\mathcal{C}) \stackrel{\text{def}}{=} \liminf_i \left\{ \frac{\mathrm{radius}_L(C_i)}{n_i} \right\} , \tag{1}$$

where n_i is the blocklength of C_i.

We now define the function which measures the trade-off achievable between rate and erasure list decoding radius.

Definition 3 *For an integer L and $0 \leq p \leq 1$, the maximum rate of a code family with list-of-L erasure decoding radius at least p, denoted $R_L(p)$, is defined as*

$$R_L(p) \stackrel{\text{def}}{=} \sup_{\mathcal{C}:\ \mathsf{ELDR}_L(\mathcal{C})\geq p} R(\mathcal{C}) . \tag{2}$$

1.2 Main Results

Combinatorial Bounds One of the objectives of this paper is to establish lower and upper bounds on this function $R_L(p)$ (Theorem 1 and Theorem 2). Since the list-of-1 ELDR of a code family equals its relative distance δ, $R_1(p) = R(\delta)$ is the central function in coding theory that studies the trade-off between the rate and minimum distance of a code. One of the consequences of results of this paper is that in the limit of large L, the function $R_L(p)$ tends to $1 - p$, thus matching the singleton bound. This result has the following nice interpretation. It is a classical result in coding theory that the capacity of the erasure channel where each codeword symbol is randomly and independently erased with probability p, equals $(1 - p)$ [5]. Thus our results show that using list decoding with large enough (but constant-sized) lists, one can approach capacity even if the symbols that are erased are *adversarially* (and not randomly) chosen. Our upper bound on $R_L(p)$ also shows that $R_L(p) < 1 - p$ for every p and every fixed list size L (this result holds even for general, non-linear codes). Thus one *needs* unbounded list size in order to approach the capacity of the adversarial erasure channel using list decoding. We remark that analogous statements also hold for the errors case from earlier results in [7, 20, 4, 11].

Code Constructions For any constant $\varepsilon > 0$, our lower bound on $R_L(p)$ implies the existence of a binary linear code family $\mathcal{C}_\varepsilon = \{C_i\}_{i\geq 1}$ of rate $\Omega(\varepsilon/\lg(1/\varepsilon))$ such that each C_i is $((1-\sigma)n_i, O(1/\sigma))$-erasure list-decodable for every σ satisfying $\varepsilon \leq \sigma \leq 1$. One can use such codes as inner codes in a concatenated scheme with outer code being any code of rate $\Omega(\varepsilon)$ and relative distance $1 - O(\varepsilon)$

(like Reed-Solomon codes, or, if one wants codes over constant alphabet size, algebraic-geometric codes or the expander based codes of [2]). Such a concatenation scheme gives a *uniformly constructive* family of binary linear codes of rate $\Omega(\varepsilon^2/\lg(1/\varepsilon))$ that can be list decoded from a fraction $(1-\varepsilon)$ of erasures using lists of size $O(1/\varepsilon)$.[1] The previous best result in this vein from [14] achieved a rate of $\Omega(\varepsilon^3)$, by exploiting the fact that any family of codes with relative distance greater than $(1-\varepsilon)/2$ has a small list size for decoding from $(1-\varepsilon)$ erasures. [2] The best construction of such high distance codes achieves a rate of $\Omega(\varepsilon^3)$ (see [2]). We stress that the novelty of our result is that it is **not** obtained by appealing to this distance to ELDR relation, and indeed no polynomial time constructions of binary code families of relative distance $(1/2 - O(\varepsilon))$ and rate about ε^2 are known. In fact such a construction (which will asymptotically match the Gilbert-Varshamov bound at low rates) will be a *major* breakthrough in coding theory.

Moreover, by using Reed-Solomon codes as outer codes, we can get a code with slightly worse construction time, but which can be encoded in near-linear time and decoded in near-quadratic time. The decoding is based on a list decoding algorithm due to [13] and its fast implementation due to [16, 17].

2 Combinatorial Bounds on $R_L(p)$

2.1 Lower bound on $R_L(p)$

All logarithms in this paper are to the base 2 and will be denoted by $\lg x$. The following folklore combinatorial lemma gives a useful linear-algebraic characterization of when a linear code is (s, L)-erasure list-decodable.

Lemma 1. *An $[n, k]_2$ linear code C is (s, L)-erasure list-decodable if and only if its $n \times k$ generator matrix G has the property that every $(n - s) \times k$ submatrix of G has rank at least $(k - \lfloor \lg L \rfloor)$.*

Proof: Let $T \subseteq \{1, 2, \ldots, n\}$ with $|T| = n - s$ and $r \in \{0,1\}^{n-s}$, the number of codewords $c \in C$ with $[c]_T = r$ is precisely the number of solutions $x \in \{0,1\}^k$ to the system $G_T x = r$ where G_T is the submatrix of G consisting of all rows indexed by elements in T. By standard linear algebra, the number of solutions x to the linear system $G_T x = \mathbf{0}$ is precisely 2^ℓ where $\ell = k - \mathrm{rank}(G_T)$, and for any $r \in \{0,1\}^{n-s}$, the number of solutions x to $G_T x = r$ is at most 2^ℓ (in fact, it is always either 0 or 2^ℓ). Hence C is (s, L)-erasure list-decodable if and only if for every $T \subseteq \{1, \ldots, n\}$ with $|T| = n - s$, G_T has rank at least $k - \lfloor \lg L \rfloor$. □

[1] The term "uniformly constructive" means a polynomial time construction where the exponent of the polynomial is an absolute constant independent of ε.

[2] More generally, any binary code of relative distance δ has a small (linear in block-length) number of codewords for up to a fraction 2δ of erasures, and this is the best possible in the sense that for every $\gamma > 0$, there exist linear codes C_γ of relative distance δ, and a received word y with a fraction $(2 + \gamma)\delta$ of erasures, such that C_γ has exponentially many codewords consistent with y.

Theorem 1. *For every integer $L \geq 1$ and every p, $0 \leq p \leq 1$, we have*

$$R_L^{\text{lin}}(p) \geq 1 - p - \frac{H(p)}{\lceil \lg(L+1) \rceil} \ . \tag{3}$$

The above theorem follows from Part (i) of the stronger claim of the lemma below. The second part of the lemma will be used in the application to concatenated codes in Section 3. It asserts the existence of codes that exhibit a "gradual" increase in list size as more and more codeword symbols are erased, up to a total of $(1-\varepsilon)n$ erasures.

Lemma 2. *For all large enough n, the following hold:*

(i) For every integer $L \geq 1$ and every p, $0 \leq p \leq 1$, there exists an $[n,k]_2$ linear code C with $k = \lfloor \big(1 - p - \frac{H(p)}{\lceil \lg(L+1) \rceil}\big)n - \sqrt{n} \rfloor$ that is (pn, L)-erasure list-decodable.

(ii) There exists absolute constants $A, B > 0$ such that for every ε, $0 \leq \varepsilon \leq 1$, there exists an $[n,k]_2$ linear code C with $k = \lfloor \frac{\varepsilon n}{\lg(A/\varepsilon)} \rfloor$ that is $(s, \frac{Bn}{n-s})$-erasure list-decodable for every *$s \leq (1-\varepsilon)n$.*

Proof: We first prove Part (i) of the lemma. The proof is based on the probabilistic method. We will pick a code C generated by a random $n \times k$ generator matrix G where k is as specified in the statement of the lemma.[3] We will prove that such a random code is (pn, L)-erasure list-decodable with high probability. By Lemma 1, this amounts to estimating the probability that a random $n \times k$ matrix over $\mathbb{F}_2$ has the property that every $t \times k$ submatrix has rank at least $(k - \lfloor \lg L \rfloor)$, where $t = (1-p)n$. For any J', $0 \leq J' \leq k$, the probability that a random matrix M of order $t \times k$ has rank $(k - J')$ is at most $2^{kJ'} 2^{-tJ'}$. This follows since for a fixed subspace S of $\mathbb{F}_2^k$ of dimension $(k - J')$, the probability that all t rows of M lie in S is at most $2^{-J't}$, and furthermore the number of subspaces of $\mathbb{F}_2^k$ of dimension $(k-J')$ is at most $2^{kJ'}$ as one can specify such a subspace as the null-space of a $J' \times k$ matrix over $\mathbb{F}_2$. Let $J \stackrel{\text{def}}{=} \lceil \lg(L+1) \rceil = 1 + \lfloor \lg L \rfloor$. Then, for $t \geq k$, the probability that a random matrix M of order $t \times k$ has rank $(k - J)$ or less is at most

$$\sum_{J'=J}^{k} 2^{(k-t)J'} \leq k 2^{(k-t)J} \ .$$

Now, by a union bound, the probability that some $t \times k$ submatrix of G has rank at most $(k - J)$ is at most

$$\begin{aligned} \binom{n}{t} \cdot k 2^{(k-t)J} &= \binom{n}{(1-p)n} \cdot k 2^{(k-t)J} \\ &\leq k \cdot 2^{H(p)n} \cdot 2^{kJ - (1-p)nJ} \end{aligned}$$

[3] We will assume for simplicity that C has dimension k, i.e. G has full column rank, since this happens with very high probability.

$$
\begin{aligned}
&\le k \cdot 2^{H(p)n} \cdot 2^{\left((1-p)J - H(p)\right)n} \cdot 2^{-J\sqrt{n}} \cdot 2^{-(1-p)nJ} \\
&= k \cdot 2^{-J\sqrt{n}} \\
&= o(1) ,
\end{aligned}
$$

where in the second step we have used the fact that $\binom{n}{(1-p)n} = \binom{n}{pn} \le 2^{H(p)n}$ (cf. [18, Chap. 1, Theorem 1.4.5]), and in the third step we substituted the value of k from the statement of Part (i) of the lemma.

Hence an $n \times k$ matrix G every $t \times k$ submatrix of which has rank at least $k - J + 1 = k - \lfloor \lg L \rfloor$ exists, and by Lemma 1, therefore, there exists a (pn, L)-erasure list-decodable code of blocklength n and dimension k. This proves Part (i) of the lemma.

To prove Part (ii), we apply the above proof for $p = (1-\varepsilon)$, and $k = \lfloor \frac{\varepsilon}{\lg(8/\varepsilon)} n \rfloor$. Let us estimate the probability that for a fixed s, where $n/2 \le s \le (1-\varepsilon)n$, C is not $(s, \frac{8n}{n-s})$-erasure list-decodable. By Lemma 1, this happens only if some $(n-s) \times k$ submatrix of G has rank less than $k - \lg(\frac{8n}{n-s})$. Let $\sigma = (n-s)/n$; we have $\varepsilon \le \sigma \le 1/2$. As in the proof of Part (i), the probability that some $(n-s) \times n$ submatrix of G has rank less than $k - \lg(8/\sigma)$ is at most

$$
\begin{aligned}
\binom{n}{n-s} \cdot k \cdot 2^{(k-n+s)\lg(8/\sigma)} &\le k \cdot \left(\frac{e}{\sigma}\right)^{\sigma n} \cdot 2^{\left((\varepsilon/\lg(8/\varepsilon)) - \sigma\right)\lg(8/\sigma)n} \\
&= k \cdot 2^{-n\left(\sigma \lg(8/e) - \frac{\varepsilon \lg(8/\sigma)}{\lg(8/\varepsilon)}\right)} \\
&= 2^{-\Theta(n)}
\end{aligned}
$$

where the last step follows since $\sigma \ge \varepsilon$, and in the first step we used the inequality $\binom{n}{\sigma n} \le (e/\sigma)^{\sigma n}$ for $\sigma \le 1/2$.

Now, by a union bound, the probability that for some s, $n/2 \le s \le (1-\varepsilon)n$, C is not $(s, \frac{8n}{n-s})$-erasure list-decodable is also exponentially small. Hence there exists a linear code C of blocklength n and rate $\varepsilon/\lg(8/\varepsilon)$ that is $(s, \frac{8n}{n-s})$-erasure list-decodable for every s that satisfies $n/2 \le s \le (1-\varepsilon)n$. Since the list size for $s < n/2$ erasures is at most the list size for $n/2$ erasures, such a code is also $(s, \frac{16n}{n-s})$-erasure list-decodable for every s, $0 \le s \le (1-\varepsilon)n$. This proves Part (ii) of the claim (with the choice $A = 8$ and $B = 16$ for the absolute constants). □

Remark: Note that the above lemma not only guarantees the existence of codes with the required properties, but also proves that a random code has these properties with very high probability.

2.2 Upper bound on $R_L(p)$

We now turn to upper bounds on $R_L(p)$. It is easy to prove that for any fixed L (in fact even for a list size L that is allowed to grow polynomially in the blocklength), we must have $R_L(p) \le 1-p$. Indeed, let C be a code of blocklength n and rate r, and let $T = \{1, 2, \ldots, (1-p)n\}$. Pick a *random* $y \in \{0,1\}^{(1-p)n}$

and consider the set S_y of codewords $c \in C$ that satisfy $[c]_T = y$. The expected number of such codewords equals $2^{rn}2^{-(1-p)n}$, and hence there must exist a $y \in \{0,1\}^n$ for which $|S_y| \geq 2^{(r-(1-p))n}$. Since we want $|S_y| \leq L \leq \mathrm{poly}(n)$, we must have $r \leq (1-p)$. Hence $R_L(p) \leq 1-p$. Below, we are interested in a better upper bound on $R_L(p)$, which in particular bounds it strictly away from $(1-p)$ for every fixed L (and thereby shows that one requires unbounded list size in order to approach the "capacity" $(1-p)$ of the erasure channel). The proof uses standard techniques similar to those used in the Plotkin and Elias-Bassalygo bounds for the rate vs. minimum distance trade-off. It will appear in the full version of the paper and can also be found in [10, Chapter 10].

Theorem 2. *For every integer $L \geq 1$ and every p, $0 \leq p \leq 1 - 2^{-L}$, we have*

$$R_L(p) \leq \min\Big\{1 - H(\lambda), 1 - \frac{p}{1-2^{-L}}\Big\} \ , \tag{4}$$

where λ is the unique root of the equation $\lambda^{L+1} + (1-\lambda)^{L+1} = 1-p$ in the range $0 \leq \lambda \leq 1/2$. For $p \geq 1 - 2^{-L}$, we have $R_L(p) = 0$.

Corollary 3 *For every integer L and every p, $0 < p < 1$, $R_L(p) < 1 - p$.*

Remark: For $L = 1$, the two bounds proven in Theorem 2 are precisely the Elias-Bassalygo and Plotkin bounds on the rate of a code in terms of its minimum distance, which state that $R(\delta) \leq 1 - H(\frac{1-\sqrt{1-2\delta}}{2})$ and $R(\delta) \leq 1 - 2\delta$, respectively. For large L, the bound $R_L(p) \leq 1 - p/(1-2^{-L})$ is better than the other bound $R_L(p) \leq 1 - H(\rho)$ *except for* very small p (less than about $L/2^L$).

3 Application to Concatenated Codes

We now use the codes guaranteed by Lemma 2 (specifically Part (ii) of the lemma) as inner codes in a suitable concatenation scheme to construct binary code families of rate $\Omega(\varepsilon^2/\lg(1/\varepsilon))$ that can be efficiently list decoded using lists of size $O(1/\varepsilon)$ when up to a fraction $(1-\varepsilon)$ of its symbols are erased. This follows from the following theorem (Theorem 4), which is the main result of this section. The proof uses the simple, yet powerful, tool of code concatenation [8]. Given an $[N, K, D]_{2^m}$ code C_1, and an $[n, m, d]_2$ code C_2, the concatenated code $C_1 \oplus C_2$ is an $[Nn, Km, \geq Dd]_2$ binary linear code whose codewords are obtained by encoding a message x first using C_1, and then encoding each of the symbols of $C_1(x)$ using C_2 (note that each symbol of $C_1(x)$ is an element of $\mathrm{GF}(2^m)$ and can hence be interpreted as an m-bit vector, and m is exactly the dimension of C_2).

Theorem 4. *For every $\varepsilon > 0$, there exists a family of binary linear codes of rate $\Omega(\varepsilon^2/\log(1/\varepsilon))$ such that a code of blocklength in the family can be constructed in $2^{\varepsilon^{-O(1)}} + \mathrm{poly}(N, 1/\varepsilon)$ time, and can be list decoded in polynomial time using lists of size $O(1/\varepsilon)$ when up to a fraction $(1-\varepsilon)$ of its symbols are erased.*

The above result follows immediately from the following lemma.

Lemma 3. *There exist absolute constants b, d such that for all large enough integers K and all small enough $\varepsilon > 0$, there exists a binary linear code C_K that has the following properties:*

(i) *C_K has dimension K and blocklength $N \le \frac{bK\log(1/\varepsilon)}{\varepsilon^2}$.*
(ii) *The generator matrix of C_K can be constructed in $2^{\varepsilon^{-O(1)}} + \mathrm{poly}(N, 1/\varepsilon)$ time.*
(iii) *C_K is $((1-\varepsilon)N, \frac{d}{\varepsilon})$-erasure list-decodable (and since C_K is linear there is an $O(N^3)$ time list decoding algorithm to decode up to $(1-\varepsilon)N$ erasures using lists of size $O(1/\varepsilon)$).*

Proof: The code C_K is constructed by concatenating an outer code C_{out} over $\mathrm{GF}(2^m)$ of blocklength n_0, dimension $k_0 = K/m$ and minimum distance d_0, with an inner code C_{in} as guaranteed by Part (ii) of Lemma 2. By using a construction in [2], we can pick parameters so that $k_0 = K/m = \Omega(\varepsilon n_0)$, $m = O(1/\varepsilon)$ and $d_0 = (1-O(\varepsilon))n_0$ (for convenience, we hide constants using big-Oh notation, but we stress that these are absolute constants that do not depend on ε).[4] We note that this choice of C_{out} is not linear over $\mathrm{GF}(2^m)$ (though it *is* additive), but a (succinct) description of C_{out} can be constructed in time $\mathrm{poly}(n, 1/\varepsilon)$. Moreover, after concatenation with an inner binary linear code, the overall concatenated code will be a binary linear code.

The inner code C_{in} will be a code as guaranteed by Lemma 2 of dimension m and blocklength $n_1 = O(\frac{m\lg(1/\varepsilon)}{\varepsilon})$ that is $((1-\sigma)n_1, B/\sigma)$-erasure list-decodable for every $\sigma \ge \varepsilon/2$. We can construct such a code by a brute-force search in $2^{O(mn_1)} = 2^{\varepsilon^{-O(1)}}$ time. Once we have descriptions of C_{out} and C_{in}, one can construct the generator matrix of the concatenated code C_K in $\mathrm{poly}(N, 1/\varepsilon)$ time where $N = n_0 n_1$ is the blocklength of C_K. Now $N = n_0 \cdot n_1 = O(\frac{K}{m\varepsilon}) \cdot O(\frac{m\lg(1/\varepsilon)}{\varepsilon}) = O(\frac{K\lg(1/\varepsilon)}{\varepsilon^2})$. This proves Parts (i) and (ii) of the statement of the lemma.

We now prove that C_K is $((1-\varepsilon)N, O(1/\varepsilon))$-erasure list-decodable. Let y be a received word with $(1-\varepsilon)N$ erasures, and let $c'_1, c'_2, \ldots, c'_M$ be the codewords in C_K "consistent" with y, and let c_j be the codeword of C_{out} corresponding to c'_j, for $1 \le j \le M$. We wish to prove that $M = O(1/\varepsilon)$. Let y_i be the portion of y corresponding to the i'th outer codeword position, for $1 \le i \le n_0$. Let the number of symbols in y_i be $\sigma_i n_1$. We have $\sum_i \sigma_i = \varepsilon n_0$. Define $Q = \{i : \sigma_i \ge \varepsilon/2\}$. Clearly we have

$$\sum_{i \in Q} \sigma_i \ge \frac{\varepsilon n_0}{2} . \tag{5}$$

[4] Actually, we can use certain families of algebraic-geometric codes and even have $m = O(\log(1/\varepsilon))$. But we prefer to use the codes from [2] as they are easier to construct and have lower construction complexity, and we do not want to give the impression that we need complicated AG-codes for our construction. We will highlight the improvement in performance attainable using AG-codes in Section 3.2.

Now define "weights" $w_{i,\beta}$ for $1 \le i \le n_0$ and $\beta \in \mathrm{GF}(2^m)$ as follows. If $i \notin Q$, set $w_{i,\beta} = 0$ for all $\beta \in \mathrm{GF}(2^m)$. If $i \in Q$, then $\sigma_i \ge \varepsilon/2$ and hence by construction C_{in} is $((1-\sigma_i)n_1, B/\sigma_i)$-erasure list-decodable. In other words, if $J_i = \{\beta : C_{\text{in}}(\beta) \text{ is consistent with } y_i\ \}$, then $|J_i| \le B/\sigma_i$. We set (for $i \in Q$):

$$w_{i,\beta} = \begin{cases} \sigma_i & \text{if } \beta \in J_i \\ 0 & \text{if } \beta \notin J_i \end{cases}$$

Since $|J_i| \le B/\sigma_i$, our choice of weights clearly satisfy

$$\sum_{i,\beta} w_{i,\beta}^2 \le B \sum_{i \in Q} \sigma_i \ . \tag{6}$$

We now use a combinatorial result that gives an upper bound on the number of codewords of C_{out} that satisfy a certain weighted condition depending on the distance d_0 of C_{out}. This result is a "weighted" version of the classical "Johnson bounds", and appears, for example, in [15] and [10, Section 3.4]. Specifically, this result states that, for any $\varepsilon' > 0$, the number of codewords $(\alpha_1, \alpha_2, \ldots, \alpha_{n_0}) \in C_{\text{out}}$ that satisfy

$$\sum_{i=1}^{n_0} w_{i,\alpha_i} \ge \Big((n_0 - d_0(1-\varepsilon')) \sum_{i,\beta} w_{i,\beta}^2 \Big)^{1/2} \tag{7}$$

is at most $1/\varepsilon'$.

Now for each c_j, $1 \le j \le M$, and each $i \in Q$, we have $c_{j,i} \in J_i$, where $c_{j,i} \in \mathrm{GF}(2^m)$ denotes the i'th symbol of c_j. Now, $w_{i,c_{j,i}} = \sigma_i$ for every $i \in Q$ and $w_{i,c_{j,i}} = 0$ for $i \notin Q$. Thus, we have

$$\sum_{i=1}^{n_0} w_{i,c_{j,i}} = \sum_{i \in Q} \sigma_i \ . \tag{8}$$

Combining Equations (5), (6) and (8), we have that the codeword c_j satisfies Condition (7) as long as

$$\frac{\varepsilon n_0}{2} \ge (n_0 - d_0(1-\varepsilon'))B \ ,$$

which can be satisfied provided $d_0 \ge n_0\big(1 - \frac{\varepsilon}{2B}\big)(1-\varepsilon')^{-1}$. Picking $\varepsilon' = \frac{\varepsilon}{4B}$, we only need $d_0 = n_0(1 - O(\varepsilon))$ which is satisfied for our choice of C_{out}. Hence the number of codewords consistent with the received word y is at most $1/\varepsilon' = O(1/\varepsilon)$, thus proving Part (iii) of the lemma as well. □ *(Lemma 3)*

Remark: The previous results from [14] (obtained using code constructions from [2, 3]) along the lines of the above theorem achieved a blocklength of $N = \min\{O(\frac{K^2}{\varepsilon^2}), O(\frac{K}{\varepsilon^3})\}$. Thus, our result achieves the best features of both these bounds and gets $N = \tilde{O}(K/\varepsilon^2)$ (hiding the $\lg(1/\varepsilon)$ factor).

3.1 Obtaining near-linear encoding and quadratic decoding times

The codes constructed in Lemma 3 being linear can be encoded in $O(N^2)$ time and can be list decoded in $O(N^3)$ time by solving a linear system. By using Reed-Solomon codes as outer code, the encoding can be accomplished in $\tilde{O}(N)$ time (here $\tilde{O}(N)$ denotes $O(N\mathrm{poly}(\lg N)))$. This is because the outer encoding can be performed in $\tilde{O}(N)$ using FFT based methods and then the inner encoding just takes $O(\lg^2 N)$ time for each of the N outer codeword symbols. For decoding, the inner codes can be decoded in $O(\lg^3 N)$ time by solving a linear system, and then the Reed-Solomon code can be decoded from a set of weights that satisfy Condition (7) using a near-quadratic time implementation (cf. [16,17]) of the weighted Reed-Solomon decoding algorithm from [13]. However, the code family is no longer uniformly constructive as one must now search and find a good inner code of dimension $\Omega(\lg N)$ as opposed to the construction in Lemma 3 where the inner code has dimension which is a constant dimension depending only on ε. In fact, a naive search for the inner code among all inner codes will take time which is quasi-polynomial in N. However, one can use the method of conditional expectations to achieve a deterministic construction in $N^{O(\log^2(1/\varepsilon))}$ time — we omit the details.

3.2 Improving construction time using Algebraic-geometric codes

Despite the impressive encoding and decoding times, the drawback of the above construction with Reed-Solomon codes as the outer code is the rather high exponent of N in the construction time. In particular, the code family ceases to be *uniformly constructive*, since the construction time is no longer of the form $O(f(\varepsilon)N^a)$ where a is an absolute constant independent of ε.

Instead of Reed-Solomon codes, we can use AG-codes of distance $(1-O(\varepsilon))$ as the outer code. Such codes of rate $\Omega(\varepsilon)$ exist over an alphabet size of $O(1/\varepsilon^2)$ (any family of AG-codes that meet the so-called Drinfeld-Vlădut bound will satisfy this, cf. [10, Section 6.3.9]). Now, the dimension k of the inner binary code is only $O(\log(1/\varepsilon))$, and one can deterministically find a binary linear inner code which has the required erasure list decodability properties in $2^{O(\log^3(1/\varepsilon))}$ time.

Since the overall code is linear, the claims of quadratic encoding time and cubic decoding time still holds. Thus, we can obtain the same properties as the codes from Lemma 3 with an improved construction time of $2^{O(\log^3(1/\varepsilon))} + \mathrm{poly}(N, 1/\varepsilon)$. We formally record this improvement below.

Theorem 5. *For every $\varepsilon > 0$, there exists a family of binary linear codes of rate $\Omega(\varepsilon^2/\log(1/\varepsilon))$ such that a code of blocklength in the family can be constructed in $2^{\varepsilon^{-O(1)}} + \mathrm{poly}(N, 1/\varepsilon)$ time, and can be list decoded in polynomial time using lists of size $O(1/\varepsilon)$ when up to a fraction $(1-\varepsilon)$ of its symbols are erased.*

4 Concluding Remarks

Our lower bound on $R_L(p)$ from Theorem 1 guarantees the *existence* of binary linear code families of rate $\Omega(\varepsilon)$ which can be list decoded from up to a fraction $(1-\varepsilon)$ of erasures, using lists of size, say, ε^{-2}. The construction of Theorem 4, however, only achieves a rate of about ε^2. It was recently observed in [12] that the result of Theorem 1 is actually "tight" in the "high-noise regime"– specifically they show that for linear codes, in order to have positive rate, one requires a list size of at least $\Omega(1/\varepsilon)$ to list decode from $(1-\varepsilon)$ erasures. This indicates that our results cannot be improved by the choice of a better linear code as inner code.

Furthermore, as pointed out to us by Alon [1] (see also the formal discussion in [10, Section 10.6.4]), getting a poly$(N, 1/\varepsilon)$ time construction of a binary *linear* code family of rate better than ε^2 (like $\varepsilon^{3/2}$, for example) with poly$(1/\varepsilon)$-sized lists for up to a fraction $(1-\varepsilon)$ of erasures, would imply a significant improvement for explicit constructions of certain bipartite Ramsey graphs. In this sense, improving our results further seems to be a significant challenge: there is, however, some hope since we allow for a construction that runs in time exponential in $1/\varepsilon$ and/or a list size that is exponential in $1/\varepsilon$. It is an interesting open question whether this can be exploited to beat the "ε^2 barrier" for the rate.

If one allows larger alphabets, known constructions of large distance algebraic-geometric codes imply code families of ELDR $(1-\varepsilon)$ and rate $\Omega(\varepsilon)$ over an alphabet of size $O(1/\varepsilon^2)$, even for a list size of 1. Explicit constructions of linear codes of rate better than $\Omega(\varepsilon^2)$ with ELDR $(1-\varepsilon)$ even for a slightly smaller alphabet do not appear to be known. However, in recent work, [12] present constructions of *non-linear* codes over large alphabets (of size *independent* of ε) that can be efficiently list decoded from a fraction $(1-\varepsilon)$ of erasures and which have rate better than $\Omega(\varepsilon^2)$ (in fact they have rate approaching $\Omega(\varepsilon)$ as the alphabet size grows) — details on this construction appear also in [10, Section 10.7].

It is also an interesting question whether a code construction similar to Theorem 5 can be obtained in time which is polynomial in both N and $1/\varepsilon$.

Acknowledgments

I would like to thank Piotr Indyk, Dan Spielman and Madhu Sudan for useful discussions.

References

1. N. Alon. Personal Communication, October 1999.
2. 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.
3. 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.

4. V. M. Blinovsky. Bounds for codes in the case of list decoding of finite volume. *Prob. Information Transmission*, **22** (1), pp. 11-25 (in Russian), 1986; pp. 7-19 (in English), 1986.
5. P. Elias. Coding for two noisy channels. *Information Theory*, Third London Symposium, September 1955, Buttersworth's Scientific Publications, pp. 61-76.
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, 1996.
9. 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.
10. V. Guruswami. *List Decoding of Error-Correcting Codes.* Ph.D thesis, Massachusetts Institute of Technology, August 2001.
11. V. Guruswami, J. Håstad, M. Sudan and D. Zuckerman. Combinatorial Bounds for List Decoding. *Proceedings of the 38th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, October 2000.
12. V. Guruswami and P. Indyk. Untitled Manuscript, April 2001.
13. V. Guruswami and M. Sudan. Improved Decoding of Reed-Solomon and Algebraic-geometric codes. *IEEE Transactions on Information Theory*, 45 (1999), pp. 1757-1767.
14. V. Guruswami and M. Sudan. List decoding algorithms for certain concatenated codes. *Proc. of 32nd ACM Symposium on the Theory of Computing (STOC)*, May 2000, pp. 181-190.
15. V. Guruswami and M. Sudan. The Johnson Bound: Revisited and Improved. *Submitted for publication*, December 2000.
16. R. R. Nielsen and T. Hϕholdt. Decoding Reed-Solomon codes beyond half the minimum distance. *Proceedings of the International conference on Coding Theory and Cryptography*, 1998.
17. R. Roth and G. Ruckenstein. Efficient decoding of Reed-Solomon codes beyond half the minimum distance. *IEEE Transactions on Information Theory*, **46**(1), pp. 246-257, January 2000.
18. J. H. van Lint. *Introduction to Coding Theory.* Graduate Texts in Mathematics **86**, (Third Edition) Springer-Verlag, Berlin, 1999.
19. J. M. Wozencraft. List Decoding. *Quarterly Progress Report*, Research Laboratory of Electronics, MIT, Vol. 48 (1958), pp. 90-95.
20. 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.