# Improved Decoding of Interleaved AG Codes

Andrew Brown, Lorenz Minder, and Amin Shokrollahi

Laboratoire des mathematiques algorithmiques (LMA),
Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne
{andrew.brown, lorenz.minder, amin.shokrollahi}@epfl.ch

**Abstract.** We analyze a generalization of a recent algorithm of Bleichenbacher et al. for decoding interleaved codes on the $Q$-ary symmetric channel for large $Q$. We will show that for any $m$ and any $\epsilon$ the new algorithms can decode up to a fraction of at least $\frac{\beta m}{\beta m+1}(1-R-2Q^{-1/2m})-\epsilon$ errors, where $\beta = \frac{\ln(q^m-1)}{\ln(q^m)}$, and that the error probability of the decoder is upper bounded by $O(1/q^{\epsilon n})$, where $n$ is the block-length. The codes we construct do not have a-priori any bound on their length.

## 1 Introduction

The general $Q$-ary symmetric channel of communication has not been as prominently featured in the literature as the binary symmetric channel. While the case of small $Q$ has been investigated by some authors in connection with belief-propagation algorithms, the case of large $Q$ has been largely untouched.

Perhaps one reason for this omission is the complexity of belief-propagation type algorithms which increases with the alphabet size $Q$, rendering the design of efficient decoding algorithms impossible for large $Q$. Another possible reason is the observation that for large $Q$ the code design problem can be reduced to the code design problem for the binary *erasure* channel, albeit at the expense of some loss in the rate of the transmission. This reduction is for example employed in the Internet: in this case the symbols are packets; each packet is equipped with a checksum, or more generally, a hash value. After the transmission, the hash value of each symbol is checked, and a symbol is declared as erased if the hash value does not match. If $h$ bits are used for the hash value, and if $Q = 2^{mh} = q^m$, then, each symbol's effective information rate is reduced by a factor of $(m-1)/m$. If the error rate of the $Q$-ary symmetric channel is $p$, and if the erasure code operates at a rate of $1-p-\varepsilon$ for some $\varepsilon$, then the effective rate of the transmission is about $1-(p+\varepsilon+1/m)$, and the error probability is upper bounded by $n/2^h = n/q$, where $n$ is the block-length of the erasure code, when an erasure correcting code such as a Tornado code [9] is used.

A linear time decoding algorithm for the $Q$-ary symmetric channel using LDPC codes was recently proposed by Luby and Mitzenmacher [8]. They did not exhibit codes that come arbitrarily close to the capacity of the $Q$-ary symmetric channel, but it is possible to extend their methods to find such codes [11]. In their construction, the error probability of the decoder is at most $O(n/Q)$, which can be much smaller than the error probability obtained using the hashing method.

N.P. Smart (Ed.): Cryptography and Coding 2005, LNCS 3796, pp. 37–46, 2005.

Recently, Bleichenbacher et al. [1] invented a new decoding algorithm for Interleaved Reed-Solomon Codes over the $Q$-ary symmetric channel. As the name suggests, the codes are constructed with an interleaving technique from $m$ Reed-Solomon codes defined over $\mathbb{F}_q$, if $Q = q^m$. These codes are similar to well-known product code constructions with Reed-Solomon codes as inner codes, but there is an important improvement: interleaved codes model the $Q$-ary channel more closely than a standard decoder for the product code would. It follows that interleaved codes achieve much better rates: interleaved Reed-Solomon Codes can asymptotically have rates as large as $1 - p(1 + 1/m)$, which is much more than the rate $1 - 2p$ achieved with a standard product code decoder. Bleichenbacher et al. prove that the error probability of their decoder is upper bounded by $O(n/q)$, where $n$ is the block length of the code. Compared to the hashing method, this decoder has about the same error probability, but the rate of the code is closer to the capacity of the channel.

A general method for decoding of interleaved codes has been discussed in [3]. The gist of the algorithm is to find a polynomial in $m + 1$ variables that passes through the points given by the interpolation points of the code and the coordinate positions of the received words. The polynomial can then be used to scan the received word, and probabilistically identify the incorrect positions. The method can decode up to a fraction of $1 - R - R^{m/(m+1)}$ errors, with an error probability of $O(n^{O(m)}/q)$, where $R$ is the rate of the code. Note that the error probability of this algorithm increases with $n$. Note also that this algorithm is superior to that of Bleichenbacher et al. for small rates. The interleaved decoding algorithm has also been used in conjunction with concatenated coding [6].

Another class of algorithms to which the interleaved decoding algorithm can be compared is that of list-decoding algorithms [13,12,5]. However, this comparison is not fair, since these decoding algorithms work under adversarial conditions, i.e., recover a list of closest codewords without any restriction on the noise (except the number of corrupted positions). The best known codes to-date (in terms of error-correction capability) with a polynomial time decoding algorithm are given in [10]. For these codes the authors provide a decoding algorithm which can correct up to a fraction of $1 - \epsilon$ errors with a code of length $n$ and rate $\Omega(\epsilon/\log(1/\epsilon))$ over an alphabet of size $n^{O(\log(1/\epsilon))}$. The codes provided in this paper improve upon this bound considerably, when the rate is not too small.

We have recently shown in [2] that the error probability of the decoder in [1] is in fact $O(1/q)$, independent of $n$. In this paper, we present a slightly different algorithm than that of Bleichenbacher et al. for the class of algebraic-geometric codes (AG-codes). We will show that the algorithm can successfully decode $e$ errors with an error probability that is proportional to

$$\left(\frac{1}{q}\right)^{\beta m(n-k-2g)-(\beta m+1)e}$$

where $g$ is the genus of the curve underlying the AG-code, $R$ is the rate, $Q = q^m$, and $\beta = \frac{\ln(q^m-1)}{\ln(q^m)}$.

Since the error probability of our algorithm does not increase with $n$, it is possible to consider long codes over the alphabet $\mathbb{F}_q$. In particular, using codes from asymptotically

optimal curves over $\mathbb{F}_{q^2}$ [7,4], and assuming that $m$ is large enough, our codes will be able to reliably decode over a $Q$-ary symmetric channel with error probability $p$, and maintain a rate close to $1 - p - \frac{2}{\sqrt{q}-1}$.

Despite the proximity to channel capacity we can gain with this algorithm, the construction of codes on the $Q$-ary channel with both rate close to the capacity and polynomial-time decoding complexity (where we measure complexity relative to the size of the received input, i.e. as a function of $n \log(Q)$), is still an open challenge.

In the next two sections of this paper we will introduce interleaved codes and the main decoding algorithm, and analyze this algorithm. The last section gives a detailed comparison of our method with various hashing methods. For the rest of the paper we will assume familiarity with the basic theory of AG-codes.

## 2 Interleaved AG-Codes and Their Decoding

Let $\mathcal{X}$ be an absolutely irreducible curve over $\mathbb{F}_q$, and let $D, P_1, \ldots, P_n$ denote $n+1$ distinct $\mathbb{F}_q$-rational points of $\mathcal{X}$. Let $g$ denote the genus of $\mathcal{X}$. For a divisor $A$ of $\mathcal{X}$ we denote by $\mathcal{L}(A)$ the associated linear space. The theorem of Riemann states that the dimension of this space, denoted $\dim(A)$, is at least $\deg(A) - g + 1$.

Fix a parameter $\alpha$ with $2g - 2 < \alpha < n$. A (one-point) AG-code associated to $D, P_1, \ldots, P_n$ and $\alpha$ is the image of the evaluation map $\mathrm{Ev}\colon \mathcal{L}(\alpha D) \to \mathbb{F}_q^n$, $\mathrm{Ev}(f) = (f(P_1), \ldots, f(P_n))$.

Suppose that $Q = q^m$, and let $\beta_1, \ldots, \beta_m$ denote a basis of $\mathbb{F}_Q$ over $\mathbb{F}_q$. We define a code over $\mathbb{F}_Q$ of length $n$ in the following way: the codewords are

$$\left( \sum_{j=1}^{m} f_j(P_1)\beta_j, \ldots, \sum_{j=1}^{m} f_j(P_n)\beta_j \right),$$

where $(f_1, \ldots, f_m) \in \mathcal{L}(\alpha D)^m$. (This algebraic interpretation of interleaved coding was communicated to us by A. Vardy [14].) Note that this code does not necessarily form an $\mathbb{F}_Q$-vector space, but it does form an $\mathbb{F}_q$-vector space.

Suppose that such a codeword is sent over a $Q$-ary symmetric channel, and that $e$ errors occur during the transmission. Denote by $E$ the set of these error positions. Because of the properties of the $Q$-ary symmetric channel, each of the $m$ codewords of the constituent code is independently subjected to a $q$-ary symmetric channel under the additional assumption that for each of these positions, at least one of the codewords is corrupted.

Our task is to decode the codeword. We proceed in a way similar to [1]: let $t$ be a parameter to be determined later, and let $W$ and $V$ be defined by

$$W := \begin{pmatrix} \phi_1(P_1) & \phi_2(P_1) & \cdots & \phi_d(P_1) \\ \phi_1(P_2) & \phi_2(P_2) & \cdots & \phi_d(P_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(P_n) & \phi_2(P_n) & \cdots & \phi_d(P_n) \end{pmatrix}, \quad V := \begin{pmatrix} \psi_1(P_1) & \psi_2(P_1) & \cdots & \psi_s(P_1) \\ \psi_1(P_2) & \psi_2(P_2) & \cdots & \psi_s(P_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(P_n) & \psi_2(P_n) & \cdots & \psi_s(P_n) \end{pmatrix},$$

where $\phi_1, \ldots, \phi_d$ form a basis of $\mathcal{L}\big((t+g)D\big)$, and $\psi_1, \ldots, \psi_s$ form a basis of $\mathcal{L}\big((t+g+\alpha)D\big)$. Let $\left(\sum_{j=1}^m y_{1j}\beta_j, \ldots, \sum_{j=1}^m y_{nj}\beta_j\right)$ be the received word, and let

$$A := \left(\begin{array}{ccc|c} V & & & -D_1W \\ & V & & -D_2W \\ & & \ddots & \vdots \\ & & V & -D_mW \end{array}\right), \tag{1}$$

where $D_j$ is the diagonal matrix with diagonal entries $y_{1j}, \ldots, y_{nj}$. The decoding process is now as follows:

- Find a non-zero element $v = (v_1 \mid \cdots \mid v_m \mid w)$ in the right kernel of $A$, where $v_1, \ldots, v_m \in \mathbb{F}_q^s$ and $w \in \mathbb{F}_q^d$. If $v$ does not exist, output a decoding error.
- Identify $v_1, \ldots, v_m$ with functions in the space $\mathcal{L}\big((t+g+\alpha)D\big)$, and $w$ with a function in $\mathcal{L}\big((t+g)D\big)$. If $w$ divides $v_j$ for each $j = 1, \cdots, m$, then set $f_1 = v_1/w, \cdots, f_m = v_m/w$, and output $f_1, \cdots, f_m$. Otherwise, output a decoding error.

The value of $t$ determines the error probability of the decoder, as the following main theorem suggests.

**Theorem 1.** *Suppose we have $t$ that satisfies*

$$g - 1 \le t \le \frac{\beta m}{\beta m + 1}(n - \alpha - g - 1) - \frac{c}{\beta m + 1},$$

*for some $c > 0$, and where $\beta = \frac{\ln(q^m - 1)}{\ln(q^m)}$. Let $e$ denote the number of errors incurred during transmission, and suppose that $e \le t$. Then we have:*

(1) *If $e + t < n - \alpha - g$, then the error probability of the above decoder is zero.*
(2) *For general $e \le t$ the error probability of the above decoder is at most $\frac{q}{q-1} \cdot q^{-c}$.*

This theorem will be proved in the next section.

## 3 Analysis of the Decoder

To analyze the decoder of the last section, we make the following simplifying assumptions:

(a) The error positions are $1, 2, \ldots, e$.
(b) The functions $f_1, \ldots, f_m$ sent over the channel are all zero.

It is easily seen that we can assume (a) and (b) without loss of generality. This goes without saying for (a); as for (b), note that since the code is invariant under addition, the behavior of the matrix $A$ in (1) with respect to the algorithm is the same no matter which codeword is sent.

The assumptions imply the following:

(1) For each $i = e+1, \ldots, n$, and for each $j = 1, \ldots, m$, we have $y_{ij} = 0$. Equivalently, for each $j$, the last $n-e$ diagonal entries of $D_j$ are zero.
(2) For each $i = 1, \ldots, e$, the vector $(y_{i1}, \ldots, y_{im})$ is chosen uniformly at random from $\mathbb{F}_q^m \backslash \{(0, \ldots, 0)\}$.
(3) The probability of a decoding error is upper bounded by the probability that there exists a vector $(v_1 \mid \cdots \mid v_m \mid w)$ in the right kernel of $A$ for which at least one of the $v_i$ is non-zero, plus the probability that the right kernel of $A$ is trivial.

Note that because both the number of errors and the error positions have been fixed, the only randomness in the matrix $A$ comes from the values $y_{ij}$ for $i = 1, \ldots, e$ and $j = 1, \ldots, m$.

We will show that if $e \leq t$, then the right kernel of $A$ is nontrivial. Hence, we only need to bound the probability that there exists a vector $(v_1 \mid \cdots \mid v_m \mid w)$ in the right kernel of $A$ for which at least one of the $v_i$ is non-zero. Let us call such a vector *erroneous*. Note that if the right kernel of $A$ is nontrivial and does not contain any erroneous vectors, then the algorithm is successful with probability one.

We bound the probability of the existence of an erroneous vector in the following way: for each non-zero $w \in \mathcal{L}\big((t+g)D\big)$, we calculate the expected number of $(v_1 \mid \cdots \mid v_m)$ such that $v = (v_1 \mid \cdots \mid v_m \mid w)$ is in the right kernel of $A$. An upper bound on the desired probability can then easily be obtained using Markov's inequality.

**Proof of Theorem 1.** Throughout we use the notation $v_j$ to denote both a vector in $\mathbb{F}_q^s$ and the corresponding element of $\mathcal{L}\big((t+\alpha+g)D\big)$ (obtained with the basis $\psi_1, \ldots, \psi_s$). Likewise $w$ can denote both a vector in $\mathbb{F}_q^d$ and an element of $\mathcal{L}\big((t+g)D\big)$ (with the basis $\phi_1, \ldots, \phi_d$).

First we will show that if $e \leq t$, then the right kernel of $A$ is nontrivial. To this end, note that by the Theorem of Riemann $\dim\big((t+g)D - \sum_{i=1}^{e} P_i\big) \geq t-e+1 > 0$, hence $\mathcal{L}\big((t+g)D - \sum_{i=1}^{e} P_i\big)$ is nontrivial. Let $w$ be a non-zero function in this space. Setting $v_j := wf_j \ (= 0)$, we see that the vector $v = (v_1 \mid \cdots \mid v_m \mid w)$ is in the right kernel of $A$, and is nontrivial, as required. It follows that the error probability of the decoder is upper bounded by the probability that the right kernel of $A$ contains erroneous vectors.

If $v = (v_1 \mid \cdots \mid v_m \mid w) \in \ker(A)$ then we have

$$\forall i = 1, \ldots, n, \ \forall\, j = 1, \ldots, m\colon \quad v_j(P_i) = y_{ij} \cdot w(P_i). \tag{2}$$

Furthermore, since we are assuming that the zero codeword was transmitted, we have $y_{ij} = 0$ for $i > e$ (since $i$ is not an error position). From this and (2) we can deduce that

$$\forall i = e+1, \ldots, n, \ \forall\, j = 1, \ldots, m, : \quad v_j(P_i) = 0. \tag{3}$$

This implies that

$$\forall\, j = 1, \ldots, m\colon \quad v_j \in \mathcal{L}\left((t+\alpha+g)D - \sum_{i=e+1}^{n} P_i\right) =: \mathcal{L}\big(T\big). \tag{4}$$

In particular, this proves part (1) of the theorem: if $t+\alpha+g-n+e < 0$, or equivalently, if $t+e < n-\alpha-g$, then this linear space is trivial, and hence any element in the right kernel of $A$ is non-erroneous (since it has the property that $v_j = 0$ for all $j = 1, \ldots, m$).

For $w \in \mathcal{L}((t+g)D)$ let $Z(w) = \{P_i \mid 1 \le i \le e, w(P_i) = 0\}$. Let $\overline{Z}(w)$ be its complement in $\{P_1, \ldots, P_e\}$, and let $\gamma(w) = |\overline{Z}(w)|$. If $v$ is erroneous, then there is some $j$ with $v_j \neq 0$. This $v_j$ cannot have more than $\ell_1 := \deg(T) = t + g + \alpha - n + e$ zeros. This implies that the number of points in the set $\{P_1, \ldots, P_e\}$ at which $v_j$ does not vanish is at least $\ell_2 := e - \ell_1 = n - \alpha - g - t$. Furthermore from (2) we see that if $v_j(P_i) \neq 0$ then $w(P_i) \neq 0$, and so $w$ must be also be non-zero on at least $\ell_2$ of the points $P_1, \ldots, P_e$. So if $v$ is erroneous then $\gamma(w) \ge \ell_2$.

If $v = (v_1 \mid \cdots \mid v_m \mid w) \in \ker(A)$ then for all $P_i \in Z(w)$ we have $v_j(P_i) = y_{ij} \cdot w(P_i) = 0$. From this and (4) we obtain

$$\forall j = 1, \ldots, m : \; v_j \in \mathcal{L}\left(T - \sum_{P \in Z(w)} P\right) =: \mathcal{L}(S). \tag{5}$$

Fix $w \in \mathcal{L}\big((t+g)D\big)$, with $\gamma(w) \ge \ell_2$. We will count the expected number of non-zero $(v_1 \mid \ldots \mid v_m)$ for which $v = (v_1 \mid \cdots \mid v_m \mid w) \in \ker(A)$.

If $P_i \in \overline{Z}(w)$ then $y_{ij} = \frac{v_j(P_i)}{w(P_i)}$, and so for a given a non-zero $(v_1, \ldots, v_m) \in \mathcal{L}\big(S\big)^m$, we will have $v \in \ker(A)$ if and only if $y_{ij}$ has the appropriate values for all $j = 1, \ldots, m$ and for all $i$ with $P_i \in \overline{Z}(w)$. Since these $i$ are all error positions, for each one there must be a $j$ with $y_{ij} \neq 0$. So for each $i$, $(y_{i1}, \ldots, y_{im})$ can take $q^m - 1$ different values uniformly (of which exactly one will satisfy $y_{ij} = \frac{v_j(P_i)}{w(P_i)}$ for all $j$). Let $(v_1, \cdots, v_m) \in \mathcal{L}\big(S\big)^m$ be a nonzero vector. Using the fact that $|\overline{Z}(w)| = \gamma(w)$, we obtain

$$\Pr\left[(v_1 \mid \cdots \mid v_m \mid w) \in \ker(A)\right] \le \left(\frac{1}{q^m - 1}\right)^{\gamma(w)}. \tag{6}$$

If $v \in \ker(A)$ then $v_j(P_i) = y_{ij} w(P_i)$, so if $v_j(P_i) = 0$ then $y_{ij} = 0$ for all $i$ with $P_i \in \overline{Z}(w)$. Since $y_{ij}$ cannot be 0 for all $j$, for each $P_i \in \overline{Z}(w)$ there must be some $j$ with $v_j(P_i) \neq 0$. Since $t < n - g - \alpha$ by assumption, we have $\mathcal{L}\big((t + g + \alpha)D - \sum_{i=1}^n P_i\big) = 0$, and since $v_j \in \mathcal{L}(S)$, there exists a subset $U \subseteq \overline{Z}(w)$ of size $\dim(S)$ for which the values of $v_j$ on the points in $U$ uniquely determines $v_j$. So picking $(v_1, \ldots, v_m) \in \mathcal{L}(S)^m$ is the same as picking $v_j(P_i)$ for $j = 1, \ldots, m$ and for $P_i \in U$. Furthermore, as stated above if $v$ is in $\ker(A)$ then for all $P_i$ there must be some $j$ with $v_j(P_i) \neq 0$. So the number of choices for $(v_j, \ldots, v_m) \in \mathcal{L}(S)^m$ for which there exists $j$ with $v_j(P_i) \neq 0$ for all $P_i \in U$ is at most $(q^m - 1)^{\dim(S)}$ and hence for a fixed $w$, the expected number of erroneous vectors $v = (v_1 \mid \cdots \mid v_m \mid w) \in \ker(A)$ is at most

$$\left(\frac{1}{q^m - 1}\right)^{\gamma(w)} \cdot \big(q^m - 1\big)^{\deg(T) - e + \gamma(w) + 1} = q^{\beta m(\deg(T) - e + 1)}, \tag{7}$$

using the fact that $\dim(S) \le \deg(S) + 1 = \deg(T) - e + \gamma(w) + 1$, where $\beta = \frac{\ln(q^m - 1)}{\ln(q^m)}$.

Since $t + g \ge 2g - 1$ by assumption, by the Theorem of Riemann-Roch we have $\dim \mathcal{L}\big((t+g)D\big) = t + 1$, and so there are at most $q^{t+1}$ possible choices for $w$ (there

may be considerably less since we consider only those with $\gamma(w) \geq \ell_2$). The expected number of erroneous vectors in $\ker(A)$ is therefore at most

$$\begin{aligned} q^{t+1} \cdot q^{\beta m(\deg(T)-e+1)} &= q^{t+1+\beta m(t+g+\alpha-n+e-e+1)} \\ &= q^{(\beta m+1)t-\beta m(n-\alpha-g-1)+1}, \end{aligned} \tag{8}$$

and so if $t \leq \frac{\beta m}{\beta m+1}(n-\alpha-g-1) - \frac{c}{\beta m+1}$ then the expected number of erroneous vectors in $\ker(A)$ is at most $q^{1-c}$. If a vector is erroneous, then any non-zero $\mathbb{F}_q$-scalar multiple of that vector is also erroneous. Thus, the probability that the number of erroneous vectors is larger than 0 equals the probability that the number of erroneous vectors is at least $q-1$. By Markov's inequality, this probability is at most the expected number of erroneous vectors divided by $q-1$. This implies that

$$\Pr[\text{exists erroneous vector in } \ker(A)] \leq \frac{q^{1-c}}{q-1} = \frac{q}{q-1} \cdot q^{-c}. \quad \square$$

We conclude the section with the following observation: Setting $t = \frac{\beta m}{\beta m+1}(n-\alpha-g-1) - \frac{2g}{\beta m+1}$ (so $c = 2g$ in the bound above), and observing that the dimension $k$ of the code is at least $\alpha + 1 - g$, we get

$$\begin{aligned} t &\geq \tfrac{\beta m}{\beta m+1}(n-k-2g) - \tfrac{2g}{\beta m+1} \\ &= \tfrac{\beta m}{\beta m+1}(n-k) - 2g \\ &= n\big(\tfrac{\beta m}{\beta m+1}(1-R) - \tfrac{2g}{n}\big). \end{aligned}$$

Since our algorithm can correct up to $t$ errors, the error probability of the $Q$-ary symmetric channel we consider can be at most $\frac{t}{n}$, which is about $1 - R - \frac{2g}{n}$ when $m$ is very large (recall that $\beta = \frac{\ln(q^m-1)}{\ln(q^m)}$). Therefore if $m$ is very large, if $q$ is a square, and if a sequence of very good algebraic curves is used to construct the underlying AG-code, then on a $Q$-ary symmetric channel with error probability $p$ the maximum achievable rate for vanishing error probability of the decoder is roughly

$$1 - p - \frac{2}{\sqrt{q}-1}.$$

(This follows from the fact that for a very good sequence of AG-codes the ratio $g/n$ tends to $1/(\sqrt{q}-1)$.) This shows that these codes and these decoding algorithms can get very close to the capacity of the $Q$-ary symmetric channel.

## 4 Comparison to the Hashing Method

In this final chapter of this paper we give an extensive comparison of our method to other hashing methods. These methods effectively reduce the number of errors, albeit at the expense of reducing the rate of transmission.

The classical method for coding over large alphabets is to dedicate a part of each symbol as check positions. These positions can be used to store a hash value of the symbol. The advantage is that the hash can be used at the receiver side to detect corrupted symbols: If it does not match the symbol, the symbol is corrupted and can be discarded. This way, the decoding problem is effectively reduced to an erasure decoding problem. There is a downside however: each corrupted symbol has a small probability of having a matching hash. The decoder will fail if such a symbol is used, and therefore such decoders have an error term which is linear in the blocklength $n$.

If we use an $[n, k, n-k+1-g]$ AG-code over $\mathbb{F}_Q$, and $\ell$ bits are used in each symbol for the hashing value, then only the remaining $\log(Q) - \ell$ bits can be used per symbol, so the effective rate of the code is

$$r = \frac{\log(Q) - \ell}{\log Q} \cdot \frac{k}{n}.$$

There are two possible failure modes for this decoder. First, if too many symbols have to be discarded, then decoding will fail. A Chernoff-bound argument can be used to show that this happens with exponentially small probability if the symbol error probability bounded away from

$$\frac{n-k-1+g}{n} = \left(1 - \frac{\log Q}{\log(Q) - \ell} r\right) + \frac{g-1}{n}.$$

The second failure mode is when an incorrect symbol passes the hashing test and is therefore used in the decoding process. This happens with probability at most

$$\frac{np}{2^\ell},$$

where $p$ is the symbol error probability. Note that this error probability is linear in $n$, unlike the bounds we get for interleaved codes.

However, it is possible to do better also with the hashing method by adding a second stage to the decoder. After removing symbols with mismatching hash, a few erroneous symbols remain; if there are not too many such symbols, those can be corrected in a second step with the decoder for AG codes. The reasoning is as follows. Let $X_1$ be the number of received erroneous symbols which have mismatching hash values, and let $X_2$ be the number of received erroneous symbols for which the hash matches. Then after removing the mismatching $X_1$ symbols, we are left with an $[n-X_1, k, n-X_1-k+1-g]$ AG-code. Such a code is correctable for errors up to half the minimum distance, hence the correctability condition is

$$n - k + 1 - g > X_1 + 2X_2.$$

If $p$ is the symbol error probability, then we have

$$E[X_1 + 2X_2] = np \cdot (1 - 2^{-\ell}) + 2np \cdot 2^{-\ell}$$

A Chernoff-bound can then be used to show that if the symbol error probability $p$ is bounded away from

$$\frac{(1-R)n + 1 - g}{(1 + 2^{-\ell})n},$$

the resulting failure probability will be exponentially small ($R = k/n$). To summarize, such codes are decodable, if the overall rate (including loss via hashing) is chosen such that

$$r < \frac{\log(Q) - \ell}{\log(Q)} \left(1 - (1 + 2^{-\ell})p - \frac{g}{n} + \frac{1}{n}\right).$$

To compare this to interleaved AG codes, note that the factor

$$\frac{\log(Q) - \ell}{\log(Q)}$$

corresponds to the $(\beta m - 1)/(\beta m)$ term we have for interleaved codes. So, hashing is away by the factor $(1 + 2^{-\ell})$. On the other hand, the advantage of hashing is that $g/n$ can be made much smaller than in the interleaved case, since we are working in a much larger field.

Unfortunately, this fact has another downside in itself: Working on the larger field increases the complexity of the decoder considerably. For interleaved code, it is $O(n^{1+\varepsilon} \log(q)^2)$ where for the hashing method, it is $O(n^{1+\varepsilon} \log(Q)^2)$.

Hashing can also be combined with an interleaved code to produce a much faster decoder which is also extremely simple. The idea is as follows: We dedicate the first of the $m$ interleaved words just for error detection. That is, the first codeword will always be a transmitted zero. On the receiver side, symbols which have a non-zero value in this first interleaved word are again considered erasures. The other interleaved words can then all be decoded separately, using the standard decoder. That way, it is possible to get a decoder which operates on the small field only, and which thus has decoding complexity similar to the interleaved decoder. The error analysis is the same as for the hashing code over the large field; the downside is that we are back to the case where $g/n$ tends to $1/(\sqrt{q} - 1)$. Hence, these codes have slightly worse rates than interleaved AG codes.

## References


1. D. Bleichenbacher, A. Kiyayias, and M. Yung. Decoding of interleaved Reed-Solomon codes over noisy data. In *Proceedings of ICALP 2003*, pages 97–108, 2003.
2. A. Brown, L. Minder, and A. Shokrollahi. Probabilistic decoding of interleaved Reed-Solomon-codes on the $Q$-ary symmetric channel. In *Proceedings of the IEEE International Symposium on Information Theory*, page 327, 2004.
3. D. Coppersmith and M. Sudan. Reconstructing curves in three (and higher dimensional) space from noisy data. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC)*, 2003.
4. A. Garcia and H. Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound. *Invent. Math.*, 121:211–222, 1995.
5. V. Guruswami and M. Sudan. Improved decoding of Reed-Solomon and algebraic-geometric codes. In *Proceedings of the 39th IEEE Symposium on Foundations of Computer Science*, pages 28–37, 1998.
6. J. Justesen, Ch. Thommesen, and T. Høholdt. Decoding of concatenated codes with interleaved outer codes. In *Proc. International Symposium on Information Theory*, page 328, 2004.

7. G.L. Katsman, M.A. Tsfasman, and S.G. Vladut. Modular curves and codes with a polynomial construction. *IEEE Trans. Inform. Theory*, 30:353–355, 1984.
8. M. Luby and M. Mitzenmacher. Verification codes. In *Proceedings of the 40th Annual Allerton Conference on Communication, Control, and Computing*, 2002.
9. M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman. Efficient erasure correcting codes. *IEEE Trans. Inform. Theory*, 47:569–584, 2001.
10. F. Parvaresh and A. Vardy. Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *Proceedings of the 46th Annual IEEE Symposium on the Foundations of Computer Science (FOCS)*, 2005. To appear.
11. A. Shokrollahi and W. Wang. LDPC codes with rates very close to the capacity of the $q$-ary symmetric channel for large $q$. In *Proceedings of the International Symposium on Information Theory, Chicago*, 2004.
12. A. Shokrollahi and H. Wasserman. List decoding of algebraic-geometric codes. *IEEE Trans. Inform. Theory*, 45:432–437, 1999.
13. M. Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *J. Compl.*, 13:180–193, 1997.
14. A. Vardy. Private communication. 2005.