Implementing Gradient Descent Decoding

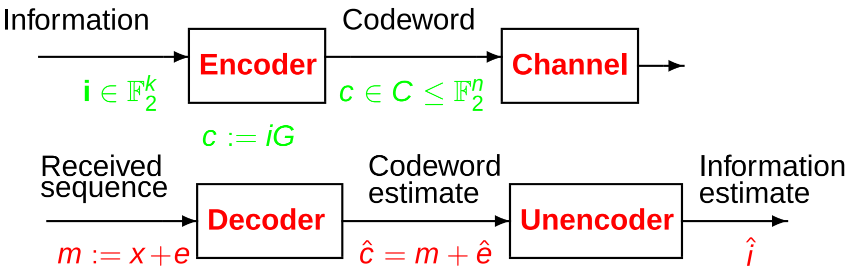
Robert A. Liebler
Colorado State University

A **Provably accurate** efficient optimal decoding algorithm for block codes on channels with BSC input and output is given. But extremely careful choice of a possibly highly redundant parity check matrix is required. A variety of examples are presented.

The fact that such matrices actually exist seems to be new. It demonstrates that trade off between design complexity and implementation cost is possible.

12 October 2007 Perth

Binary error correcting block codes with rate k/n



Information

$\mathbf{i} \in \mathbb{F}_2^k$

Encoder

$c := iG$

Codeword

$c \in C \leq \mathbb{F}_2^n$

Channel

Received sequence

$m := x + e$

Decoder

Codeword estimate

$\hat{c} = m + \hat{e}$

Unencoder

Information estimate

$\hat{i}$

The encoder multiplies i by the generating matrix $G \in Mat_{k,n}(\mathbb{F}_2)$. It has the code C as row space.

The channel adds a random error e.

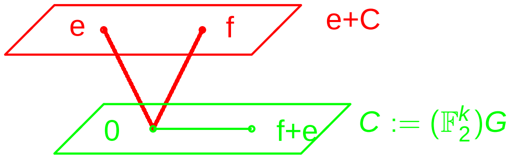
The decoder determines a most likely error $\hat{e}$ and removes it.

Decoding

A parity check matrix: H has maximal rank such that $GH^T = 0$: The syndrome $s := mH^T = (x+e)H^T = eH^T$ depends only on the error e.

The (Hamming) **weight** $wt(e)$ of $e \in \mathbb{F}_2^n$ counts non-zero coordinates and **distance** of $e, f \in \mathbb{F}_2^n$ is $d(e,f) := wt(e+f)$.
Say e, f are **Hamming neighbors** when $d(e,f) = 1$.
The most likely errors, **coset leaders** are those with smallest weight.

If a coset has many coset leaders, there are several equally likely code words given the message.



Optimal vs complete decoding, Gradient functions

The **Voronoi region** $\mathcal{V}(C)$ of a code C is the set of messages $\mathbf{m} \in \mathbb{F}_2^n$ that are coset leaders. Its **interior** is the subset that have $\mathbf{0}$ as the unique closest codeword.

A decoder is **optimal** if it correctly decodes all messages with coset leader in the interior of $\mathcal{V}(C)$. It is **complete** or a **list** decoder if further, it returns a list of nearest codewords other messages.

An iterative decoding algorithm $\mathcal{A}$ is **guided by the function** $\gamma : \mathbb{F}_2^n \to \mathbb{Z}$ if given a received sequence $\mathbf{m}$ for which $\gamma(\mathbf{m}) > \gamma(\mathbf{0})$, $\mathcal{A}$ searches the Hamming neighbors of $\mathbf{m}$ for one that satisfies its **selection criterion** $\mathcal{C}_{\mathcal{A}}$. Then replaces $\mathbf{m}$ with the found neighbor.

A **gradient function** is a function $\gamma : \mathbb{F}_2^n \to \mathbb{Z}$ such that $\gamma(\mathbf{m})$ is an increasing function of the weight of (any) coset leader of $\mathbf{m} + C$.

Quick and Greedy

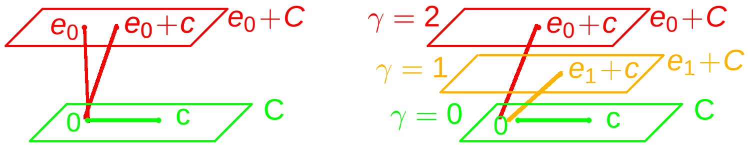
Call the algorithm $\mathcal{A}$ **quick** if its selection criterion $\mathcal{C}_{\mathcal{A}}$ is: the first neighbor for which $\gamma(\mathbf{n}) < \gamma(\mathbf{m})$ and it never "looks back." If no such neighbor exists $\mathcal{A}$ fails.

Call the algorithm $\mathcal{A}$ **greedy** if $\mathcal{C}_{\mathcal{A}}$ requires that $\gamma(\mathbf{n})$ is as small as possible.

NOTES: Greedy $\mathcal{A}$ can fail to converge but only if γ is not gradient.

If γ is gradient and $\mathcal{A}$ quick, $\mathcal{A}$ only fails for messages in the boundary of $\mathcal{V}(C)$, so it is optimal. Also the probability of s-steps in decoding is np^s, where p is the bit error probability, so decoding complexity depends on error weight.

Contrast with classical decoding, The dual code



Use the bilinear form $\mathbf{x} \circ \mathbf{y} := \mathbf{x}\mathbf{y}^T$ on $\mathbb{F}_2^n$ to define the **dual code** $C^\perp$. It is the row space of (any) parity check matrix transpose H for C.

For a vector space W of dimension d, call 1-dimensional (d–1 dimensional) subspaces **points** $\in P_1(W)$ (**hyperplanes** $\in P^1(W)$).

There is a natural bijection β between the **points** of $C^\perp =: P_1(C^\perp)$ and the **hyperplanes** of $\mathbb{F}_2^n/C =: P^1(\mathbb{F}_2^n/C)$ given by

$$\mathbf{p} \overset{\beta}{\leftrightarrow} p^\perp := \{\overline{\mathbf{m}} : p \cdot m = 0\}; \text{ where } \overline{\mathbf{m}} := \mathbf{m} + C.$$

Incidence matrices, $\mathbb{F}_2^n/C$; The weight vector

Given an incidence relation $\mathcal{I} \subset A \times B$, the associated **incidence matrix** N has rows labeled by A and columns labeled by B and $N_{a,b} = 1$ if $(a, b) \in \mathcal{I}$ and 0 else.
Note that $NN^T_{(x,y)}$ counts elements of B incident with both $x, y \in A$.

Let $A = P^1(\mathbb{F}_2^n/C)$, $B = P_1(\mathbb{F}_2^n/C)$ and $\mathcal{I}$ be the **does NOT contain**

$$N^T N = 2^{s-2}(I + J) = NN^T \text{ s:=n-k} \tag{1}$$

where I is the identity matrix on X and J is all ones.

Define the weight vector $\mathbf{wt} \in \mathbb{Z}\mathcal{P}_1(\mathbb{F}_2^n/C)$ to have as $\overline{\mathbf{m}}$-th coordinate the Hamming weight of (any) coset leader of $\overline{\mathbf{m}}$.

A widest parity check matrix; Notation

Set $s = n - k$. Build a (highly redundant) parity check matrix H^T with n rows and 2^s–1 columns indexed by $B = P^1(\mathbb{F}_2^n/C)$ (in the same order as columns of N) having p-th column p^T; $p \in \mathbb{F}_2^s \setminus \{\mathbf{0}\}$.
No parity check matrix can have more columns without repeats.

Consider the "characteristic crossing" function $\blacktriangle : \mathbb{F}_2^s \to \mathbb{Z}^s$ that maps 0 to 0 and 1 to 1. (here s is determined by context).
Use the companion $\blacktriangledown$ for reduction modulo 2 from $\mathbb{Z}$ to $\mathbb{F}_2$.

We use row vectors, so matrices act on the right.
Also sometimes use subscripts to emphasize the characteristic of various objects.
For example, for $M_{\mathbb{Z}} \in Mat_{c,s}(\mathbb{Z})$, write $M_2 = \blacktriangle M_{\mathbb{Z}} \blacktriangledown$ for the same matrix the entries in $\mathbb{F}_2$.

Gradient function construction

Suppose the received sequence $\mathbf{m}$ has coset leader $\mathbf{e}$, so $\overline{\mathbf{m}} = \overline{\mathbf{e}} \in P_1(\mathbb{F}_2^n/C)$. Then $\mathbf{m}H^T\blacktriangle = \mathbf{e}H^T\blacktriangle \in \mathbb{Z}^{2^{n-k}-1}$ has h-th coordinate 1 iff the corresponding hyperplane h **misses** $\overline{\mathbf{e}}$ and 0 else.

Thus $\mathbf{m}H_2^T\blacktriangle$ has coordinates indexed by $P^1(\mathbb{F}_2^n/C)$ AND matches the $\mathbf{e}$-th row of $N_{\mathbf{Z}}^T$. Therefore

$$\mathbf{m}H_2^T\blacktriangle = \chi(\mathbf{e})N_{\mathbb{Z}}^T \qquad (2)$$

where $\chi(\mathbf{e})$ the characteristic function of $\mathbf{e} \in B = P_1(\mathbb{F}_2^n/C)$.

Use equation (2) and the incidence matrix equation (1):

$$\begin{aligned}(mH_2^T\blacktriangle)(N\mathbf{wt}) &= (\chi(\mathbf{e})N^T)(N\mathbf{wt}) = \chi(\mathbf{e})(N^TN)\mathbf{wt} \\ &= \chi(\mathbf{e})(2^{n-k-2}(I+J))\mathbf{wt} = 2^{n-k-2}(wt(\mathbf{e})+K).\end{aligned}$$

where $K = \sum_{\overline{\mathbf{f}} \in P_1(\mathbb{F}_2^n/C)} \mathbf{wt}_{\overline{\mathbf{f}}}$ is a constant.

A Theorem; History

This completes the proof of

THEOREM $\mu(\mathbf{m}) := \mathbf{m}H^T\blacktriangle N\mathbf{wt}$ is a gradient function relative to Hamming weight of coset leaders.

NOTES: The quick algorithm guided by μ is "one step majority logic decoding" (from the 50's) exactly when the code is a Hamming code.

Justensen, Horholdt and Hjaltason (2005) call the function $\mathbf{m} \mapsto \mathbf{m}H^T\blacktriangle\mathbf{j}^T$ **syndrome weight**. They observe that the greedy algorithm guided by syndrome weight is gradient for the [73,45,10] code with parity check matrix the incidence matrix of the projective plane of order 8.

Complete decoding; The challenge

It may be of theoretical interest to point out where the above argument uses Hamming weight of $\overline{\mathbf{e}}$, for coordinates of $\mathbf{wt}$, one could use any numerical tag.

In particular, one could tag each $\overline{\mathbf{e}}$ with a number whose binary expansion is a list of coset leaders with "commas" removed. This number is sufficient to construct a list of most likely errors.

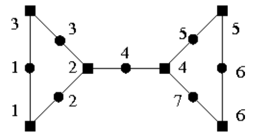
Unfortunately H has rows of length $|C^\perp|$ which is deadly for fast decoding.

The challenge is to find ways to reduce the size of H dramatically without giving up provably optimal decoding guided by μ.

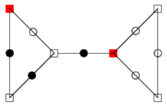
Counting unsatisfied checks with Syndrome Weight

Consider the parity check matrix as an incidence matrix of a bipartite **Tanner graph** $\mathcal{T}$

$$H^T = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

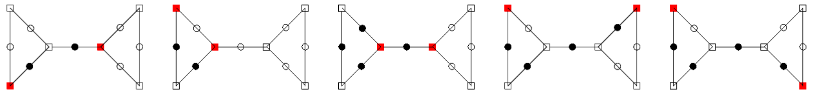


If the received sequence is 11010000



then the check vector is 00110000

Here greedy guided by syndrome weight has a severe dilemma:



Koetter, Li, Vontobel, and Walker's "dumbell"

The symmetry group of $\mathcal{T}$ is dihedral of order 8 (generated by the received sequence bit permutations (16)(2537) and (23)).

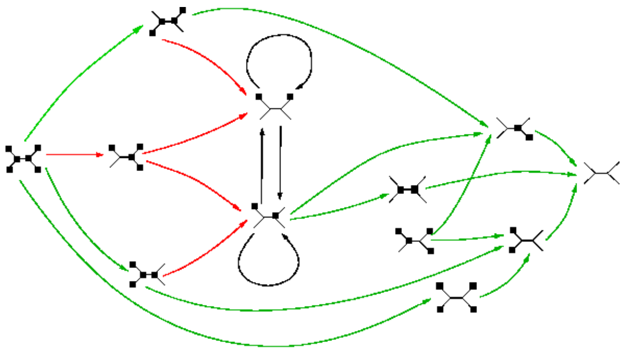
Orbit representatives, orbit size and syndrome weight of the coset leaders in the Voronoi region are:

0000000	1	0	1001000	2	4			
1000000	2	2	1000100	4	4	1001100	4	4
0100000	4	2	1000010	1	4	1001010	1	6
0001000	1	2	0101000	4	2	**0101100**	4	**2**
			0100100	4	**4**			

This syndrome weight is **NOT** gradient. The red entries indicate Hamming neighbors where the one of higher weight has lower syndrome weight.

The Dumbbell Check State Digraph/ Aut($\mathcal{T}$)

The Check State Digraph has as vertices the possible parity check vectors and arcs from determined by the greedy algorithm guided by syndrome weight.

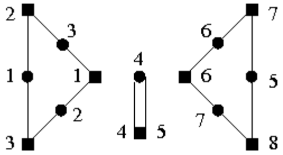


Note looping occurs with black arrows so the algorithm fails to converge.

An "Engineered" Smartbell

Apply the construction in our theorem's proof. The elements of $C^\perp$ have weight multiplicity distribution: $(24^1, 20^6, 16^{24})$. After scaling, an alternative parity check matrix and Tanner multi-graph emerges.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$



This parity check matrix has associated syndrome weight that is a gradient function. The descent algorithm converges in at most 3 steps.

Bonus: This Tanner multi-graph admits the full symmetry group, $S_3 \wr S_2$.

Choosing codes

For what codes is it easy to construct a practical gradient function of coset leader Hamming weight?
Because we do optimal decoding, classical parameters. like minimum distance, are not necessarily indicators of value.

How do you enumerate of the Voronoi region of a code?.
A practical implementation need not be linear in Hamming weight but cannot use more than a small fraction of the possible parity check functions. How to choose possible parity check functions?

Published Heuristic: **concentrate on low weight words in** $C^\perp$

Both of these questions are made easier when the codes considered have large symmetry groups.
Also codes arising from well understood combinatorial structures fall in families and earlier cases give partial answers to these questions.

The Code of the Projective plane of order 4

Consider the incidence map of $\eta : \mathbb{Z}\mathcal{P}_1(\mathbb{F}_4^3) \to \mathbb{Z}\mathcal{P}^1(\mathbb{F}_4^3)$, and let H be its matrix. Take $H_{\blacktriangledown 2}$ to be the parity check matrix of a binary code C.

The simple group $G = L3(4)$ of order 20160 acts as an automorphism group. This well known code C has parameters [21,11,6].

The minimum weight code words are hyperovals.
(The tangents to a non-degenerate conic are concurrent in $PG(2, 2^e)$.
The resulting set of $2^e + 2$ points has the property that any line that is not a secant is **exterior** and contains none of its points.
Any set with this combinatorial property is called a **hyperoval**.)

The Code of the Projective Plane of order 4 (cont)

The Voronoi region can be identified by first computing the number of G-orbits on the elements of $\mathcal{P}_1(\mathbb{F}_2^{21}/C)$. This number is the same as the number of G-orbits on the elements of $\mathcal{P}_1(C^\perp)$.

There are 8 orbits. Data listed in the table: descriptions, orbit size, coset leader weight, number of coset leaders, syndrome weight

null	1	0	1	0	triangle	280	3	4	9
point	21	1	1	5	four, three collinear	280	4	12	12
two points	210	2	1	8	four collinear	21	4	5	16
three collinear	210	3	1	13	five collinear	1	5	21	21

This syndrome weight function is not gradient. Nonetheless greedy guided by syndrome weight decoding is optimal and ends in ≤ 5 steps.

The Lander codes for the Projective Plane of order 4

There is a less well known construction of a second binary code D associated with this plane. To understand this [21,9,8] code, use the characterization:

$$\mathbf{x} \in C \Leftrightarrow \left\{ \begin{matrix} \text{(the support of) } \mathbf{x} \text{ intersects every} \\ \text{line in an even number of points} \end{matrix} \right\};$$

that is $C = (Im(\eta^T)_{\blacktriangledown 2})^\perp$.

Let $\mathcal{L}$ be the set of 6 exterior lines to a hyperoval $\mathcal{O}$. Then $\mathcal{L}^{\eta^T}$ is the multi-set $2\mathcal{E}$ where $\mathcal{E}$ is the formal sum of the 15 points not in $\mathcal{O}$. Therefore (the characteristic function of)

$$\frac{1}{2}\mathcal{L}^{\eta^T} = \mathcal{E} \in \Lambda := \frac{1}{2}\left[Im(\eta^T) \cap 2\mathbb{Z}\mathcal{P}_1(\mathbb{F}_4^3)\right].$$

Lander codes defined

Moreover (the support of) any element of $D := (\Lambda_{\blacktriangledown 2})^\perp$ must meet every line AND every hyperoval complement in an even number of points. There are only two G-orbits on $D^\perp \cap C$ and every non-trivial coset leader for a D coset in C is a hyperoval.

For $e \leq n$, I call codes like these of the form

$$\frac{1}{2^e}\left[Im(\eta^T) \cap 2^e\mathbb{Z}\mathcal{P}_1(\mathbb{F}_{2^n}, m)\right]$$

Lander Codes because E. Lander introduced them and worked out their parameters in the late 1970's (unpublished masters thesis at Oxford 1978?).

Lander Codes arising from higher dimensional finite projective geometries.

Let H be incidence map of $\mathbb{Z}\mathcal{P}_1(\mathbb{F}_4^4) \rightarrow \mathbb{Z}\mathcal{P}^1(\mathbb{F}_4^4)$. The parameters of the code C are [85,68,6], [85,60,8] and they admit the group $L4(4)$.

There are 10 $L4(4)$-orbits on $\mathcal{P}_1(\mathbb{F}_2^{85}/C)$. These are (listed with descriptions, coset leader weight, number of coset leaders and syndrome weight):

null	0	1	0	triangle	3	4	37
point	1	1	21	four collinear	4	21	64
two points	2	1	40	four, 3 collinear	4	12	48
three collinear	3	1	53	four, indep.	4	27	40
				five collinear	5	100	85
				five, gen. pos.	5	135	35

Syndrome weight isn't gradient but greedy syndrome weight descent decoding is optimal.

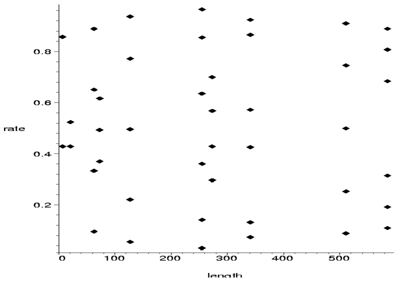
Future work?

The second Lander code D has parameters [85,60,8] and (again) only two L4(4)-orbits on $\mathcal{P}_1(C/D)$. The nontrivial orbit (still) has hyperovals (in a plane) as coset leaders.
One can decode the second code by computing the syndrome weight for the first code and only when it takes this value, $\mu_C = 9$ use a second set of parity checks (possibly based on an L3(4)-orbit of hyperovals) and the associated syndrome weight function to determine which bit to flip.

null	1	0	1	0	triangle	280	3	4	9
point	21	1	1	5	four, three collinear	280	4	12	12
two points	210	2	1	8	four collinear	21	4	5	16
three collinear	210	3	1	13	five collinear	1	5	21	21

Future work?

The rate and lengths of the first few binary Lander codes:



The challenge of Shannon's channel coding theorem.

Earlier work of which I have become aware

Lucas, Bossert and Breitbach (1998) study a "soft output channel" but use only parity check matrices generated from the minimum weight dual code words.
They define a "gradient function" γ that reduces to something equivalent to "syndrome weight" for a BSC but use the Gallager update rule where **All** b are revised at every step using γ. The algorithm iterates until stability or time runs out. Their function does not have property we have taken for the definition of a gradient function.

Kim, Lin and Fossorier (2001) seem to be the first to suggest an algorithm that updates the syndrome after each bit flip.

If you are aware of other historical connections, PLEASE LET ME KNOW.