Extractors from Reed-Muller codes

Sven Laur
swen@math.ut.ee

Helsinki University of Technology

Brief outline

- Quick recap of coding theory
  - Reed-Muller codes
  - State of the art binary codes
- Alternative view to randomness
  - Block-sources with low conditional min-entropy
  - Block-sources with small failure probability
- Close look to the engine under the hood
  - TZS-construction
  - Double counting proof technique
- Bells and whistles—wrapper for binary inputs

Essentials of Reed-Muller codes

Consider all multivariate polynomials $f : \mathbb{F}_q^d \to \mathbb{F}_q$ with $\deg f \leq h$.

$$\mathcal{M} = \{f \in \mathbb{F}_q[x_1, \ldots, x_d] : \deg f \leq h\}$$

Now, fix a clever set $\mathcal{S} \subseteq \mathbb{F}_q^d$. To code $f \in \mathcal{M}$ evaluate f on $\mathcal{S}$, i.e.

$$\mathcal{M} \ni f \longmapsto (f(s_1), \ldots, f(s_k)) \in \mathbb{F}_q^k.$$

$\textsc{Trivia}$: If $\mathcal{S} = \mathbb{F}_q^d$ then

- code length is q^d;
- code dimension $\binom{h+d}{d}$.

State of the art binary codes

We need really sparse codes that are efficiently constructible.

Combinatorial list decoding property

A code has combinatorial list decoding property α

- if every Hamming ball of relative radius $\frac{1}{2} - \alpha$ has $\mathcal{O}(1/\alpha^2)$ codewords.

There are polynomial-time constructible $[n, k]$ codes with combinatorial list decoding property α, where $n = \mathcal{O}(k/\alpha^4)$.

Uniform distribution has no memory!

Consider a partition of random random source into blocks $Z = Z_1 \circ \cdots \circ Z_b$.

How much information we gain if we know $Z_1 \circ \cdots \circ Z_{i-1}$?

$$k_i(z_1 \circ \cdots \circ z_{i-1}) = H_\infty(Z_i | z_1 \circ \cdots \circ z_{i-1})$$

Block source

Random source Z is a $(n_1, k_1), \ldots, (n_b, k_b)$ block source iff

$$k_i = \max_{z_1 \circ \cdots \circ z_{i-1}} H_\infty(Z_i | z_1 \circ \cdots \circ z_{i-1}) \qquad i = 1, \ldots, b.$$

- In case of uniform distribution and $k_i = n_i$.
- Hence the values k_i characterise how far is the distribution from uniform.

Block sources with small failures

Block source with tolerated failure β

Random source Z is a $(n_1, k_1), \ldots, (n_b, k_b)$ β-almost block source iff

$$\Pr_{z_1 \circ \cdots \circ z_{i-1}} [H_\infty(Z_i | z_1 \circ \cdots \circ z_{i-1}) < k_i] \leq \beta \qquad i = 1, \ldots, b.$$

Lemma 1. *A β-almost $(n_1, k_1), \ldots, (n_\beta, k_b)$ block source is $b\beta$ close to $(n_1, k_1), \ldots, (n_\beta, k_b)$ block source.*

Proof. Standard hybrid argument technique:

- Substitute failures with uniform distribution.
- Do some simple calculations to verify result.

Block sources with large min-entropies

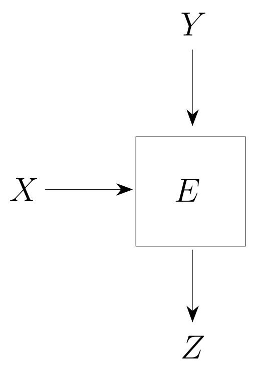
Lemma 2. *A* $(1, \kappa(\alpha)), \ldots, (1, \kappa(\alpha))$ *block source with* $2^{-\kappa(\alpha)} = \frac{1}{2} + \alpha$ *is* $b\alpha$ *close to the uniform distribution.*

Proof. Standard hybrid argument technique:

- Substitute blocks one by one with uniform distribution.
- Do some simple calculations to verify result.

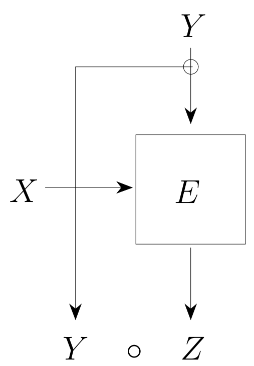
Corollary 1. *A* β*-almost* $(1, \kappa(\alpha)), \ldots, (1, \kappa(\alpha))$ *block source with* $2^{-\kappa(\alpha)} = \frac{1}{2} + \alpha$ *is* $b(\alpha + \beta)$ *close to uniform distribution.*

Extractor specification



Standard extractor

- Random source X has a min-entropy $H_\infty(X) \geq k$.
- Random source Y is uniform.
- Output source Z is close to uniform.



Strong extractor

- Random source X has a min-entropy $H_\infty(X) \geq k$.
- Random source Y is uniform.
- Output source $Y \circ Z$ is close to uniform.

TZS extractor. Type specification

We construct a strong extractor where

- X ranges over all two-variate Reed-Muller polynomials

$$\mathcal{M} = \{f \in \mathbb{F}_q[x_1, x_2] : \deg f < h\} ;$$

- Y ranges uniformly over triples

$$(a_1, a_2, j) \in \mathbb{F}_q \times \mathbb{F}_q \times [1, \ell];$$

- output Z is m bit string.

TZS extractor. Auxiliary structures

Given two error parameters:

- α min-entropy bound to a bit, i.e. $2^{-\kappa(\alpha)} = \frac{1}{2} + \alpha$;
- failure bound β on block source.

We can set

- q to the first prime $q \geq \Omega\left(\frac{h}{\alpha^4\beta^4}\right)$;
- $\mathbf{C}$ to linear binary code of dimension $d = \lceil \log q \rceil$ with combinatorial list decoding property $\frac{\alpha\beta}{4}$;
- ℓ to the code-length of $\mathbf{C}$.

TZS extractor. Construction

Given: $f \in \mathcal{M}$ and $(a_1, a_2, j) \in \mathbb{F}_q \times \mathbb{F}_q \times [1, \ell]$

Output: a bit string $z = z_1, \ldots z_m$ such that

$$z_i = \mathbf{C}\left(f(a_1 + i, a_2)\right)_j \qquad i = 1, \ldots, m.$$

Visualisation

	1	2	$\cdots$	m
1 2 3 $\vdots$ ℓ	$\mathbf{C}\left(f(a_1 + 1, a_2)\right)$	$\mathbf{C}\left(f(a_1 + 2, a_2)\right)$	$\vdots$	$\mathbf{C}\left(f(a_1 + m, a_2)\right)$

General framework

Function E is (k, ϵ) strong extractor iff for any distribution X over $\mathcal{M}$

$$H_\infty(X) \geq k \qquad \Rightarrow \qquad \mathsf{SDiff}(U_Y \circ E(X, U_Y) \| U_Z) \leq \epsilon.$$

<u>**Necessary and sufficient test**</u>

Choose $X \subseteq \mathcal{M}$ such that $|X| \geq 2^k$ and use uniform distribution over X. Show that

$$\mathsf{SDiff}(U_Y \circ E(U_X, U_Y) \| U_Z) \leq \epsilon$$

or alternatively

$$\mathsf{SDiff}(U_Y \circ E(U_X, U_Y) \| U_Z) > \epsilon \qquad \Rightarrow \qquad |X| < 2^k.$$

Double counting argument

Assume that $U_Y \circ E(U_X, U_Y)$ is not a block source with:

- relatively small failure probability β;
- min-entropy bound $\kappa(\alpha)$ such that $2^{-\kappa(\alpha)} = \frac{1}{2} + \alpha$.

Deduce a contradiction

- Derive a short description for some polynomials $f \in X^* \subseteq X$.
- Show that polynomials with short description form a large fraction of X.
- Compute the information-theoretical upper bound to $|X^*|$.
- Expose the contradiction in sizes.

Setting the stage

To get a contradiction assume that the output $U_Y \circ E(U_X, U_Y)$ is not β-almost $(\star,\star),(1,\kappa(\alpha)),\ldots,(1,\kappa(\alpha))$ block-source.

In other words exists i_0 such that

$$\Pr_{\substack{z_1\circ\cdots\circ z_{i_0-1}\\ a_1,a_2\in\mathbb{F}_q\\ j\in[1,\ell]}}\left[H_\infty(Z_{i_0}|a_1\circ a_2\circ j\circ z_1\circ\cdots\circ z_{i_0-1})<\kappa(\alpha)\right]\geq\beta.$$

Or even more explicitly

$$\Pr\left[a_1,a_2,j:\Pr\left[Z_{i_0}|a_1,a_2,j,z_1\circ\cdots\circ z_{i_0-1}\right]\geq\frac{1}{2}+\alpha\right]\geq\beta.$$

How to reconstruct f from output?

$a_1 - i_0$	a_2	j	$z_1 \ldots, z_{i_0-1}$	?	$\cdots$

$$\begin{array}{|c|}\hline f_1 \\ f_2 \\ \vdots \\ f_r \\ \hline\end{array} \quad \begin{array}{c} \mapsto \\ \mapsto \\ \vdots \\ \mapsto \end{array} \quad \begin{array}{|c|c|}\hline z_1 \ldots, z_{i_0-1} & z_{i_0}^1 \\ z_1 \ldots, z_{i_0-1} & z_{i_0}^2 \\ \vdots & \vdots \\ z_1 \ldots, z_{i_0-1} & z_{i_0}^r \\ \hline\end{array}$$

- Consider set of all polynomials that are consistent with prefix $z_1 \ldots z_{i_0-1}$

$$X_{j,z_1 \ldots z_{i_0-1}} = \left\{ f \in \mathcal{M} \ : \ \mathbf{C}\left(f(a_1 - i_0 + i, a_2)\right)_j = z_i, \ i = 1, \ldots, i_0 - 1 \right\}$$

- Predict the next bit $z_{i_0} = \mathbf{C}\left(f(a_1, a_2)\right)_j$ by majority voting.

How to predict an $f(a_1, a_2)$?

<u>Aim</u>: Minimise the number of evaluations of $f \in X$.

- Compute values $f(a_1 - i_0 + 1, a_2), \ldots, f(a_1 - 1, a_2)$ and corresponding output bits $z_1(j), \ldots, z_{i_0-1}(j)$ of the output $\boxed{a_1 - i_0 \mid a_2 \mid j \mid \ldots}$
- Find $X_{j, z_0(j) \ldots z_{i_0-1}(j)}$ use majority voting to guess the next bit $z_{i_0}(j)$.
- Form a codeword $z = z_{i_0}(1) \ldots z_{i_0}(\ell)$ and use list decoding to find all $g \in \mathcal{M} \cap X$ such that

$$\mathbf{EP} = \left\{ g(a_1, a_2) \ : \ H(\mathbf{C}\left(g(a_1, a_2)\right), z) \leq \left(\frac{1}{2} - \frac{\alpha\beta}{4} \right) \ell \right\}$$

- Output $\mathbf{EP}$.

What is the probability of a correct guess?

Consider: $a_1, a_2 \in \mathbb{F}_q$, $j \in [1, \ell]$ and a prefix $z_1(j) \ldots z_{i_0}(j)$

(a) Block source fails $H_\infty(Z_{i_0} | a_1 \circ a_2 \circ j \circ z_1(j) \ldots z_{i_0 - 1}(j)) < \kappa(\alpha)$

$$\Pr\left[z_{i_0} \text{ coincides with majority} | \text{failure}\right] = ?$$

(b) Block source works $H_\infty(Z_{i_0} | a_1 \circ a_2 \circ j \circ z_1(j) \ldots z_{i_0 - 1}(j)) > \kappa(\alpha)$

$$\Pr\left[z_{i_0} \text{ coincides with majority} | \neg\text{failure}\right] = ?$$

(c) As the failure probability is less than β

$$\Pr\left[z_{i_0} \text{ coincides with majority}\right] \geq ?$$

Working with averages

As $\Pr\limits_{a_1,a_2,j}[z_{i_0}\text{coincides with majority}] \geq \frac{1}{2} + \alpha\beta$

(a) Show that there exists $X' \subset X$ such that

- $|X'| \geq \frac{\alpha\beta}{2}|X|$;
- $\Pr\limits_{a_1,a_2,j}[\ z_{i_0} \text{ coincides with majority}|f \in X'] \geq \frac{1}{2} + \frac{\alpha\beta}{2}$

(b) Show that $H(\mathbf{C}\left(f(a_1,a_2)\right), z) < \frac{\alpha\beta}{4}$ is not rare

$$\Pr_{a_1,a_2}\left[\Pr_j [z_{i_0} \text{ coincides with majority}|f \in X'] \geq \frac{1}{2} + \frac{\alpha\beta}{4}\right] > \frac{\alpha\beta}{4}$$

Short summary

Exists a mythical set X' such that

- $|X'| \geq \frac{\alpha\beta}{2} |X|$;
- $\Pr [f \in \mathbf{EP} | f \in X'] \geq \frac{\alpha\beta}{4}$.

Cleverly chosen list-decoding property $\frac{\alpha\beta}{4}$ assures that $|\mathbf{EP}| = \mathcal{O}\left(\frac{1}{\alpha^2\beta^2}\right)$.

The computation takes ages, but we do not care.

A short description of f restricted to a line

Description of f_L:

- List tuples of values $f(a_1 - i_0 + 1, a_2), \ldots, f(a_1 - 1, a_2)$ for h line points $L(1), \ldots L(h)$. The number of values is less than $(m-1)h$.
- Give index i of correct polynomial f_L in the set of all consistent candidates G. We prove that $|G| = \mathcal{O}\left(\frac{1}{\alpha^3\beta^3}\right)$.

Decoding procedure

- Use polynomial interpolation to restore values at all points

$$f(a_1 - i_0 + 1, a_2 + j), \ldots, f(a_1 - 1, a_2 + j) \qquad j \in \mathbb{F}_q$$

- Compute predictor sets $S_j = \mathbf{EP}\big(L(j)\big),\ j \in \mathbb{F}_q$.
- Compute the list G of all univariate polynomials that are consistent at least with $\frac{\alpha\beta}{8}q$ sets.
- Output the ith polynomial or $\perp$ if $i = 0$.

When does decoding fail if $f \in X'$?

Given a random line L over the $\mathbb{F}_q \times \mathbb{F}_q$ the failure probability is $\mathcal{O}\big(\frac{1}{\alpha\beta q}\big)$.

Proof.

- Let $Y_i = [f(L(i)) \in \mathbf{EP}(L(i))]$ and $Y = Y_1 + \cdots + Y_q$.
- Now $E(Y) \geq \frac{\alpha\beta q}{4}$ and $D(Y) = \mathcal{O}(\alpha\beta q)$.
- Chebyshev inequality gives

$$\Pr\left[Y \leq \frac{\alpha\beta q}{8}\right] \leq \mathcal{O}\left(\frac{1}{\alpha\beta q}\right).$$

How large is the set of all candidates G?

- We know that $|S_j| = \mathcal{O}\big(\frac{1}{\alpha^2\beta^2}\big)$ and $q = \Omega\big(\frac{h}{\alpha^4\beta^4}\big)$.
- A nice interpolation lemma by Sudan assures that $|G| = \mathcal{O}\big(\frac{1}{\alpha^3\beta^3}\big)$.
  - The result follows from clever trade-off between q and h.

A short description of f

Description of f:

- Choose a line L. Compute corresponding description.
- Advance line one step forward, i.e. $L = L + (1, 0)$. Compute description. Store only the index i of the correct candidate polynomial.
- Repeat second step $h - 1$ times.

Recovery procedure

Decoding:

- Restore the first line f_L or output $\perp$ on failure.
- Set $L = L+(1,0)$ and restore f_L using precomputed values of $f(a_1-i, a_2)$ or halt with $\perp$ on failure.
- Repeat second step $h-1$ times.
- Interpolate f over horizontal lines.

When does decoding fail if $f \in X'$?

Given a random line L over the $\mathbb{F}_q \times \mathbb{F}_q$ the failure probability is $\mathcal{O}\left(\frac{h}{\alpha\beta q}\right)$. The latter can be made less than $\frac{1}{2}$ by tuning the parameters q, h and α.

Proof. Union bound.

There exist a line L and set X^* such that decoding is always successful and $|X^*| \geq \frac{1}{2}\,|X'|$

Proof. Summation reordering technique.

The Promised Contradiction

Information-theoretic bound
The number of description states is bounded by

$$q^{(m-1)h} \times \mathcal{O}\left(\frac{1}{\alpha^3\beta^3}\right)^h = q^{(m-1)h} o(q)^h = \frac{\alpha\beta}{4} o(q^{mh}).$$

Thus $|X| \leq \frac{4}{\alpha\beta} |X^*| = o(q^{mh})$.

Thus if we assume that $|X| \geq q^{mh}$, we get the promised β-almost block source.

Adjusting the first input

- Let $m \leq \sqrt{n}$ then we can find an embedding from $\{0,1\}^n \rightarrow \mathcal{M}$ and we are done.
- After some calculations one can deduce $t = 2\log q + \log \ell$ is actually less than $\log n + \mathcal{O}\big(\log \frac{1}{\alpha\beta}\big)$.
- Some clever tweaking with parameters gives the following theorem.

Theorem 1. *For every $m = m(n)$, $k = k(n)$ and $\epsilon = \epsilon(n) \leq 1/2$ such that $3m\sqrt{n}\log(n/\epsilon) \leq k \leq n$ there exist an explicit family of (k, ϵ) strong extractors $E_n : \{0,1\}^n \times \{0,1\}^t \rightarrow \{0,1\}^m$ with $t = \log n + \mathcal{O}(\log m) + \mathcal{O}(\log \frac{1}{\epsilon})$*

Post-processing. Multivariate polynomials

By Post-processing one can slightly improve the result.

Code	Min-entropy k	Additional randomness t	m
Two-variate No preprocessing	$3m\sqrt{n}\log(n/\epsilon)$	$\log n + \mathcal{O}(\log m)$ $+\mathcal{O}(\log \frac{1}{\epsilon})$	m
Two-variate with preprocessing	$m\sqrt{n}\log^2 n$	$\log n + \mathcal{O}(\log^* m)$	m
Multi-variate with preprocessing	$n^{1/c}m$	$\log n + \mathcal{O}(c^2 \log m)$	$\Omega(k)$
Multi-variate with preprocessing	$\Omega(n)$	$\log n + \log\log n$	$\Omega(k)$

State of the art

Enhancement of TZS extractor by Shaltiel and Umans gives better results.

Min-entropy k	Pure randomness t	Output size m
$\log^{\mathcal{O}(1/\delta)} n$	$\mathcal{O}(\log n)$	$k^{1-\delta}$
$\log^{\mathcal{O}(1/\delta)} n$	$(1+\delta)\log n$	$k^{\Omega(\delta)}$
any	$(1+\alpha)\log n$	$k/(\log^{\mathcal{O}(1/\alpha)} n)$