

Reed-Solomon Codes Against Adversarial Insertions and Deletions

RONI CON

JOINT WORK WITH AMIR SHPILKA AND ZACHI TAMO



Insertions, Deletions, and Edit Distance

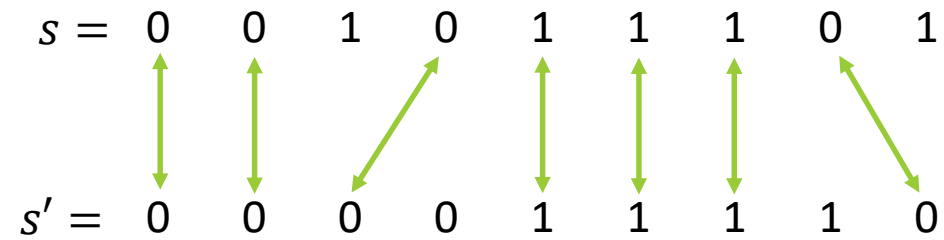
- ❖ Deletion: $10010 \rightarrow 100$
- ❖ Insertion: $0110 \rightarrow 01010$
- ❖ $ED(s, s')$: The minimal number of insertions and deletions for $s \rightarrow s'$.

Why Insertions and Deletions?

- ❖ Model synchronization errors.
- ❖ Appear naturally in DNA and RNA.
- ❖ DNA storage.
- ❖ Trace reconstruction problem.

Longest Common Subsequence

An LCS between s and s' is a maximal length subsequence of both. The length is denoted $|LCS(s, s')|$.



LCS and ED

❖ It holds that

$$ED(s, s') = |s| + |s'| - 2|LCS(s, s')|$$

❖ Small LCS \leftrightarrow large edit distance.

❖ $C \subseteq \Sigma^n$.

$$ED(C) = \min_{\substack{c, c' \in C \\ c \neq c'}} ED(c, c') \quad \text{and} \quad LCS(C) = \max_{\substack{c, c' \in C \\ c \neq c'}} LCS(c, c')$$

$$LCS(C) < n - \delta n \leftrightarrow ED(C) > 2\delta n.$$

❖ Rate:

$$R = \frac{\log(|C|)}{n \cdot \log(|\Sigma|)}$$

Background on Insdel Codes

- ❖ [VT65, Lev66]: Optimal binary codes correcting 1 deletion/insertion.
- ❖ [HS17]: Efficient codes of rate $1 - \delta - \varepsilon$ with alphabet size $O_\varepsilon(1)$ correcting δ -fraction of insdel.
- ❖ [Hae19], [CJLW19]: Efficient binary codes of rate $1 - O(\delta \log^2 \frac{1}{\delta})$ correcting δ -fraction of insdel.
- ❖ [GS19, SRB20, GH21]: Best binary codes correcting 2 deletion/insertion.

Not linear codes!

Linear Codes

❖ A linear code of *length* n and *dimension* k :

$C \subseteq \mathbb{F}_q^n$ linear subspace of dimension k .

❖ Notation: $[n, k]_q$ code

❖ Rate: $R = \frac{k}{n}$

We Love Linear Codes

- ❖ Compact representations.

- ❖ Generating matrix, G : rows are basis of the linear codes

$$C = \{xG \mid x \in \mathbb{F}_q^k\}.$$


- ❖ Parity check matrix, H :

$$C = \{x \in \mathbb{F}_q^n \mid Hx = 0\}.$$

- ❖ Efficiently encodable.

- ❖ Familiar, easier to analyze, and sometimes even efficiently decodable.

Previous Results - Insdel Linear Codes

- ❖ [CGHL21]: There are linear codes with rate $\frac{1-\delta}{2} - \frac{h(\delta)}{\log_2 q}$.
- ❖ [CGHL21]: Every linear code has rate $\leq \frac{1-\delta}{2} + o(1)$.  **Half-Singleton bound.**

	q	δ	R
[CGHL21]	2	$< 1/400$	$\approx 2^{-80}$
[CST22]	$poly(\varepsilon^{-1})$	$< 1/4$	$(1 - 4\delta)/8 - \varepsilon$
[CST22]	2	$< 1/54$	$(1 - 54\delta)/1216$

- ❖ What about RS codes against insdel errors?


Reed-Solomon Codes

Let $\alpha_1, \dots, \alpha_n \in \mathbb{F}_q$ be distinct. The $[n, k]_q$ RS code defined with $\alpha_1, \dots, \alpha_n$ is

$$C := \{(f(\alpha_1), \dots, f(\alpha_n)) \mid f \in \mathbb{F}_q[X], \deg(f) < k\}$$

❖ Linear code.

❖ Rate: $\frac{k}{n}$

❖ In the Hamming metric: $\delta = 1 - R + \frac{1}{n}$  **Matches the Singleton bound.**

❖ $q \geq n$.

Vandermonde Matrix

❖ Generating matrix of RS codes:

$$V = \begin{pmatrix} 1 & \alpha_1 & \dots & \alpha_1^{k-1} \\ 1 & \alpha_2 & \dots & \alpha_2^{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \alpha_n & \dots & \alpha_n^{k-1} \end{pmatrix}$$

$$m = (a_0, \dots, a_{k-1})$$

$$f = \sum_{i=0}^{k-1} a_i x^i$$

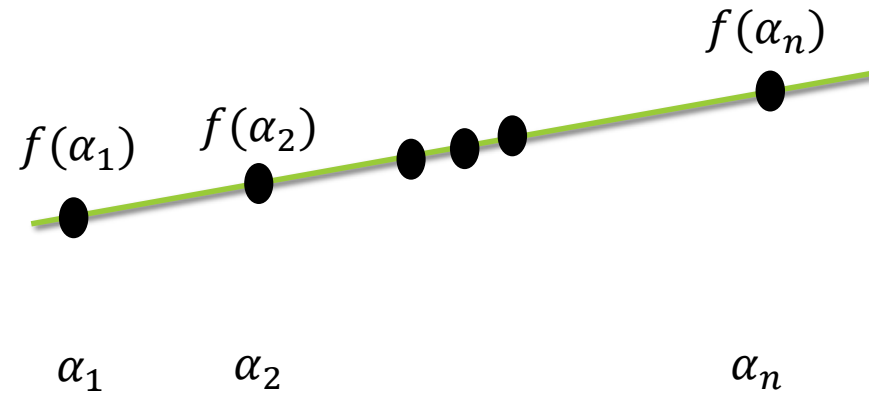
$$V \cdot m = c$$

$$\longrightarrow c = (f(\alpha_1), \dots, f(\alpha_n))$$

We Love RS codes

- ❖ Linear and have efficient decoding!
- ❖ Data storage
 - ❖ CDs, DVDs, etc..
 - ❖ Facebook, Google, etc..
- ❖ QR codes
- ❖ Space transmission.
- ❖ Crypto: Shamir's secret sharing, MPC, etc..

The $k = 2$ case.



Classical:

$$(\alpha_i, f(\alpha_i)), (\alpha_j, f(\alpha_j))$$

\downarrow
 f

This work:

$$f(\alpha_i), f(\alpha_j), f(\alpha_s) \quad i < j < s$$

\downarrow
 f

Previous & Our Results $k = 2$.

	q	Correcting deletions
[WMSN04]	$q > 9$ ($n = 5$)	1
[DLTX19]	$\exp(n)$	$n - 3$
[CST21]	$O(n^4)$	$n - 3$
[LX21]	$O(n^5)$	$n - 3$

[CST21] Lower Bound: $q = \Omega(n^3)$.

Previous & Our Results $k > 2$.

	q	Correcting deletion
[TSN07]	$O(n)$	$\log_{k+1} n - 1$
Existence [CST21]	$n^{O(k)}$	$n - 2k + 1$
Construction [CST21]	$\approx n^{k^k}$	$n - 2k + 1$

Match the half-Singleton Bound*

* Half-Singleton Bound [CGHL21]: Any linear code that corrects δ fraction of deletions must have rate $\leq \frac{1-\delta}{2} + o(1)$

Algebraic Condition

Definition (Increasing vectors):

$I \subseteq [n]^s$ is an increasing vector if $1 \leq I_1 < I_2 < \dots < I_s \leq n$.

Notation

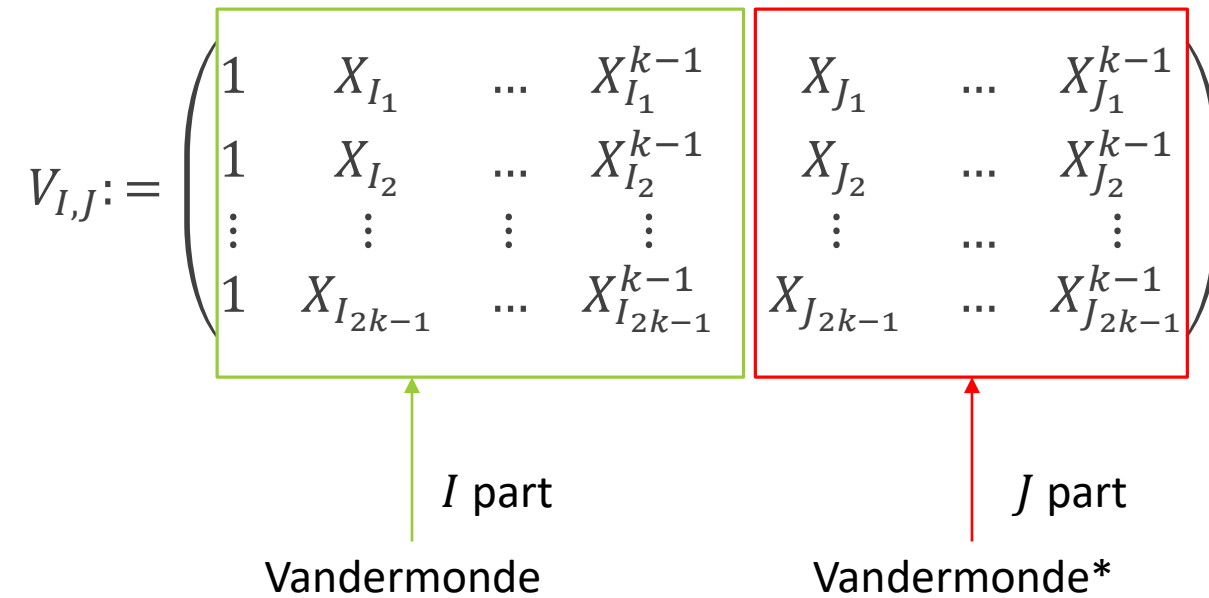
$$c_I := (c_{I_1}, c_{I_2}, \dots, c_{I_s})$$

❖ c_I is a subsequence of c of length s .

Algebraic Condition

Definition: Let $I, J \subseteq [n]^{2k-1}$ be increasing vectors and X_1, \dots, X_n be formal variables.

$$V_{I,J} := \begin{pmatrix} 1 & X_{I_1} & \dots & X_{I_1}^{k-1} & X_{J_1} & \dots & X_{J_1}^{k-1} \\ 1 & X_{I_2} & \dots & X_{I_2}^{k-1} & X_{J_2} & \dots & X_{J_2}^{k-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{I_{2k-1}} & \dots & X_{I_{2k-1}}^{k-1} & X_{J_{2k-1}} & \dots & X_{J_{2k-1}}^{k-1} \end{pmatrix}$$



I part J part
Vandermonde Vandermonde*

$$\det(V_{I,J}) \in \mathbb{F}_q[X_1, \dots, X_n].$$

Algebraic Condition

$$V_{I,J}(\alpha_1, \dots, \alpha_n) = \begin{pmatrix} 1 & \alpha_{I_1} & \dots & \alpha_{I_1}^{k-1} & \alpha_{J_1} & \dots & \alpha_{J_1}^{k-1} \\ 1 & \alpha_{I_2} & \dots & \alpha_{I_2}^{k-1} & \alpha_{J_2} & \dots & \alpha_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \alpha_{I_{2k-1}} & \dots & \alpha_{I_{2k-1}}^{k-1} & \alpha_{J_{2k-1}} & \dots & \alpha_{J_{2k-1}}^{k-1} \end{pmatrix}$$

Claim:

Let $\alpha = (\alpha_1, \dots, \alpha_n)$ and consider the $[n, k]_q$ RS code defined with α .

If for **every two increasing vectors** $I, J \subseteq [n]^{2k-1}$ that agree on at most $k - 1$ coordinates, it holds that $\det(V_{I,J}(\alpha)) \neq 0$, then the code **can correct any $n - 2k + 1$ deletions**.

Proof of Claim

❖ Assume that the claim does not hold.

❖ There are distinct

$$c = (f(\alpha_1), \dots, f(\alpha_n)), \quad f = \sum_{i=0}^{k-1} f_i x^i$$

$$c' = (g(\alpha_1), \dots, g(\alpha_n)), \quad g = \sum_{i=0}^{k-1} g_i x^i$$

such that $|LCS(c, c')| \geq 2k - 1$.

❖ There are two increasing $I, J \subseteq [n]^{2k-1}$ such that $c_I = c'_J$.

Claim:

$$V_{I,J}(\alpha) = \begin{pmatrix} 1 & \alpha_{I_1} & \dots & \alpha_{I_1}^{k-1} & \alpha_{J_1} & \dots & \alpha_{J_1}^{k-1} \\ 1 & \alpha_{I_2} & \dots & \alpha_{I_2}^{k-1} & \alpha_{J_2} & \dots & \alpha_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \alpha_{I_{2k-1}} & \dots & \alpha_{I_{2k-1}}^{k-1} & \alpha_{J_{2k-1}} & \dots & \alpha_{J_{2k-1}}^{k-1} \end{pmatrix}$$

For every two increasing vectors $I, J \subseteq [n]^{2k-1}$ that agree on at most $k - 1$ coordinates, it holds that $\det(V_{I,J}(\alpha)) \neq 0$, then the code can correct any $n - 2k + 1$ deletions.

Proof of Claim

- ❖ $f(\alpha_{I_s}) = g(\alpha_{J_s})$ for every $1 \leq s \leq 2k - 1$.
- ❖ If I, J agree on $\geq k$ coordinates then $f \equiv g$.
- ❖ Otherwise, the vector

$$(f_0 - g_0, f_1, \dots, f_{k-1}, -g_1, \dots, -g_{k-1})$$

is nonzero and in the kernel of $V_{I,J}(\alpha)$.

Claim:

$$V_{I,J}(\alpha) = \begin{pmatrix} 1 & \alpha_{I_1} & \dots & \alpha_{I_1}^{k-1} & \alpha_{J_1} & \dots & \alpha_{J_1}^{k-1} \\ 1 & \alpha_{I_2} & \dots & \alpha_{I_2}^{k-1} & \alpha_{J_2} & \dots & \alpha_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \alpha_{I_{2k-1}} & \dots & \alpha_{I_{2k-1}}^{k-1} & \alpha_{J_{2k-1}} & \dots & \alpha_{J_{2k-1}}^{k-1} \end{pmatrix}$$

For every two increasing vectors $I, J \subseteq [n]^{2k-1}$ that agree on at most $k - 1$ coordinates, it holds that $\det(V_{I,J}(\alpha)) \neq 0$, then the code can correct any $n - 2k + 1$ deletions.

Existence – Road Map (Informal)

$$V_{I,J} = \begin{pmatrix} 1 & X_{I_1} & \dots & X_{I_1}^{k-1} & X_{J_1} & \dots & X_{J_1}^{k-1} \\ 1 & X_{I_2} & \dots & X_{I_2}^{k-1} & X_{J_2} & \dots & X_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{I_{2k-1}} & \dots & X_{I_{2k-1}}^{k-1} & X_{J_{2k-1}} & \dots & X_{J_{2k-1}}^{k-1} \end{pmatrix}$$

- ❖ Show: $\det(V_{I,J})$ is not the zero polynomial for any two increasing I, J that agree on $\leq k - 1$.
- ❖ There are at most $\binom{n}{2k-1}^2$ pairs I, J .



$$P(X_1, \dots, X_n) := \prod_{I,J} \det(V_{I,J}) \neq 0$$

- ❖ $\deg P(X_1, \dots, X_n) \approx n^{4k-2} \cdot k^2$.
- ❖ Schwarz-Zippel lemma.

The $k = 2$ case.

Theorem

There exists an explicit $[n, 2]_q$ RS code that can correct from $n - 3$ insdel errors where $q = O(n^4)$.

Sidon Spaces

Definition (Sidon Space): $S \subseteq \mathbb{F}_q^n$ such that

- ❖ S is linear subspace over \mathbb{F}_q .
- ❖ $\forall a, b, c, d \in S$. If $ab = cd$, then $\{a\mathbb{F}_q, b\mathbb{F}_q\} = \{c\mathbb{F}_q, d\mathbb{F}_q\}$.

Theorem [RRT17]:

Let $q \geq 3$ prime power, m an integer. There is an explicit m dimensional Sidon space $S \subseteq \mathbb{F}_{q^{2m}}$ (over \mathbb{F}_q).

Long Code

Theorem [GS86]:

For every integer $m \geq 1$, there exists an explicit $\left[\frac{(3^m+1)}{2}, \frac{(3^m+1)}{2} - 2m \right]_3$ linear code with hamming distance ≥ 5 .

Construction

- ❖ m : positive integer.
- ❖ $S \subseteq \mathbb{F}_{3^{4m}}$: Sidon space of dimension $2m$. s_1, \dots, s_{2m} is basis of S .
- ❖ $H \in \mathbb{F}_3^{2m \times \binom{3^m+1}{2}}$: the parity check matrix of the long code.
- ❖ Define the points of the $\left[\left(\frac{3^m+1}{2}, 2\right)_{3^{4m}}\right]$ RS code:

$$(\alpha_1, \dots, \alpha_n) = (s_1, \dots, s_{2m}) \cdot H$$

- ❖ **Lemma**: Any 4 points of our code are linearly independent over \mathbb{F}_3 .

Proof

- ❖ Assume that the code does not correct $n - 3$ deletions.
- ❖ There are $(x_1, x_2, x_3), (y_1, y_2, y_3)$ that agree on at most 1 coordinate such that

$$\det \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} = 0$$

- ❖ Equivalently

$$(y_1 - y_2)(x_2 - x_3) = (y_2 - y_3)(x_1 - x_2)$$

- ❖ $\exists \lambda \in \mathbb{F}_q$ such that

$$\lambda(y_1 - y_2) = y_2 - y_3 \quad \text{or} \quad \lambda(y_1 - y_2) = x_1 - x_2$$

- ❖ Contradicts the lemma.

Proof of the lemma

Lemma: Any 4 points of our code are linearly independent over \mathbb{F}_q .

- ❖ Assume $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are dependent.
- ❖ There are $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{F}_q$ such that $\sum_{i=1}^4 \beta_i \alpha_i = 0$.

$$0 = \sum_{i=1}^4 \beta_i \alpha_i = \sum_{i=1}^4 \beta_i \sum_{j=1}^{2m} s_j H_{j,i} = \sum_{j=1}^{2m} s_j \sum_{i=1}^4 \beta_i H_{j,i}$$

- ❖ The s_j 's are independent over \mathbb{F}_q .
- ❖ For all j , $\sum_{i=1}^4 \beta_i H_{j,i} = 0$. Contradiction! (the distance of the code is ≥ 5).

$k > 2$ case

Mason-Stothers Thm:

Let $a(t)$, $b(t)$, and $c(t)$ be relatively prime polynomials such that $a + b = c$ and not all of them have vanishing derivative. Then,

$$\max\{\deg(a), \deg(b), \deg(c)\} \leq \deg(\text{rad}(abc)) - 1$$

($\text{rad}(f)$) is the product of the distinct irreducible factors of f)

Thm [VW03]:

Let $m \geq 2$ and $Y_0(x) = Y_1(x) + \cdots + Y_m(x)$ where $Y_j(x) \in \mathbb{F}_p[x]$. Assume that

- ❖ $\gcd(Y_0, Y_1, \dots, Y_m) = 1$
- ❖ Y_1, \dots, Y_m are linearly independent over $\mathbb{F}_p(x^p)$

Then,

$$\deg(Y_0) \leq -\binom{m}{2} + (m-1) \sum_{j=0}^m \nu(Y_j)$$

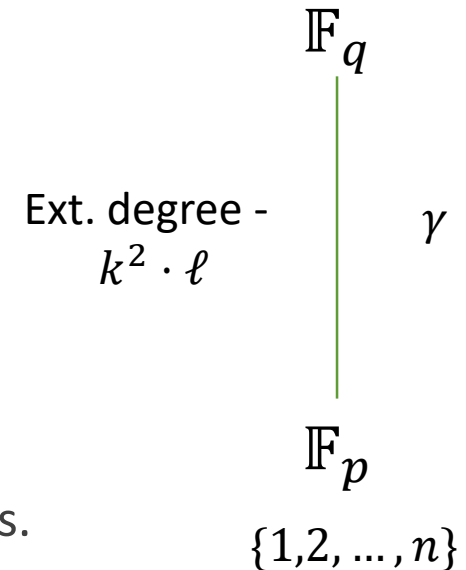
($\nu(Y_j)$ is the number of distinct roots of Y_j with multiplicity not divisible by p)

Construction

- ❖ $\ell = ((2k)!)^2$.
- ❖ \mathbb{F}_p where $p > k^2 \cdot \ell$.
- ❖ \mathbb{F}_q such that $[\mathbb{F}_q : \mathbb{F}_p] = k^2 \cdot \ell$.
- ❖ $\gamma \in \mathbb{F}_q$ such that $\mathbb{F}_q = \mathbb{F}_p(\gamma)$.
- ❖ Construction points: $\forall i \in [n]$ (where $n < p$):

$$\alpha_i = (\gamma - i)^\ell$$

$$\{\alpha_1 = (\gamma - 1)^\ell, \dots, \alpha_n = (\gamma - n)^\ell\}$$



Claim:

The $[n, k]_q$ RS code with $\alpha_1, \dots, \alpha_n$ can correct from $n - 2k + 1$ deletions.

Proof idea (informal)

$$V_{I,J}(\alpha) = \begin{pmatrix} 1 & \alpha_{I_1} & \cdots & \alpha_{I_1}^{k-1} & \alpha_{J_1} & \cdots & \alpha_{J_1}^{k-1} \\ 1 & \alpha_{I_2} & \cdots & \alpha_{I_2}^{k-1} & \alpha_{J_2} & \cdots & \alpha_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \alpha_{I_{2k-1}} & \cdots & \alpha_{I_{2k-1}}^{k-1} & \alpha_{J_{2k-1}} & \cdots & \alpha_{J_{2k-1}}^{k-1} \end{pmatrix}$$

If $\det(V_{I,J}(\alpha)) = 0$, then

$$\det(V_{I,J}(\alpha)) = P_0(\gamma) + P_1(\gamma) + \cdots + P_{(2k-1)!-1}(\gamma) = 0.$$

Proof idea (very informal)

$$\det(V_{I,J}(\alpha)) = P_0(\gamma) + P_1(\gamma) + \cdots + P_{(2k-1)!-1}(\gamma) = 0.$$

❖ $\deg(P_j) = \ell \cdot k \cdot (k - 1)$

❖ $v(P_j) \leq 2k - 2.$

$$\begin{pmatrix} 1 & \alpha_{I_1} & \cdots & \alpha_{I_1}^{k-1} & \alpha_{J_1} & \cdots & \alpha_{J_1}^{k-1} \\ 1 & \alpha_{I_2} & \cdots & \alpha_{I_2}^{k-1} & \alpha_{J_2} & \cdots & \alpha_{J_2}^{k-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \alpha_{I_{2k-1}} & \cdots & \alpha_{I_{2k-1}}^{k-1} & \alpha_{J_{2k-1}} & \cdots & \alpha_{J_{2k-1}}^{k-1} \end{pmatrix}$$

$$\alpha_i = (\gamma - i)^\ell$$

❖ By thm, it holds that

$$\deg(P_0) < (2k - 1)! \sum_{j=0}^{(2k-1)!-1} v(P_j) \leq ((2k - 1)!)^2 \cdot (2k - 2) < \ell$$

Conclusion & Open Questions

- ❖ Explicit construction with smaller field size.
- ❖ A tighter lower bound on the field size for $k > 2$.
- ❖ Decoding algorithms?

Thank You!
