The Geometry of Codes for Random Access in DNA Storage

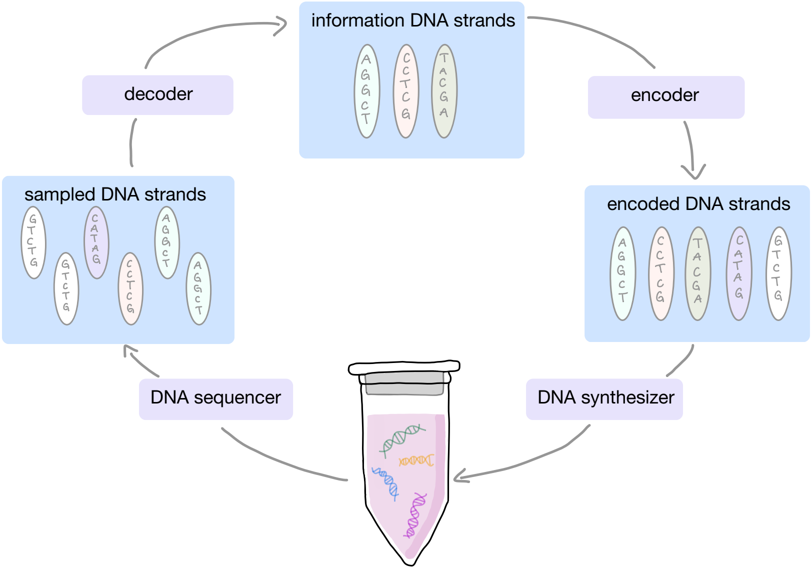
Anina Gruica, Technical University of Denmark

The Seventh Irsee Conference, Finite Geometries 2025

September 3

joint work with Maria Montanucci and Ferdinando Zullo

DNA storage systems



The random access problem

- $1 \leq k \leq n$ integers, q a prime power,
- $\mathcal{G} = \{P_1, \ldots, P_n\} \subseteq \mathsf{PG}(k-1,q)$ with $\langle P_1, \ldots, P_n \rangle = \mathsf{PG}(k-1,q)$,
- E_i denotes the point corresponding to the i-th basis vector (fundamental point),
- points in $\mathcal{G}$ are drawn uniformly at random,
- $\forall i \in \{1, \ldots, k\}$, $\tau_{E_i}(\mathcal{G})$ – random variable counting the number of points of $\mathcal{G}$ that are drawn until E_i is in their $\mathbb{F}_q$-span,
- More generally: $\tau_P(\mathcal{G})$ – random variable counting the number of points of $\mathcal{G}$ that are drawn until P is in their $\mathbb{F}_q$-span.

The random access problem

- $1 \le k \le n$ integers, q a prime power,
- $\mathcal{G} = \{P_1, \ldots, P_n\} \subseteq \mathrm{PG}(k-1, q)$ with $\langle P_1, \ldots, P_n \rangle = \mathrm{PG}(k-1, q)$,
- E_i denotes the point corresponding to the i-th basis vector (fundamental point),
- points in $\mathcal{G}$ are drawn uniformly at random,
- $\forall i \in \{1, \ldots, k\}$, $\tau_{E_i}(\mathcal{G})$ – random variable counting the number of points of $\mathcal{G}$ that are drawn until E_i is in their $\mathbb{F}_q$-span,
- More generally: $\tau_P(\mathcal{G})$ – random variable counting the number of points of $\mathcal{G}$ that are drawn until P is in their $\mathbb{F}_q$-span.

The Random Access Problem

- For any $i \in \{1, \ldots, k\}$ compute the expectation $\mathbb{E}[\tau_{E_i}(\mathcal{G})]$.
- Find the maximal expected number of samples to retrieve an information strand

$$T_{\max}(\mathcal{G}) \triangleq \max_{i \in \{1, \ldots, k\}} \mathbb{E}[\tau_{E_i}(\mathcal{G})].$$

The random access problem - example

Example (points in Fano plane): Let

$$\begin{aligned}\mathcal{G} &= \{(1:0:0),(0:1:0),(0:0:1),(1:1:0),(0:1:1),(1:0:1),(1:1:1)\}\\ &= \{E_1,E_2,E_3,P_4,P_5,P_6,P_7\} = \mathsf{PG}(2,2).\end{aligned}$$

A possible (first part) of a sequence of reads is

$$\omega = (P_4,\ E_2,\ E_2,\ P_5,\ P_7,\ E_1,\ldots).$$

Then $\tau_{E_2}(\mathcal{G})(\omega) = 2$, $\tau_{E_1}(\mathcal{G})(\omega) = 2$, $\tau_{E_3}(\mathcal{G})(\omega) = 4$.

What is known – results from [BLSGY23]

- identity code achieves $T_{\max}(\mathcal{G}) = k$,
- simple parity code achieves $T_{\max}(\mathcal{G}) = k$,
- non-systematic $[n, k]$ MDS codes achieves $T_{\max}(\mathcal{G}) \approx n \log\left(\frac{n}{n-k}\right) > k$,
- systematic $[n, k]$ MDS codes achieves $T_{\max}(\mathcal{G}) = k$,
- construction of $[2k, k]$ codes for which $T_{\max}(\mathcal{G}) \approx 0.95k$,
- construction of 2-dim. code (with rate $= 0$) for which $T_{\max}(\mathcal{G}) \approx 0.91 \cdot 2$, and of 3-dim. code (with rate $= 0$) for which $T_{\max}(\mathcal{G}) \approx 0.89 \cdot 3$.

[BLSGY23] D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, **"Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems"**, IEEE Transactions on Information Theory (2024).

General formula for expectation

Proposition [G., Bar-Lev, Ravagnani, & Yaakobi, 2024]: Let $\mathcal{G} = \{P_1, \ldots, P_n\}$ and $H_i := 1 + 1/2 + \cdots + 1/i$ (the i-th harmonic number). We have

$$\mathbb{E}[\tau_P(\mathcal{G})] = nH_n - \sum_{s=1}^{n-1} \frac{|\{S \subseteq \{1, \ldots, n\} : |S| = s,\ P \in \langle P_i : i \in S \rangle\}|}{\binom{n-1}{s}}.$$

General formula for expectation

Proposition [G., Bar-Lev, Ravagnani, & Yaakobi, 2024]: Let $\mathcal{G} = \{P_1, \dots, P_n\}$ and $H_i := 1 + 1/2 + \cdots + 1/i$ (the i-th harmonic number). We have

$$\mathbb{E}[\tau_P(\mathcal{G})] = nH_n - \sum_{s=1}^{n-1} \frac{|\{S \subseteq \{1, \dots, n\} : |S| = s,\ P \in \langle P_i : i \in S \rangle\}|}{\binom{n-1}{s}}.$$

Example: Assume $\mathcal{G} = \{P_1, \dots, P_n\}$ is an n-arc. Then

$$|\{S \subseteq \{1, \dots, n\} : |S| = s,\ P_i \in \langle P_j : j \in S \rangle\}| = \begin{cases} \binom{n-1}{s-1} & \text{if } s \in [k-1], \\ \binom{n}{s} & \text{if } s \geq k. \end{cases}$$

From above proposition we get

$$\mathbb{E}[\tau_{P_i}(\mathcal{G})] = nH_n - \sum_{s=1}^{k-1} \frac{\binom{n-1}{s-1}}{\binom{n-1}{s}} - \sum_{s=k}^{n-1} \frac{\binom{n}{s}}{\binom{n-1}{s}} = nH_n - \sum_{s=1}^{k-1} \frac{s}{n-s} - \sum_{s=k}^{n-1} \frac{n}{n-s} = k.$$

Theorem [G., Bar-Lev, Ravagnani, & Yaakobi, 2024]: Let $\mathcal{G} = \{P_1, \dots, P_n\}$. We have

$$\sum_{i=1}^{n} \mathbb{E}\left[\tau_{P_i}(\mathcal{G})\right] = kn.$$

Theorem [G., Bar-Lev, Ravagnani, & Yaakobi, 2024]: Let $\mathcal{G} = \{P_1, \dots, P_n\}$. We have

$$\sum_{i=1}^{n} \mathbb{E}\left[\tau_{P_i}(\mathcal{G})\right] = kn.$$

Proof sketch: Let $\mathcal{G} = \{P_1, \dots, P_7\} = \mathsf{PG}(2,2)$,
T_i – random variable counting number of draws until we sample a new point, having already recovered $i-1$ of them. Since $T_i \sim \mathsf{Geom}(\frac{7-i+1}{7})$, $\mathbb{E}[T_i] = \frac{7}{7-i+1}$.
There always exists some ordering $\{i_1, \dots, i_7\} = \{1, \dots, 7\}$ s.t.

$$\begin{aligned}
\tau_{P_{i_1}}(\mathcal{G}) &= T_1 = 1, \\
\tau_{P_{i_2}}(\mathcal{G}) &= T_1 + T_2, \\
\tau_{P_{i_3}}(\mathcal{G}) &= T_1 + T_2, \\
\tau_{P_{i_4}}(\mathcal{G}) &= \tau_{P_{i_5}}(\mathcal{G}) = \tau_{P_{i_6}}(\mathcal{G}) = \tau_{P_{i_7}}(\mathcal{G}) = T_1 + T_2 + T_4.
\end{aligned}$$

Theorem [G., Bar-Lev, Ravagnani, & Yaakobi, 2024]: Let $\mathcal{G} = \{P_1, \dots, P_n\}$. We have

$$\sum_{i=1}^{n} \mathbb{E}\left[\tau_{P_i}(\mathcal{G})\right] = kn.$$

Proof sketch: Let $\mathcal{G} = \{P_1, \dots, P_7\} = \mathrm{PG}(2,2)$,
T_i – random variable counting number of draws until we sample a new point, having already recovered $i-1$ of them. Since $T_i \sim \mathrm{Geom}(\frac{7-i+1}{7})$, $\mathbb{E}[T_i] = \frac{7}{7-i+1}$.
There always exists some ordering $\{i_1, \dots, i_7\} = \{1, \dots, 7\}$ s.t.

$$\begin{aligned}
\tau_{P_{i_1}}(\mathcal{G}) &= T_1 = 1, \\
\tau_{P_{i_2}}(\mathcal{G}) &= T_1 + T_2, \\
\tau_{P_{i_3}}(\mathcal{G}) &= T_1 + T_2, \\
\tau_{P_{i_4}}(\mathcal{G}) &= \tau_{P_{i_5}}(\mathcal{G}) = \tau_{P_{i_6}}(\mathcal{G}) = \tau_{P_{i_7}}(\mathcal{G}) = T_1 + T_2 + T_4.
\end{aligned}$$

$\implies \sum_{i=1}^{7} \mathbb{E}\left[\tau_{P_i}(\mathcal{G})\right] = 7 \cdot \mathbb{E}[T_1] + 6 \cdot \mathbb{E}[T_2] + 4 \cdot \mathbb{E}[T_4] = 7 \cdot \frac{7}{7} + 6 \cdot \frac{7}{6} + 4 \cdot \frac{7}{4} = 3 \cdot 7 = 21.$

Recovery balanced codes

$\mathcal{G}$ is **recovery balanced** if

$$\mathbb{E}[\tau_{P_1}(\mathcal{G})] = \cdots = \mathbb{E}[\tau_{P_n}(\mathcal{G})] = \boldsymbol{k}.$$

Recovery balanced codes

$\mathcal{G}$ is **recovery balanced** if

$$\mathbb{E}[\tau_{P_1}(\mathcal{G})] = \cdots = \mathbb{E}[\tau_{P_n}(\mathcal{G})] = \boldsymbol{k}.$$

Some examples:
MDS codes, simplex code, Hamming code, Reed-Muller code, binary Golay code.

$\Longrightarrow$ for these codes the random access expectation is k.

Recovery balanced codes

$\mathcal{G}$ is **recovery balanced** if

$$\mathbb{E}[\tau_{P_1}(\mathcal{G})] = \cdots = \mathbb{E}[\tau_{P_n}(\mathcal{G})] = \boldsymbol{k}.$$

Some examples:
MDS codes, simplex code, Hamming code, Reed-Muller code, binary Golay code.

$\Longrightarrow$ for these codes the random access expectation is k.

Point sets that are recovery balanced are not “good”!!

Small values of k

$k = 2$

- In [BLSGY23] authors give construction of $\mathcal{G}$ with $T_{\max}(\mathcal{G}) \approx 0.914 \cdot k$.
- In [BEGGTY25] we show that their construction is optimal, i.e., one can not obtain lower random access expectation for $k = 2$.

[BLSGY23] D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, **"Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems"**, IEEE Transactions on Information Theory (2024).
[BEGGTY25] A. Boruchovsky, O. Elishco, R. Gabrys, A. G., I. Tamo, and E. Yaakobi, **"Making it to First: The Random Access Problem in DNA Storage"**, arXiv preprint arXiv:2501.12274.

Small values of k

$k = 2$

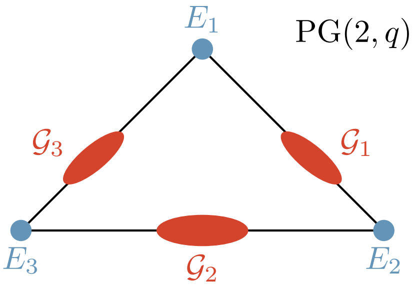
- In [BLSGY23] authors give construction of $\mathcal{G}$ with $T_{\max}(\mathcal{G}) \approx 0.914 \cdot k$.
- In [BEGGTY25] we show that their construction is optimal, i.e., one can not obtain lower random access expectation for $k = 2$.

[BLSGY23] D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, **"Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems"**, IEEE Transactions on Information Theory (2024).
[BEGGTY25] A. Boruchovsky, O. Elishco, R. Gabrys, A. G., I. Tamo, and E. Yaakobi, **"Making it to First: The Random Access Problem in DNA Storage"**, arXiv preprint arXiv:2501.12274.

$k = 3$

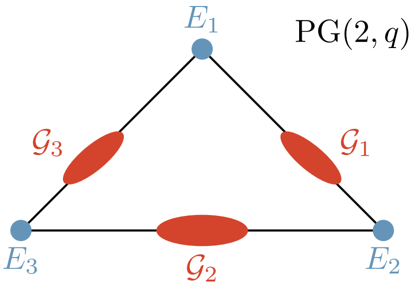
- In [BLSGY23] authors give construction of $\mathcal{G}$ with $T_{\max}(\mathcal{G}) \approx 0.89 \cdot k$.
- In [GMZ24] we give construction of $\mathcal{G}$ with $T_{\max}(\mathcal{G}) \approx 0.88 \cdot k$.

[GMZ24] A. G., M. Montanucci and F. Zullo **"The Geometry of Codes for Random Access in DNA Storage"**, arXiv preprint arXiv:2411.08924.



Balanced quasi-arc of weight x:

- $\mathcal{G} = \{E_1, E_2, E_3\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$,
- $|\mathcal{G}_1| = |\mathcal{G}_2| = |\mathcal{G}_3| = x$,
- $\mathcal{G}_i \subseteq E_i E_{i+1}$ for any i,
- $|\ell \cap \mathcal{G}| \leq 2$, for any line $\ell \neq E_1E_2, E_2E_3, E_1E_3$.

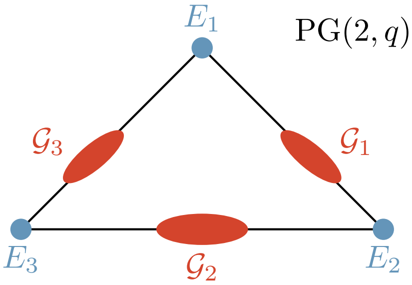


Balanced quasi-arc of weight x:

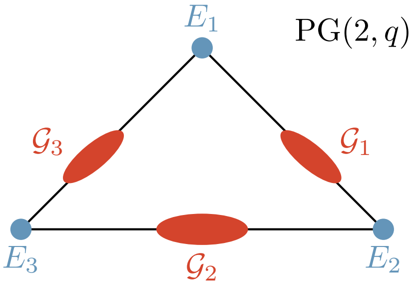
- $\mathcal{G} = \{E_1, E_2, E_3\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$,
- $|\mathcal{G}_1| = |\mathcal{G}_2| = |\mathcal{G}_3| = x$,
- $\mathcal{G}_i \subseteq E_iE_{i+1}$ for any i,
- $|\ell \cap \mathcal{G}| \leq 2,$ for any line $\ell \neq E_1E_2, E_2E_3, E_1E_3$.

Properties:

- $|\mathcal{G}| = 3x + 3$
- $|\mathcal{G} \cap E_iE_{i+1}| = x + 2$
- $x \leq \frac{q-1}{2}$



- $H \subseteq \mathbb{F}_q^*$ a subgroup,
- $\tilde{H} \subseteq \mathbb{F}_q^* \setminus H$,
- $|H| = |\tilde{H}| = x$,
- $\mathcal{G}_1 = (1, -\tilde{h}, 0)$, $\tilde{h} \in \tilde{H}$,
  $\mathcal{G}_2 = (0, 1, -h)$, $h \in H$,
  $\mathcal{G}_3 = (-h, 0, 1)$, $h \in H$,

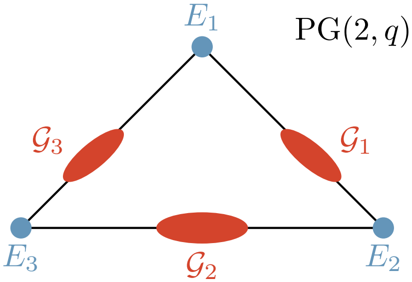


- $H \subseteq \mathbb{F}_q^*$ a subgroup,
- $\tilde{H} \subseteq \mathbb{F}_q^* \setminus H$,
- $|H| = |\tilde{H}| = x$,
- $\mathcal{G}_1 = (1, -\tilde{h}, 0)$, $\tilde{h} \in \tilde{H}$,
  $\mathcal{G}_2 = (0, 1, -h)$, $h \in H$,
  $\mathcal{G}_3 = (-h, 0, 1)$, $h \in H$,

More explicitly:

- q **odd:** $(H, \tilde{H}) = (\blacksquare_q, \mathbb{F}_q^* \setminus \blacksquare_q)$
- q **even:** $(H, \tilde{H}) = (\mathbb{F}^*_{q/2}, \subset \mathbb{F}_q^* \setminus \mathbb{F}^*_{q/2})$

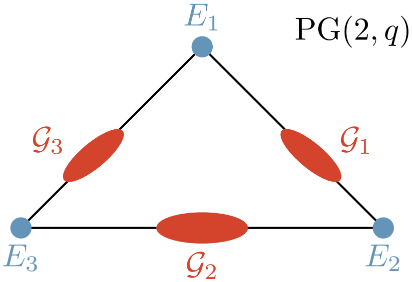
$\Longrightarrow x = \frac{q-1}{2}$ in both cases.



$\mathcal{G}_x := \{E_1, E_2, E_3\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3.$

$\alpha(\mathcal{G}_x, s) := |\{S \subseteq \mathcal{G}_x : |S| = s,\ E_i \in \langle P : P \in S\rangle\}|$

$$\alpha(\mathcal{G}_x, s) = \begin{cases} 1, & s = 1, \\ 2\binom{x+2}{2} + x, & s = 2, \\ \binom{3x+3}{s} - \binom{x+2}{s} & 3 \le s \le x+2, \\ \binom{3x+3}{s} & x+2 < s \le 3x+2. \end{cases}$$



$$\mathcal{G}_x := \{E_1, E_2, E_3\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3.$$

$$\alpha(\mathcal{G}_x, s) := |\{S \subseteq \mathcal{G}_x : |S| = s,\ E_i \in \langle P : P \in S \rangle\}|$$

$$\alpha(\mathcal{G}_x, s) = \begin{cases} 1, & s = 1, \\ 2\binom{x+2}{2} + x, & s = 2, \\ \binom{3x+3}{s} - \binom{x+2}{s} & 3 \leq s \leq x+2, \\ \binom{3x+3}{s} & x+2 < s \leq 3x+2. \end{cases}$$

$$\implies \mathbb{E}[\tau_{E_i}(\mathcal{G}_x)] = 3 + \frac{2}{3x+1} - \frac{2((x+2)(x+1)+x)}{(3x+2)(3x+1)} + \sum_{s=3}^{x+2} \prod_{i=0}^{s-1} \frac{x+2-i}{3x+2-i}.$$

$$\implies \lim_{x\to\infty} \mathbb{E}[\tau_{E_i}(\mathcal{G}_x)] \leq 3 - 1/6 \approx 0.9\overline{44}k.$$

Improvement of previous construction

By adding multiplicities to fundamental points we can improve previous construction.

Let $\mathcal{G}_{x,y} := \{{E_1}^y, {E_2}^y, {E_3}^y\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$. Then

Improvement of previous construction

By adding multiplicities to fundamental points we can improve previous construction.

Let $\mathcal{G}_{x,y} := \{{E_1}^y, {E_2}^y, {E_3}^y\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$. Then

$$\mathbb{E}[\tau_{E_i}(\mathcal{G}_{x,y})] = 3 + \frac{2}{3x+3y-2} - \frac{y-1}{3x+3y-1} - \frac{2\left(xy + \binom{x}{2}\right) + y(3x+2y) + y(y-1)/2}{\binom{3x+3y-1}{2}}$$
$$+ \sum_{s=3}^{x+2y} \prod_{i=0}^{s-1} \frac{x+2y-i}{3x+3y-i-1} + \sum_{s=3}^{y+1} \frac{2\binom{y}{s-1}x}{\binom{3x+3y-1}{s}}$$

$$\Longrightarrow \lim_{x\to\infty} \mathbb{E}[\tau_{E_i}(\mathcal{G}_{x,0.834x})] \leq 0.881\overline{66} \cdot k.$$

Improvement of previous construction

By adding multiplicities to fundamental points we can improve previous construction.

Let $\mathcal{G}_{x,y} := \{{E_1}^y, {E_2}^y, {E_3}^y\} \cup \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$. Then

$$\mathbb{E}[\tau_{E_i}(\mathcal{G}_{x,y})] = 3 + \frac{2}{3x+3y-2} - \frac{y-1}{3x+3y-1} - \frac{2\left(xy + \binom{x}{2}\right) + y(3x+2y) + y(y-1)/2}{\binom{3x+3y-1}{2}}$$
$$+ \sum_{s=3}^{x+2y} \prod_{i=0}^{s-1} \frac{x+2y-i}{3x+3y-i-1} + \sum_{s=3}^{y+1} \frac{2\binom{y}{s-1}x}{\binom{3x+3y-1}{s}}$$

$$\implies \lim_{x\to\infty} \mathbb{E}[\tau_{E_i}(\mathcal{G}_{x,0.834x})] \leq 0.881\overline{66} \cdot k.$$

Thank you for your attention!