Covering Arrays via Finite Fields

Charles J. Colbourn

Perth, December 2023

ARIZONA STATE UNIVERSITY

10th ACCM	University of Adelaide	23–27 August 1982
11th ACCM	University of Canterbury	29 Aug – 2 Sep 1983
12th ACCMC	University of Western Australia	13–17 August 1984
13th ACCMC	University of Sydney	26–30 August 1985

Happy (Belated) Birthday

Gordon Royle

Happy (Belated) Birthday

Gordon Royle

Covering Array
Definition

- Let N, k, t, v, and λ be positive integers.
- Let C be an $N \times k$ array with entries from an alphabet Σ of size v; we typically take $\Sigma = \{0, \ldots, v-1\}$.
- When $(\nu_1, \ldots, \nu_t)$ is a t-tuple with $\nu_i \in \Sigma$ for $1 \leq i \leq t$, $(c_1, \ldots, c_t)$ is a tuple of t column indices ($c_i \in \{1, \ldots, k\}$), and $c_i \neq c_j$ whenever $\nu_i \neq \nu_j$, the t-tuple $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ is a *t-way interaction*.
- C *λ-covers* the t-way interaction $\{(c_i, \nu_i) : 1 \leq i \leq t\}$ if, in at least λ rows $\rho_1, \ldots, \rho_\lambda$ of C, the entry in row ρ_r and column c_i is ν_i for $1 \leq r \leq \lambda$ and $1 \leq i \leq t$.
- Array C is a *covering array* $\mathrm{CA}_\lambda(N; t, k, v)$ of *strength t* and index λ when every t-way interaction is λ-covered.
- $\mathrm{CAN}_\lambda(t, k, v)$ is the minimum N for which a $\mathrm{CA}_\lambda(N; t, k, v)$ exists.

Covering Array

$CA_1(13;3,10,2)$

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
1	1	1	0	1	0	0	0	0	1
1	0	1	1	0	1	0	1	0	0
1	0	0	0	1	1	1	0	0	0
0	1	1	0	0	1	0	0	1	0
0	0	1	0	1	0	1	1	1	0
1	1	0	1	0	0	1	0	1	0
0	0	0	1	1	1	0	0	1	1
0	0	1	1	0	0	1	0	0	1
0	1	0	1	1	0	0	1	0	0
1	0	0	0	0	0	0	1	1	1
0	1	0	0	0	1	1	1	0	1

Arrays over Finite Fields

Setup I

- George Sherwood suggested a framework for constructing covering arrays using finite fields.
- Let q be a prime power, and let $\mathbb{F}_q$ be the finite field of order q.
- Let $\mathcal{R}_{t,q} = \{\mathbf{r}_0, \ldots, \mathbf{r}_{q^t-1}\}$ be the set of all (row) vectors of length t with entries from $\mathbb{F}_q$, and let $\mathcal{T}_{t,q}$ be the set of all column vectors of length t with entries from $\mathbb{F}_q$, not all 0.
- A vector $\mathbf{x} \in \mathcal{T}_{t,q}$ is a *permutation vector*, so called because the multiplication of all $\mathbf{r}_i \in \mathcal{R}_{t,q}$ with $\mathbf{x}$ can be interpreted as q^{t-1} permutations of $\mathbb{F}_q$.

Arrays over Finite Fields

Setup II

Lemma

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ *be a set of vectors from* $\mathcal{T}_{t,q}$*. The array* $A = (a_{ij})$ *formed by setting* a_{ij} *to be the product of* $\mathbf{r}_i$ *and* $\mathbf{x}_j$ *is a* $\mathrm{CA}(q^t; t, t, q)$ *if and only if the* $t \times t$ *matrix* $X = [\mathbf{x}_1 \cdots \mathbf{x}_t]$ *is nonsingular.*

Proof.

Array A contains some row $\mathbf{b}$ at least twice exactly when $\mathbf{r}X = \mathbf{b}$ has more than one solution. $\square$

Arrays over Finite Fields

Setup III

- $(0, \ldots, 0)^T$ cannot appear in a nonsingular matrix, so it is not in $\mathcal{T}_{t,q}$.
- For any nonzero $\mu \in \mathbb{F}_q$, substituting $\mu \mathbf{x}_i$ for $\mathbf{x}_i$ permutes the rows does not alter the fact that it is a covering array.
- Define $\langle \mathbf{x} \rangle = \{\mu \mathbf{x} : \mu \in \mathbb{F}_q, \mu \neq 0\}$. When $\mathbf{x}$ is not all 0, we can select as the representative of $\langle \mathbf{x} \rangle$ the unique vector whose first nonzero coordinate is the multiplicative identity element.
- Let $\mathcal{V}_{t,q}$ be the set of representatives of the column vectors in $\mathcal{T}_{t,q}$.
- Then $|\mathcal{V}_{t,q}| = \frac{q^t - 1}{q - 1} = \sum_{i=0}^{t-1} q^i$.

Covering Perfect Hash Families

Definition for Higher Index

- A *covering perfect hash family* $\mathrm{CPHF}_\lambda(n; k, q, t)$ is an $n \times k$ array $C = (\mathbf{c}_{ij})$ with entries from $\mathcal{V}_{t,q}$ so that, for every set $\{\gamma_1, \ldots, \gamma_t\}$ of distinct column indices, there are at least λ row indices $\rho_1, \ldots, \rho_\lambda$ of C for which $[\mathbf{c}_{\rho_\ell \gamma_1} \cdots \mathbf{c}_{\rho_\ell \gamma_t}]$ is nonsingular for each $1 \le \ell \le \lambda$.

Lemma

When a $\mathrm{CPHF}_\lambda(n; k, q, t)$ *exists, there exists a* $\mathrm{CA}_\lambda(n(q^t - 1) + \lambda; t, k, q)$.

CPHF Asymptotics for Covering Arrays

- Choose entries of an $n \times k$ array A uniformly at random from $\mathcal{T}_{t,q}$.
- Let T be a set of t columns of A. The probability that A does not contain a covering t-set for T can easily be computed.
- The total number of t-sets is $\left(q^t - 1\right)^t$, and the number that are covering t-sets is $\prod_{i=0}^{t-1}(q^t - q^i)$.
- So within one row of A, the probability that the columns of T are *not* covering is

$$\phi_{t,q} := 1 - \frac{\prod_{i=0}^{t-1}(q^t - q^i)}{(q^t - 1)^t} = 1 - \prod_{i=1}^{t-1} \frac{q^t - q^i}{q^t - 1}.$$

Asymptotics

- $\phi_{t,q}^{N}$ is the probability that a specified t-set of columns is covered in 0 of the N rows.
- $\phi_{t,q}^{N-\ell}(1-\phi_{t,q})^{\ell}$ is the probability that a specified t-set of columns is covered in a specified choice of exactly ℓ of the N rows.
- the probability that a specified t-set of columns is covered in fewer than λ of the N rows is

$$\psi_{N,t,q,\lambda} = \phi_{t,q}^{N} \sum_{\ell=0}^{\lambda-1} \binom{N}{\ell} \left[\frac{1-\phi_{t,q}}{\phi_{t,q}}\right]^{\ell}$$

- Solving for the smallest N in $\binom{k}{t}\psi_{N,t,q,\lambda} < 1$ leads to asymptotic bounds for covering arrays of index λ!

Blemishes

- A blemish is a t-set of columns for which fewer than λ rows are covering.
- The asymptotics essentially determine the expected number of blemishes, observing when this expectation is less than 1, a CPHF exists.

The Quality of the Bound

- Dougherty (2022) observes that the bound via CPHFs fares worse and worse as λ increases, when compared to a random construction of covering arrays directly. Why does this happen?
- Moreover, even when λ is small, the bound via CPHFs does not compare well when q is very small. Why does this happen?

The Quality of the Bound

- The notion of covering for a t-set of columns in a CPHF is all or nothing, even though many t-way interactions may be covered – possibly many times – in the covering array generated despite rows of the CPHF not covering everything individually.
- Can we exploit the partial coverage obtained when a t-set is non-covering (i.e., singular) in a row of the CPHF?

Blemishes and Flaws

- A flaw is a t-way interaction that is not covered λ or more times.
- A blemish is a t-set S of columns for which at least one of the q^t t-way interactions on the columns of S is a flaw.
- This refined notion of blemish may reduce the expected number of blemishes!

Flaws

- Consider a t-set of columns, and the entries of a CPHF-like array in a specific row.
- When these entries form a $t \times t$ matrix of rank d, in the corresponding CA-like array, the generated rows cover
  - q^d t-way interactions each q^{t-d} times, and
  - the remaining $q^t - q^d$ not at all.
- ... but not all t-way interactions are equally likely.
- To correct this, choose a random $n \times k$ array whose entries are field elements ("adders")
- As the CA is generated, add the appropriate adder to each of the q^t elements in the column generated from an entry of the CPHF.

Flaws and Blemishes

- It is easy to determine the probability that the array on a t-set of columns has rank equal to d.
- And it is "easy" to determine the expected number of flaws on a t-set of columns.
- When the expected number of flaws is less than 1, this t-set is not a blemish.
- (Skipping lots of algebra,) this improves the bounds on covering array numbers.

Oversampling

- Idea: Make more columns than desired but with more blemishes.
- Delete a column from each blemish so that
  - No blemishes remain but the number of columns is as least as large as desired.

Examples

Strength 3, #Symbols 9

Index $\lambda = 1$

k	Basic CA	Basic CPHF	Basic FF	Oversample CA	Oversample CPHF	Oversample FF
10^3	18592	6553	5832	14952	5097	5097
10^6	33691	13833	13122	25018	10193	10193

Index $\lambda = 50$

k	Basic CA	Basic CPHF	Basic FF	Oversample CA	Oversample CPHF	Oversample FF
10^3	82541	57562	56862	76000	53922	52922
10^6	106693	69210	68526	93272	62658	62658

Examples

Strength 4, #Symbols 4

Index $\lambda = 1$

	Basic			Oversample		
k	CA	CPHF	FF	CA	CPHF	FF
10^3	5296	3316	3072	4708	2806	2806
10^6	14725	11221	9728	11770	8671	7936

Index $\lambda = 50$

	Basic			Oversample		
k	CA	CPHF	FF	CA	CPHF	FF
10^3	30811	30905	26624	28884	29120	25344
10^6	41675	40085	34048	37365	36515	31232

Examples

Strength 7, #Symbols 2

Index $\lambda = 1$

	Basic			Oversample		
k	CA	CPHF	FF	CA	CPHF	FF
10^3	5695	13844	5760	5181	12447	5353
10^6	11862	30608	12160	10467	26798	10752

Index $\lambda = 50$

	Basic			Oversample		
k	CA	CPHF	FF	CA	CPHF	FF
10^3	18390	50596	18176	17607	48437	17488
10^6	26923	74091	26880	25089	69011	25088

Wrapping Up

- CPHFs make great covering arrays when q is large and λ is small, BUT
- the all-or-nothing coverage underestimates the chance that a covering array is generated!
- Accounting for partial coverage, we get the best of all worlds — competitive bounds, a compact representation, ... and even fast construction algorithms!