Small arrays of maximum coverage

Shagnik Das∗† Tamás Mészáros∗‡

April 24, 2018

Abstract

Given a set $S$ of $v \geq 2$ symbols, and integers $k \geq t \geq 2$ and $N \geq 1$, an $N \times k$ array $A \in S^{N \times k}$ is said to cover a $t$-set of columns if all sequences in $S^t$ appear as rows in the corresponding $N \times t$ subarray of $A$. If $A$ covers all $t$-subsets of columns, it is called an $(N; t, k, v)$-covering array. These arrays have a wide variety of applications, driving the search for small covering arrays. Here we consider an inverse problem: rather than aiming to cover all $t$-sets of columns with the smallest possible array, we fix the size $N$ of the array to be equal to $v^t$ and try to maximise the number of covered $t$-sets. With the machinery of hypergraph Lagrangians, we provide an upper bound on the number of $t$-sets that can be covered. A linear algebraic construction shows this bound to be tight; exactly so in the case when $v$ is a prime power and $\frac{v^t - 1}{v - 1}$ divides $k$, and asymptotically so in other cases. As an application, by combining our construction with probabilistic arguments, we match the best-known asymptotics of the covering array number $\text{CAN}(t, k, v)$, which is the smallest $N$ for which an $(N; t, k, v)$-covering array exists, and improve the upper bounds on the almost-covering array number $\text{ACAN}(t, k, v, \varepsilon)$.

1 Introduction

In the last few decades, a great deal of research has been devoted to the study of orthogonal and covering arrays, an important class of combinatorial designs. This research is motivated by numerous applications, in particular to computer science and the design of experiments, and one of the major open problems in this area is to determine how small these arrays can be. In this paper we study a related problem, considering arrays of small fixed size and asking how close to covering they can be. As we shall see, this will naturally lead us to the construction of small covering arrays recently given by Colbourn, Lanus and Sarkar [6], as well as improving the upper bounds on sizes of almost-covering arrays. We shall now provide a brief introduction to the subject, before presenting our new results.

∗Freie Universität Berlin, Institut für Mathematik, Arnimallee 3, 14195 Berlin, Germany.
†E-mail: shagnik@mi.fu-berlin.de
‡E-mail: tamas.meszaros@fu-berlin.de. Position funded by the DRS POINT Fellowship Program.
1.1 Background and previous results

Let $A$ be an $N \times k$ array, whose entries come from some set $S$ of $v$ symbols; that is, $A \in S^{N \times k}$. In the context of experimental design, $N$ represents the number of trials to be carried out, $k$ denotes the number of factors to be tested, and $S$ is the set of levels these factors can take. The objective is to determine how subsets of the factors interact with one another. To that end, given a set $Q$ of $t$ column indices, we denote by $A_Q$ the $N \times t$ subarray obtained by restricting $A$ to the columns in $Q$.

**Definition 1.1 (Orthogonal arrays).** Given a set $S$ of $v$ symbols, $k$ columns, an index $\lambda \in \mathbb{N}$ and a strength $t \in \mathbb{N}$, let $N = \lambda v^t$. An $N \times k$ array $A \in S^{N \times k}$ is an $(N; t, k, v)$-orthogonal array if for every subset $Q$ of $t$ columns, every sequence in $S^t$ appears exactly $\lambda$ times as a row of the subarray $A_Q$.

An orthogonal array is therefore a very regular structure, behaving uniformly with respect to every subset of $t$ columns, giving rise to an important application in theoretical computer science. Many randomised algorithms use some large number $k$ of independent random variables, each uniformly distributed over a set $S$ of size $v$, thus using the exponentially large probability space $S^k$. However, quite often one only requires the weaker condition that the random variables be $t$-wise independent, for some small $t$. Given an $(N; t, k, v)$-orthogonal array, a uniform distribution on the $N$ rows of this array provides a probability space with the desired independence, and, if $N$ is small, this allows for brute-force derandomisation of the algorithm.

The interest, then, is in determining how few rows an orthogonal array with a given strength and number of columns can have. In one of the early papers on the subject, Plackett and Burman [21] provided sharp bounds for orthogonal arrays of strength two, showing how the number of rows must grow with the number of columns.

**Theorem 1.2 (Plackett–Burman [21], 1946).** If an $(N; 2, k, v)$-orthogonal array exists, we must have

$$k \leq \left\lfloor \frac{N - 1}{v - 1} \right\rfloor.$$ 

Following this initial focus on orthogonal arrays of strength two, Rao [22, 23] generalised the notion to arrays of strength $t$, giving rise to the modern study of orthogonal arrays. For an account of the last half-century’s developments in the field, the reader is referred to the book of Hedayat, Sloane and Stufken [12].

However, the high level of regularity required of an orthogonal array places severe restrictions on the possible values of the parameters, and hence the fundamental question asks for which parameters an $(N; t, k, v)$-orthogonal array exists. For many applications, one is willing to make do with a smaller, yet less regular, construction, giving rise to the relaxation of orthogonal arrays to covering arrays. Here, one only requires that all sequences in $S^t$ appear at least once in every $N \times t$ subarray, dropping the condition that they appear equally often. The primary question is now an extremal one — how small can an array satisfying this weaker condition be?
**Definition 1.3** (Covering arrays). Given a set $S$ of $v$ symbols, $N$ rows, $k$ columns and a strength $t \in \mathbb{N}$, let $A \in S^{N \times k}$ be an $N \times k$ array. We say that $A$ covers a subset $Q$ of $t$ columns if every sequence in $S^t$ appears at least once as a row of the subarray $A_Q$.

The array $A$ is an $(N; t, k, v)$-covering array if it covers every $t$-subset of the $k$ columns, and the covering array number $\text{CAN}(t, k, v)$ is the minimum number $N$ of rows for which an $(N; t, k, v)$-covering array exists.

The study of the covering array numbers dates back to the early 1970’s, when one of the few exact results in this field was obtained by Rényi [24] and Katona [13], and independently by Kleitman and Spencer [14], who showed

$$\text{CAN}(2, k, 2) = \min \left\{ N : k \leq \left( \left\lfloor \frac{N - 1}{2} \right\rfloor - 1 \right) \right\} = \log_2 k + \left( \frac{1}{2} + o(1) \right) \log_2 \log_2 k.$$

For higher strengths $t \geq 3$, Kleitman and Spencer [14] showed there are constants $c_1, c_2 > 0$ such that

$$c_1 2^t \log_2 k \leq \text{CAN}(t, k, 2) \leq c_2 t^2 2^t \log_2 k.$$

These results have since been extended to covering arrays over larger sets of symbols. In the strength-two case, Gargano, Körner and Vaccaro [9] established the asymptotic result

$$\text{CAN}(2, k, v) = \left( \frac{1}{2} + o(1) \right) v \log_2 k$$

when $v$ is fixed and $k$ tends to infinity. For higher strengths $t$, just as in the binary case, there is a considerable gap between the best-known lower and upper bounds on $\text{CAN}(t, k, v)$. One can prove a lower bound by induction on $t$, reducing the problem to the $t = 2$ case. One of the first general upper bounds was given by Godbole, Skipper and Sunley [10] in 1996, using a uniformly random array. This leads to the bounds

$$\left( \frac{1}{2} + o(1) \right) v^{t-1} \log_2 (k - t + 2) \leq \text{CAN}(t, k, v) \leq (1 + o(1)) \frac{(t-1) \log_2 k}{\log_2 \frac{v}{v-1}}.$$

This shows that, for fixed $t$ and $v$ and large $k$, we have $\text{CAN}(t, k, v) = O_{t,v}(\log_2 k)$, but the correct dependence on $t$ and $v$ is unknown. Research in this area has thus been devoted to the determination of the function

$$d(t, v) = \limsup_{k \to \infty} \frac{\text{CAN}(t, k, v)}{\log_2 k}.$$

The previously-mentioned bounds show $v^{t-1}/2 \lesssim d(t, v) \lesssim (t-1)/\log_2 \frac{v^t}{v^t-1}$, with the upper bound being asymptotically $(t-1)v^t$ when $v^t$ is large. In 2016, marking the first improvement on these general bounds in twenty years, Francetić and Stevens [7] and Sarkar and Colbourn [26] improved the lower-order terms of the upper bound, which resulted in much smaller covering arrays for small values of $t$ and $v$ (for a detailed account of the best-known constructions for particular values of $t$ and $v$, we refer the reader to the excellent surveys of Lawrence, Kacker, Lei, Kuhn and Forbes [15] and Colbourn [5]).
More recently, during the preparation of this manuscript, Colbourn, Lanus and Sarkar [6] achieved a significant breakthrough. Their construction improves the upper bound on $d(t, v)$ by a logarithmic factor, showing

$$d(t, v) \leq \frac{(t - 1)v^t}{2\log_2 v - \log_2(v + 1)}$$

whenever $v$ is a prime power.

Despite these improved constructions, the lower bound on $d(t, v)$ shows that when $k$ is large, any covering array must necessarily be large as well. It is thus natural to ask how much better we can do if we are more modest in our goals — rather than seeking to cover all $t$-sets of columns, what if we only want to cover most? This line of investigation was pursued by Sarkar, Colbourn, De Bonis and Vaccaro [27], who introduced the notion of almost-covering arrays.

**Definition 1.4** (Almost-covering arrays). Given a set $S$ of $v$ symbols, $N$ rows, $k$ columns and a strength $t \in \mathbb{N}$, let $A \in S^{N \times k}$ be an $N \times k$ array. Then $A$ is an $(N; t, k, v, \varepsilon)$-almost-covering array with coverage fraction $\varepsilon \in [0, 1]$ if it covers all but at most $\varepsilon \binom{k}{t}$ $t$-subsets of the $k$ columns, and $\text{ACAN}(t, k, v, \varepsilon)$ is the smallest $N$ for which an $(N; t, k, v, \varepsilon)$-almost covering array exists.

Note that when $\varepsilon < \binom{k}{t}^{-1}$, an $(N; t, k, v, \varepsilon)$-almost-covering array is an $(N; t, k, v)$-covering array. However, when $\varepsilon$ is a small constant, Sarkar et al. [27] showed the existence of very small almost-covering arrays, whose size is independent of the number of columns $k$. More precisely, they showed

$$\text{ACAN}(t, k, v, \varepsilon) \leq v^t \ln \left( \frac{v^{t-1}}{\varepsilon} \right),$$

with an improvement to $\text{ACAN}(t, k, v, \varepsilon) \leq v^t \ln \left( \frac{2v^{t-2}}{\varepsilon} \right) + v$ whenever $v$ is a prime power.

### 1.2 Maximum coverage and our results

We study what is, in some sense, an inverse problem. Rather than seeking the smallest array that covers all $t$-sets of columns, we fix the size of the array and try to maximise the number of covered $t$-sets. This gives rise to the following extremal function.

**Definition 1.5** (Maximum coverage function). Suppose we have a set $S$ of $v$ symbols, $N$ rows, $k$ columns and a strength $t \in \mathbb{N}$. For an $N \times k$ array $A \in S^{N \times k}$, let $\text{Cov}(A)$ denote the collection of all subsets of columns that are covered by $A$, and define $\text{cov}_t(A)$ to be the number of sets of size $t$ in $\text{Cov}(A)$.

We define the maximum coverage function, $\text{cov}_{\max}(N; t, k, v)$, to be the maximum number of $t$-subsets that can be covered by such an array. That is,

$$\text{cov}_{\max}(N; t, k, v) = \max \{ \text{cov}_t(A) : A \in S^{N \times k} \}.$$
We note that similar notions have appeared previously in the literature. Hartman and Raskin [11] and Maximoff, Trela, Kuhn and Kacker [16] suggested comparable lines of study, with a focus on developing heuristics for building small arrays that cover many sets. In this paper, we place our emphasis on proving general bounds for the maximum coverage function. Note that this is a refinement of the covering array and almost-covering array numbers, since we have \( \text{CAN}(t, k, v) = \min \{ N : \text{cov}_{\text{max}}(N; t, k, v) = \binom{t}{k} \} \) and \( \text{ACAN}(t, k, v, \varepsilon) = \min \{ N : \text{cov}_{\text{max}}(N; t, k, v) \geq (1 - \varepsilon) \binom{t}{k} \} \).

Now clearly, in order for an array \( A \in S_{N \times k} \) to cover even a single \( t \)-set \( Q \) of columns, we must have at least \( v^t \) rows, as each of the \( v^t \) sequences in \( S^t \) must appear as rows in \( A_Q \). This trivially gives \( \text{cov}_{\text{max}}(N; t, k, v) = 0 \) for all \( N \leq v^t - 1 \). The first problem of interest is thus to determine \( \text{cov}_{\text{max}}(v^t; t, k, v) \), and this is the case on which we focus. The best upper bound for \( \text{CAN}(t, k, v) \) comes from random arrays, and hence it is natural to first consider probabilistic constructions for our problem as well. Accordingly, let \( A_{\text{rand}, v^t} \) be a random array chosen uniformly from \( S_{v^t \times k} \). Observe that for any subset \( Q \) of \( t \) columns, the \( v^t \) rows of \( (A_{\text{rand}, v^t})_Q \) are independent and uniformly distributed over \( S^t \). The probability that these rows are all distinct, and hence that \( Q \) is covered, is thus

\[
P(Q \in \text{Cov}(A_{\text{rand}, v^t})) = \frac{(v^t)!}{(v^t)^{v^t}}.
\]

As \( v^t \) tends to infinity, this is \( e^{-(1-o(1))v^t} \), and so in expectation \( A_{\text{rand}, v^t} \) only covers an exponentially small fraction of all \( t \)-sets of columns.

A moment’s thought reveals that this is far from optimal. Indeed, consider the following block construction, where for simplicity we suppose that \( k \) is divisible by \( t \). Partition the \( k \) columns into \( t \) equal-sized subsets \( B_1, B_2, \ldots, B_t \), and build an array \( A_{\text{block}} \) whose rows are all sequences in \( S^k \) that are constant on the sets \( B_i \), \( 1 \leq i \leq t \). We clearly have exactly \( v^t \) rows, and a set \( Q \) of columns is covered if and only if \( |Q \cap B_i| \leq 1 \) for all \( i \). Therefore, when \( k \) and \( t \) are suitably large, we have

\[
\text{cov}_t(A_{\text{block}}) = \left( \frac{k}{t} \right)^t \sim \sqrt{2\pi t} \cdot e^{-t} \binom{k}{t}. 
\]

This already gives a significantly better lower bound on \( \text{cov}_{\text{max}}(v^t; t, k, v) \) than the random array \( A_{\text{rand}, v^t} \), but, as it turns out, is still far from the truth. Perhaps surprisingly, we will show that, as soon as an array is large enough to cover even a single \( t \)-set of columns, one can cover a large proportion of all \( t \)-sets. This proportion is given by the following constant.

**Definition 1.6.** Given \( t \geq 1 \) and \( v \geq 2 \), we define

\[
c_{t,v} = \prod_{i=0}^{t-1} \frac{v^t - v^i}{v^t - 1}. 
\]

Observe that this really is a large proportion:

\[
c_{t,v} = \prod_{i=0}^{t-1} \frac{v^t - v^i}{v^t - 1} \geq \prod_{i=1}^{t-1} (1 - v^{-i}) \geq (1 - v^{-1}) \left( 1 - \sum_{i=2}^{\infty} v^{-i} \right) = 1 - \frac{v + 1}{v^2}.
\]
In our main result below we give an upper bound on \( \text{cov}_{\text{max}}(v^t; t, k, v) \) and, whenever \( v \) is a prime power, demonstrate its tightness by means of an explicit construction.

**Theorem 1.7.** For any \( v \geq 2 \) and \( k \geq t \geq 2 \), we have the bound

\[
\text{cov}_{\text{max}}(v^t; t, k, v) \leq c_{t,v} \frac{k^t}{t!}.
\]

If \( v \) is a prime power, and \( \frac{v^t - 1}{v - 1} \) divides \( k \), we have equality; that is, \( \text{cov}_{\text{max}}(v^t; t, k, v) = c_{t,v} \frac{k^t}{t!} \). Otherwise, for prime power \( v \) and all \( k \geq t \), we have

\[
\text{cov}_{\text{max}}(v^t; t, k, v) \geq c_{t,v} \left( \binom{k}{t} \right).
\]

Although the more novel contribution of this paper is the upper bound of Theorem 1.7, the constructive lower bound has implications for the covering array and almost-covering array numbers previously introduced. As shown in the following corollary, we can improve the upper bound on \( \text{ACAN}(t, k, v, \varepsilon) \) by a logarithmic factor, while also rediscovering the recent upper bound on \( d(t, v) \) due to Colbourn et al. \([6]\). We also provide some bounds in the non-prime-power case.

**Corollary 1.8.** For integers \( k \geq t \geq 2 \), real \( \varepsilon > 0 \) and prime power \( v \geq 2 \), we have

\[
\text{ACAN}(t, k, v, \varepsilon) \leq v^t \left[ \frac{\ln \frac{1}{\varepsilon}}{2 \ln v - \ln(v + 1)} \right] \quad \text{and} \quad d(t, v) \leq \frac{(t - 1)v^t}{2 \log_2 v - \log_2(v + 1)}.
\]

Moreover, there is some absolute constant \( v_0 \in \mathbb{N} \) such that for all integers \( v \geq v_0 \), we have

\[
\text{ACAN}(t, k, v, \varepsilon) \leq e^{tv^{-0.474}} v^t \left[ \frac{\ln \frac{1}{\varepsilon}}{2 \ln v - \ln(v + 1)} \right] \quad \text{and} \quad d(t, v) \leq \frac{(t - 1)e^{tv^{-0.474}} v^t}{2 \log_2 v - \log_2(v + 1)}.
\]

Observe that the optimal array for Theorem 1.7 itself shows \( \text{ACAN}(t, k, v, \varepsilon) = v^t \) for \( \varepsilon \geq 1 - c_{t,v} \), which includes the range \( \varepsilon \geq \frac{v + 1}{v^2} \).

### 1.3 Organisation and notation

The remainder of this paper is organised as follows. In Section 2 we prove the upper bound of Theorem 1.7, and in Section 3 we establish the lower bound by providing an optimal construction. Section 4 is devoted to the construction of small covering and almost-covering arrays, proving Corollary 1.8. Finally in Section 5 we provide some concluding remarks and open problems.

We use standard combinatorial notation throughout this paper. In particular, \([k]\) denotes the set of the first \( k \) positive integers, \( \{1, 2, \ldots, k\} \). Given a set \( X \) and an integer \( t \), \( \binom{X}{t} \) is the collection of all \( t \)-subsets of \( X \). As defined previously, for an array \( A \) and a subset \( Q \) of its columns, \( A_Q \) denotes the subarray of \( A \) containing only the columns in \( Q \). We denote by \( \text{Cov}(A) \) the collection of all subsets of columns that are covered by \( A \), and by \( \text{cov}_t(A) \) the number of sets of size \( t \) in \( \text{Cov}(A) \). Finally, we use \( \log_2 \) for the binary logarithm, and \( \ln \) for the natural logarithm.
2 A general upper bound

We shall start our proof of Theorem 1.7 by proving the general upper bound. In order to do this, we shall show that the number of \( t \)-sets that can be covered by an array of size \( v^t \times k \) is bounded by the Lagrangian of an auxiliary hypergraph, which we shall then estimate. First we present some useful preliminaries concerning Lagrangians in general.

2.1 Lagrangians of \( t \)-uniform hypergraphs

Lagrangians, first introduced by Motzkin and Straus \[19\] to give a proof of Turán’s theorem, have proven to be very useful in the study of extremal combinatorics. Roughly speaking, the Lagrangian of a hypergraph determines the maximum possible density of a blow-up of the hypergraph. We now define the Lagrangian more precisely.

Definition 2.1. Let \( H \subseteq \binom{U}{t} \) be a \( t \)-uniform hypergraph on a finite set \( U \) of vertices. We say that a function \( x : U \to \mathbb{R} \) is a legal weighting if \( x(u) \geq 0 \) for every \( u \in U \) and \( \sum_{u \in U} x(u) = 1 \).

The weight polynomial of \( H \) evaluated at this weighting is given by

\[
w(x, H) = \sum_{e \in H} \prod_{u \in e} x(u).
\]

The Lagrangian of \( H \) is defined to be \( \lambda(H) = \max w(x, H) \), where the maximum is taken over all legal weightings \( x \) of \( H \).

We call a legal weighting \( x \) optimal if \( w(x, H) = \lambda(H) \). The following lemma, proven by Frankl and Rödl \[8\], gives some information about the minimal supports of optimal weightings.

Lemma 2.2. Let \( H \) be a \( t \)-uniform hypergraph on the vertex set \( U \), and suppose \( x \) is an optimal weighting where the number of vertices with non-zero weight is minimal. If \( u, w \in U \) are vertices of non-zero weight, then there is an edge in \( H \) containing both \( u \) and \( w \).

Proof. Suppose for contradiction that vertices \( u \) and \( w \) have positive weight, but there is no edge of \( H \) containing both of them. We define a parametrised function \( x_\varepsilon : U \to \mathbb{R} \), where

\[
x_\varepsilon(s) = \begin{cases} x(u) + \varepsilon & \text{if } s = u, \\ x(w) - \varepsilon & \text{if } s = w, \\ x(s) & \text{otherwise.} \end{cases}
\]

Observe that \( x_\varepsilon \) is a legal weighting whenever \( \varepsilon \in [-x(u), x(w)] \), and that at the boundaries either \( x_\varepsilon(u) \) or \( x_\varepsilon(w) \) becomes zero. Moreover, since there is no edge containing both \( u \) and \( w \), \( w(x_\varepsilon, H) \) is linear in \( \varepsilon \), and hence is maximised by some \( \varepsilon^* \in \{-x(u), x(w)\} \). However, this gives a contradiction, as \( x_\varepsilon^* \) is then an optimal weighting with fewer non-zero weights. Thus \( u \) and \( w \) must appear together in some edge of \( H \). \( \square \)
The next lemma shows that we can bound the Lagrangian $\lambda(\mathcal{H})$ of a $t$-uniform hypergraph $\mathcal{H}$ in terms of the Lagrangians of smaller $(t-1)$-uniform hypergraphs. Given a vertex $u \in U$, the link hypergraph $\mathcal{H}(u)$ is a $(t-1)$-uniform hypergraph on the vertex set $U \setminus \{u\}$, with edges $e' \in \mathcal{H}(u)$ whenever $e' \cup \{u\} \in \mathcal{H}$.

**Lemma 2.3.** Given $t \geq 2$, a $t$-uniform hypergraph $\mathcal{H}$ on the vertex set $U$ and a legal weighting $x$, we have

$$w(x, \mathcal{H}) \leq \frac{1}{t} \sum_{u \in U} x(u)(1 - x(u))^t \lambda(\mathcal{H}(u)).$$

*Proof.* Note that if $x(u) = 1$ for any $u \in U$, then $w(x, \mathcal{H}) = 0$, and the inequality trivially holds. Hence we may assume $x(u) < 1$ for all $u \in U$. Since every edge of $\mathcal{H}$ has exactly $t$ vertices, double-counting gives

$$w(x, \mathcal{H}) = \sum_{e \in \mathcal{H}} \prod_{u \in e} x(u) = \frac{1}{t} \sum_{u \in U} \sum_{u \in \mathcal{H}} \prod_{e \in \mathcal{H}(u)} x(u) = \frac{1}{t} \sum_{u \in U} x(u) \sum_{e \subseteq \mathcal{H}(u)} \prod_{w \in e} x(w).$$

As $x$ is a legal weighting, for every vertex $u$, $x$ must distribute a total weight of $1 - x(u)$ on the vertices in $U \setminus \{u\}$. Thus we can rescale the weights to obtain a legal weighting for the link hypergraph $\mathcal{H}(u)$ by defining $x_u(w) = \frac{x(w)}{1 - x(u)}$ for all vertices $w \in U \setminus \{u\}$. We then have

$$w(x, \mathcal{H}) = \frac{1}{t} \sum_{u \in U} x(u) \sum_{e' \subseteq \mathcal{H}(u)} \prod_{w \in e'} x(w) = \frac{1}{t} \sum_{u \in U} x(u)(1 - x(u))^{t-1} \sum_{e \subseteq \mathcal{H}(u)} \prod_{w \in e} x_u(w)$$

$$= \frac{1}{t} \sum_{u \in U} x(u)(1 - x(u))^{t-1} w(x_u, \mathcal{H}(u)) \leq \frac{1}{t} \sum_{u \in U} x(u)(1 - x(u))^{t-1} \lambda(\mathcal{H}(u)),$$  

where the inequality follows from the definition of the Lagrangian. $\blacksquare$

The final lemma of this subsection concerns the Lagrangian of the $v$-fold tensor product of a hypergraph with itself. Given $t$-uniform hypergraphs, $\mathcal{H}_1, \ldots, \mathcal{H}_v$ on vertex sets $U_1, \ldots, U_v$ respectively, their tensor product, denoted $\otimes_{\ell=1}^v \mathcal{H}_\ell$, is a $t$-uniform hypergraph on the vertex set $\prod_{\ell=1}^v U_\ell$ with edges

$$\otimes_{\ell=1}^v \mathcal{H}_\ell = \{ \{(u_{1,1}, \ldots, u_{v,1}), \ldots, (u_{1,\ell}, \ldots, u_{v,\ell})\} : \forall \ell \in [v], \{u_{\ell,1}, \ldots, u_{\ell,t}\} \in \mathcal{H}_\ell \}.$$

Note that every $v$-tuple of edges $(e_1, \ldots, e_v) \in \prod_{\ell=1}^v \mathcal{H}_\ell$ gives rise to $(t!)^{v-1}$ edges in $\otimes_{\ell=1}^v \mathcal{H}_\ell$, as the vertices can be combined in every possible order. We write $\mathcal{H}^{\otimes v}$ for the hypergraph obtained by taking a tensor product of $v$ copies of a hypergraph $\mathcal{H}$.

**Lemma 2.4.** For every $t$-uniform hypergraph $\mathcal{H}$,

$$\lambda(\mathcal{H}^{\otimes v}) = \lambda(\mathcal{H}).$$
Proof. Let $U$ be the vertex set of $\mathcal{H}$. First consider an optimal weighting $x$ for $\mathcal{H}$, and define a legal weighting $\widehat{x}$ on $U^v$ as follows:

$$\widehat{x}((u_1, u_2, \ldots, u_v)) = \begin{cases} x(u_1) & \text{if } u_1 = u_2 = \ldots = u_v, \\ 0 & \text{otherwise.} \end{cases}$$

Now for every edge $\{u_1, \ldots, u_t\}$ in $\mathcal{H}$, the edge $\{(u_1, \ldots, u_1), \ldots, (u_t, \ldots, u_t)\}$ appears in $\mathcal{H}^{\otimes v}$ with the same weight, and so

$$\lambda(\mathcal{H}^{\otimes v}) \geq w(\widehat{x}, \mathcal{H}^{\otimes v}) \geq w(x, \mathcal{H}) = \lambda(\mathcal{H}).$$

For the reverse inequality, let $\widehat{x}$ be an optimal weighting for $\mathcal{H}^{\otimes v}$, and define the legal weighting $x$ on $U$ by $x(u) = \sum_{u_2, \ldots, u_v \in U} \widehat{x}((u, u_2, \ldots, u_v))$. We then have

$$\lambda(\mathcal{H}) \geq w(x, \mathcal{H}) = \sum_{\{u_1, \ldots, u_t\} \in \mathcal{H}} \prod_{i=1}^{t} x(u_i) = \sum_{\{u_1, \ldots, u_t\} \in \mathcal{H}} \prod_{u_j, i \in U, 2 \leq j \leq v, i \in [t]} \widehat{x}((u_i, u_{2,i}, \ldots, u_{v,i}))$$

$$\geq \sum_{\{u_1, \ldots, u_t\} \in \mathcal{H}} \prod_{i=1}^{t} \widehat{x}((u_i, u_{2,i}, \ldots, u_{v,i}))$$

$$= \sum_{\{(u_1, u_{2,1}, \ldots, u_{v,1}), \ldots, (u_t, u_{2,t}, \ldots, u_{v,t})\} \in \mathcal{H}^{\otimes v}} \prod_{i=1}^{t} \widehat{x}((u_i, u_{2,i}, \ldots, u_{v,i})) = w(\widehat{x}, \mathcal{H}^{\otimes v}) = \lambda(\mathcal{H}^{\otimes v}),$$

and thus $\lambda(\mathcal{H}) = \lambda(\mathcal{H}^{\otimes v})$. \hfill \Box

2.2 The upper bound

With these preliminaries in place, we may proceed with the proof of the upper bound in Theorem 1.7. As mentioned earlier, we shall bound the number of covered $t$-sets by the Lagrangian of an auxiliary hypergraph $\mathcal{H}_{t,v}$, which we now introduce.

The vertex set of $\mathcal{H}_{t,v}$ is $[v]^v$ and $t$ vectors $\vec{y}_1, \ldots, \vec{y}_t \in [v]^v$ form an edge in $\mathcal{H}_{t,v}$ whenever all $v^t$ vectors in $[v]^t$ appear as rows of the $v^t \times t$ matrix whose columns are $\vec{y}_1, \ldots, \vec{y}_t$. Note that this condition implies the vectors $\vec{y}_i$ are pairwise-distinct, so this is indeed a $t$-uniform hypergraph.

For a $v^t \times k$ array $A$, whose entries we shall assume to belong to $[v]$, and $\vec{y} \in [v]^v$, define $B_{\vec{y}} = \{ a \in [k] : A_{\{a\}} = \vec{y} \}$ and note that the sets $B_{\vec{y}}$ partition the set $[k]$ of columns of $A$. Observe that a $t$-set $Q \subseteq [k]$ can only be covered by $A$ if all elements of $Q$ belong to different parts $B_{\vec{y}}$, as otherwise two of the columns in $A_Q$ will be identical. However, not all such $t$-sets are covered. A $t$-set $Q \subseteq [k]$ is covered by $A$ if and only if the columns of $A_Q$ form an edge of $\mathcal{H}_{t,v}$.

We can hence count the number of covered $t$-sets, finding

$$\text{cov}_t(A) = \sum_{e \in \mathcal{H}_{t,v}} |B_{\vec{y}}|.$$

9
Since the sets \( \{ B_\vec{y} : \vec{y} \in [v]^{v^t} \} \) partition \([k]\), the function \( x : [v]^{v^t} \to \mathbb{R} \) given by \( x(\vec{y}) = \frac{1}{k} |B_\vec{y}| \) is a legal weighting of the vertices of \( \mathcal{H}_{t,v} \). Hence
\[
\text{cov}_t(A) = \sum_{e \in \mathcal{H}_{t,v}} \prod_{\vec{y} \in e} |B_\vec{y}| = k^t \sum_{e \in \mathcal{H}_{t,v}} \prod_{\vec{y} \in e} x(\vec{y}) = k^t w(x, \mathcal{H}_{t,v}) \leq k^t \lambda(\mathcal{H}_{t,v}).
\]

Now note that the hypergraph \( \mathcal{H}_{t,v} \) is in fact independent of the array \( A \) (which only determines the weighting of the vertices), and hence \( k^t \lambda(\mathcal{H}_{t,v}) \) bounds the number of \( t \)-sets that can be covered by any array of size \( v^t \times k \). The following proposition therefore gives the desired upper bound for Theorem 1.7.

**Proposition 2.5.** For all \( t \geq 1 \),
\[
\lambda(\mathcal{H}_{t,v}) \leq \frac{c_{t,v}}{t!} = \frac{1}{t!} \prod_{i=0}^{t-1} \frac{v^t - v^i}{v^t - 1}.
\]

We first establish a few simple lemmas that we shall use when proving Proposition 2.5.

**Lemma 2.6.** There is an optimal weighting \( x \) of \( \mathcal{H}_{t,v} \) for which
\[
|\text{supp}(x)| = \left| \left\{ \vec{y} \in [v]^{v^t} : x(\vec{y}) \neq 0 \right\} \right| \leq \frac{v^t - 1}{v - 1}.
\]

**Proof.** Let \( x \) be an optimal weighting of \( \mathcal{H}_{t,v} \) minimising the number of vectors with non-zero weight. By Lemma 2.2 if \( \vec{y} \) and \( \vec{z} \) are vectors with non-zero weight, then there must be an edge \( e \in \mathcal{H}_{t,v} \) containing both \( \vec{y} \) and \( \vec{z} \). Since every vector in \([v]^t\) appears as a row in the matrix whose columns are the vectors in \( e \), it follows that for each choice of \( a, b \in [v] \), there are exactly \( v^t - 1 \) coordinates \( i \) where \( y_i = a \) and \( z_i = b \). In particular, the vectors of non-zero weight form the columns of an \( (v^t; 2, k, v) \)-orthogonal array, and so by Theorem 1.2 their number is bounded from above by \( \frac{v^t - 1}{v - 1} \). \( \square \)

The next lemma describes the link hypergraphs of \( \mathcal{H}_{t,v} \).

**Lemma 2.7.** For any vertex \( \vec{y} \in [v]^{v^t} \) in \( \mathcal{H}_{t,v} \) of positive degree, \( \mathcal{H}_{t,v}(\vec{y}) \cong \mathcal{H}_{t-1,v}^{\otimes v} \).

**Proof.** Recall that we have an edge \( \{ \vec{y}_1, \ldots, \vec{y}_{t-1}, \vec{y} \} \in \mathcal{H}_{t,v} \) if and only if the \( v^t \times t \) matrix \( M \) formed with these column vectors has all vectors in \([v]^t\) as row vectors. In particular, \( \vec{y} \) must have \( v^{t-1} \) entries equal to \( a \) for every \( a \in [v] \). Given \( a \in [v] \), denote by \( M^{(a)} \) the \( v^{t-1} \times (t-1) \) matrix formed by taking those rows of \( M \) that end with \( a \), and then deleting the last (all-\( a \)) column. Since the rows of \( M \) contain every vector of \([v]^t\) ending in \( a \), it follows that the rows of \( M^{(a)} \) consist of all vectors in \([v]^{t-1} \). In particular, the columns of \( M^{(a)} \) form an edge in \( \mathcal{H}_{t-1,v} \).

In other words, after a possible reordering of the rows of \( M \), for \( 1 \leq i \leq t - 1 \) the vector \( \vec{y}_i \in [v]^{v^t} \) should be the concatenation of vectors \( \vec{w}_{j,i} \in [v]^{v^{i-1}} \), \( j \in [v] \), such that for \( 1 \leq j \leq v \) we have \( \{ \vec{w}_{j,1}, \ldots, \vec{w}_{j,t-1} \} \in \mathcal{H}_{t-1,v} \). This correspondence between the edges \( \{ \vec{y}_1, \ldots, \vec{y}_{t-1} \} \in \mathcal{H}_{t,v}(\vec{y}) \) and \( \{(\vec{w}_{1,1}, \ldots, \vec{w}_{v,1}), \ldots, (\vec{w}_{1,t-1}, \ldots, \vec{w}_{v,t-1})\} \in \mathcal{H}_{t-1,v}^{\otimes v} \) gives the desired isomorphism between \( \mathcal{H}_{t,v}(\vec{y}) \) and \( \mathcal{H}_{t-1,v}^{\otimes v} \). \( \square \)
The final lemma solves an optimisation problem that shall appear in our proof of Proposition 2.5.

**Lemma 2.8.** For \( K \geq t \geq 2 \), let \( f(x_1, \ldots, x_K) = \sum_{i=1}^{K} x_i(1 - x_i)^{t-1} \). The maximum of \( f \), subject to \( x_i \geq 0 \) for every \( 1 \leq i \leq K \) and \( \sum_i x_i = 1 \), is \( (1 - \frac{1}{K})^{t-1} \).

**Proof.** Note that \( f \) is a continuous function and the constraints define a compact set, so the maximum is well defined. By taking \( x_i = \frac{1}{K} \) for all \( i \), we find \( f \left( \frac{1}{K}, \ldots, \frac{1}{K} \right) = (1 - \frac{1}{K})^{t-1} \), and so we only need to prove the upper bound.

Set \( g(x) = x(1 - x)^{t-1} \), and observe that

\[
g'(x) = (1 - tx)(1 - x)^{t-2} \quad \text{and} \quad g''(x) = (tx - 2)(t - 1)(1 - x)^{t-3}.
\]

Hence \( g(x) \) is monotone increasing on \([0, \frac{1}{t}]\), monotone decreasing on \([\frac{1}{t}, 1]\), concave on \([0, \frac{2}{t}]\) and convex on \([\frac{2}{t}, 1]\).

Now let \( (x_1, \ldots, x_K) \) maximise \( f \), and suppose there was some index \( i_0 \) such that \( x_{i_0} > \frac{1}{t} \). Since \( K \geq t \) and \( \sum_i x_i = 1 \), the average of the weights \( x_i \) is at most \( \frac{1}{t} \), and hence there must be some index \( j_0 \) such that \( x_{j_0} < \frac{1}{t} \). Let \( \varepsilon = \min \left\{ \frac{1}{t} - x_{j_0}, x_{i_0} - \frac{1}{t} \right\} > 0 \). If we replace \( x_{j_0} \) by \( x_{j_0} + \varepsilon \) and \( x_{i_0} \) by \( x_{i_0} - \varepsilon \), the monotonicity properties of \( g \) imply the value of \( f \) would increase, contradicting the fact that \( (x_1, \ldots, x_K) \) maximises \( f \).

Hence \( x_i \in \left[0, \frac{1}{t}\right] \) for \( 1 \leq i \leq K \). As \( g \) is concave on this interval, Jensen’s inequality gives

\[
f(x_1, \ldots, x_K) = \sum_{i=1}^{K} x_i(1 - x_i)^{t-1} = \sum_{i=1}^{K} g(x_i) \leq Kg \left( \frac{1}{K} \sum_{i=1}^{K} x_i \right) = \left( 1 - \frac{1}{K} \right)^{t-1}. \quad \square
\]

We are now in position to bound the Lagrangian of the hypergraph \( H_{t,v} \), thereby completing the proof of the upper bound from Theorem 1.7.

**Proof of Proposition 2.5** For fixed \( v \), we shall prove \( \lambda(H_{t,v}) \leq \frac{c_{1,v}}{t} \) by induction on \( t \).

For the base case \( t = 1 \), the hypergraph \( H_{v,1} \) is very simple. We have \( v^v \) vertices corresponding to the vectors in \( [v]^v \). The edges of \( H_{v,1} \) are the singletons corresponding to vectors containing every \( a \in [v] \). We thus have \( v! \) edges, and the weight polynomial is simply the sum of the weights of the corresponding \( v! \) vertices, whose maximum value is trivially at most 1, which is equal to \( c_{1,v} \).

For the induction step, suppose \( \lambda(H_{t-1,v}) \leq \frac{c_{t-1,v}}{(t-1)!} \) and consider \( H_{t,v} \). Let \( x \) be an optimal weighting of \( H_{t,v} \) with minimal support. Suppose \( \vec{y}_1, \ldots, \vec{y}_K \) are the vertices of \( H_{t,v} \) with non-zero weight, and let \( x_1, \ldots, x_K \) represent their respective weights. By Lemma 2.6 \( K \leq \frac{v^{t-1}}{v-1} \).

Using Lemma 2.3 we find

\[
\lambda(H_{t,v}) = w(x, H_{t,v}) \leq \frac{1}{t} \sum_{\vec{y} \in [v]^v} x(\vec{y})(1 - x(\vec{y}))^{t-1} \lambda(H_{t,v}(\vec{y})) = \frac{1}{t} \sum_{i=1}^{K} x_i(1 - x_i)^{t-1} \lambda(H_{t,v}(\vec{y}_i)).
\]
By Lemma 2.7, $H_{t,v}(\vec{y}_i) \cong H_{t-1,v}^{c_{t-1,v}}$ for all $i$, and hence Lemma 2.4 and the induction hypothesis give $\lambda(H_{t,v}(\vec{y}_i)) = \lambda(H_{t-1,v}) \leq \frac{c_{t-1,v}}{(t-1)!}$ for all $i$. Thus

$$\lambda(H_{t,v}) \leq \frac{c_{t-1,v}}{t!} \sum_{i=1}^{K} x_i (1 - x_i)^{t-1}. $$

Since the support must span at least one edge of $H_{t,v}$, we have $K \geq t \geq 2$. Moreover, since $x$ is a legal weighting, $x_i \geq 0$ for all $i$, and $\sum_i x_i = 1$. Hence we may apply Lemma 2.8 to deduce that $\lambda(H_{t,v}) \leq \frac{c_{t-1,v}}{t!} \left( 1 - \frac{1}{K} \right)^{t-1}$. As $K \leq \frac{v^t-1}{v-1}$, this can be further bounded by $\lambda(H_{t,v}) \leq \frac{c_{t-1,v}}{t!} \left( 1 - \frac{v^t-1}{v-1} \right)^{t-1}$. Substituting in the definition of $c_{t-1,v}$, we obtain

$$\lambda(H_{t,v}) \leq \frac{c_{t-1,v}}{t!} \left( 1 - \frac{v^t-1}{v^t-1} \right)^{t-1} = \frac{1}{t!} \left( \prod_{i=0}^{t-2} \frac{(v^t-1) - v^i}{v^t-1} \right) \left( 1 - \frac{v^t-1}{v^t-1} \right)^{t-1} = \frac{1}{t!} \prod_{i=0}^{t-2} \left( \frac{v^t-1 - v^i}{v^t-1} \right) \left( v^t - v^i \right)$$

$$= \frac{1}{t!} \prod_{i=0}^{t-2} \frac{v^t - v^{i+1}}{v^t - 1} = \frac{1}{t!} \prod_{i=1}^{t-1} \frac{v^t - v^i}{v^t - 1} = \frac{1}{t!} \prod_{i=0}^{t-1} \frac{v^t - v^i}{v^t - 1} = c_{t,v},$$

completing the proof. \qed

3 \hspace{1cm} An optimal construction

In this section we prove the lower bounds of Theorem 1.7 by providing, for every prime power $v$, a linear algebraic construction of an array of size $v^t \times k$ that covers a large number of $t$-sets. The use of linear algebra in constructing such arrays is well-established; indeed, in the case $v = 2$, we were inspired by Sylvester’s \cite{31} construction of Hadamard matrices from 1867 (see \cite{4} for a linear algebraic description of these matrices).

We begin by handling the case $k = \frac{v^t-1}{v-1}$. Let $F_v$ be the $v$-element field, and consider the vector space $F_v^t$. Note that there are exactly $\frac{v^t-1}{v-1}$ 1-dimensional subspaces $L_1, \ldots, L_k$, and for each such subspace $L_i$, fix some non-zero vector $\vec{z}_i \in L_i \subseteq F_v^t$. Now let $A_{\text{opt}}$ be a $v^t \times \frac{v^t-1}{v-1}$ array whose rows are indexed by the $v^t$ vectors in $F_v^t$ and whose columns are indexed by $[k]$. Given $\vec{y} \in F_v^t$ and $i \in [k]$, we define the $(\vec{y}, i)$ entry of $A_{\text{opt}}$ to be the scalar product $\vec{y} \cdot \vec{z}_i$.

As argued by Sherwood, Martirosyan and Colbourn \cite{28}, we claim that a $t$-set $Q \subseteq [k]$ is covered by $A_{\text{opt}}$ if and only if the corresponding vectors $\{ \vec{z}_i : i \in Q \}$ are linearly independent. To see this, let $M$ be the $t \times t$ matrix containing the vectors $\{ \vec{z}_i : i \in Q \}$ as columns. For any $\vec{y} \in F_v^t$, the corresponding row of $(A_{\text{opt}})_Q$ is $\vec{y}^T M$. Now the set $Q$ is covered by $A_{\text{opt}}$ if every possible vector occurs in this way, or, equivalently, if any vector from $F_v^t$ can be obtained as $\vec{y}^T M$ for some appropriate vector $\vec{y} \in F_v^t$. This happens precisely when the matrix $M$ is invertible, i.e. when the column vectors are linearly independent.
Now how many linearly independent sets \( \{ \vec{z}_{j_1}, \ldots, \vec{z}_{j_t} \} \) are there? If we have already chosen \( i \geq 0 \) linearly independent vectors \( \vec{z}_{j_1}, \ldots, \vec{z}_{j_i} \), the next vector \( \vec{z}_{j_{i+1}} \), and hence the corresponding subspace \( L_{j_{i+1}} \) it belongs to, cannot be in the \( i \)-dimensional subspace spanned by \( \{ \vec{z}_{j_1}, \ldots, \vec{z}_{j_i} \} \). This forbids \( \frac{v^t - 1}{v - 1} \) of the possible vectors, leaving \( \frac{v^t - 1}{v - 1} - \frac{v^{t-1} - 1}{v - 1} \) choices for \( \vec{z}_{j_{i+1}} \). As the order in which the vectors are chosen does not matter, this gives a total of

\[
\frac{1}{t!} \prod_{i=0}^{t-1} \left( \frac{v^t - 1}{v - 1} - \frac{v^i - 1}{v - 1} \right) = \frac{1}{t!} \prod_{i=0}^{t-1} \frac{v^t - v^i}{v - 1} = \left( \frac{v^t - 1}{v - 1} \right)^t \frac{c_{t,v}}{t!} = \frac{c_{t,v}}{t!} k^t \]

different linearly independent sets of size \( t \), and hence this is also the number of \( t \)-sets covered by \( A_{opt} \). Note that this exactly matches the upper bound from Section 2, which implies that in Proposition 2.5 we in fact determine the Lagrangian of the hypergraph \( \mathcal{H}_{t,v} \) precisely for all prime powers \( v \).

If \( k \) is divisible by \( \frac{v^t - 1}{v - 1} \), then we can take a blow-up of this linear algebraic construction, similar to the block construction in Section 1. Partition the set \( [k] \) of column indices into \( \frac{v^t - 1}{v - 1} \) parts of equal size, so that we have parts \( \{ B_i : i \in \left[ \frac{v^t - 1}{v - 1} \right] \} \) with \( |B_i| = \frac{k(v - 1)}{v^t - 1} \) for all \( i \). We can now define the blown-up array \( A_{opt, \text{block}} \) of size \( v^t \times k \), where for \( j \in B_i \), the \( j \)th column of \( A_{opt, \text{block}} \) is the \( i \)th column of \( A_{opt} \).

It is easy to see that a set \( Q \subseteq [k] \) is covered by \( A_{opt, \text{block}} \) if and only if it contains at most one element from each block and the corresponding set of block indices is covered by \( A_{opt} \). Thus

\[
\text{Cov}(A_{opt, \text{block}}) = \bigcup_{Q' \in \text{Cov}(A_{opt})} \prod_{i \in Q'} B_i.
\]

Restricting to sets of size \( t \), we find

\[
\text{cov}_t(A_{opt, \text{block}}) = \sum_{\{j_1, \ldots, j_t\} \in \text{Cov}(A_{opt})} \prod_{i=1}^t |B_{j_i}| = \sum_{\{j_1, \ldots, j_t\} \in \text{Cov}(A_{opt})} \left( \frac{k(v - 1)}{v^t - 1} \right)^t c_{t,v} = \left( \frac{k(v - 1)}{v^t - 1} \right)^t c_{t,v} = c_{t,v} \frac{k^t}{t!},
\]

showing we have equality in Theorem 1.7 in this case.

For general \( k \), as noted by Alon (personal communication), one can take a random partition of the set \( [k] \) of column indices into the \( \frac{v^t - 1}{v - 1} \) parts \( \{ B_i : i \in \left[ \frac{v^t - 1}{v - 1} \right] \} \). More precisely, for each \( j \in [k] \) choose some \( i \in \left[ \frac{v^t - 1}{v - 1} \right] \) independently and uniformly at random, and add \( j \) to \( B_i \). We can now take the corresponding block construction \( A \) based on \( A_{opt} \), where for each \( j \in B_i \), the \( j \)th column of \( A \) is the \( i \)th column of \( A_{opt} \). We again have that a set \( Q \) is covered by \( A \) if and only if it contains at most one element from each block and the corresponding set of block indices is covered by \( A_{opt} \). Accordingly, for a fixed \( t \)-set \( Q \),

\[
\mathbb{P}(Q \in \text{Cov}(A)) = \frac{t! \text{cov}_t(A_{opt})}{\left( \frac{v^t - 1}{v - 1} \right)^t} = c_{t,v},
\]
and so by linearity of expectation the expected number of $t$-sets covered by $A$ is $c_{t,v}(k)$. Hence there must be some block partition for which the number of covered $t$-sets is at least $c_{t,v}(k)$. This concludes the proof of Theorem 1.7.

4 Covering and almost-covering arrays

In this section we will combine random copies of our optimally-constructed array of $v^t$ rows in order to construct small covering and almost-covering arrays, thus proving Corollary 1.8.

Proof of Corollary 1.8. Given integers $k \geq t \geq 2$, real $\varepsilon > 0$ and a prime power $v \geq 2$, we start by proving the upper bound on $\text{ACAN}(t, k, v, \varepsilon)$. For this, we must construct an $(rv^t; t, k, v, \varepsilon)$-almost-covering array, where $r = \left\lceil \frac{\ln \frac{1}{\varepsilon}}{2 \ln v - \ln (v+1)} \right\rceil$.

For $1 \leq \ell \leq r$, let $A^{(\ell)}$ be an independent copy of our construction from Section 3. That is, for each $j \in [k]$, let $i_{\ell,j} \in \left\lceil \frac{rv^t-1}{v} \right\rceil$ be chosen independently and uniformly at random, and take the $j$th column of $A^{(\ell)}$ to be $(A_{\text{opt}})_{i_{\ell,j}}$. We then take $A$ to be the concatenation of $A^{(1)}, A^{(2)}, \ldots, A^{(r)}$, giving an $rv^t \times k$ array. We wish to show that $A$ is the desired almost-covering array with positive probability.

Hence, for each $t$-subset $Q$ of the columns, we let $\mathcal{E}_Q$ be the event that $Q$ is not covered by $A$. Since the array $A$ contains each array $A^{(\ell)}$, $\ell \in [r]$, if $Q$ is not covered by $A$, then it is not covered by any $A^{(\ell)}$. From Section 3 we have $\mathbb{P}(Q \notin \text{Cov}(A^{(\ell)})) = 1 - c_{t,v}$, and by construction these subarrays are independent of one another. Therefore, we have

$$\mathbb{P}(\mathcal{E}_Q) \leq \mathbb{P}(\bigwedge_{\ell \in [r]} \{Q \notin \text{Cov}(A^{(\ell)})\}) = (1 - c_{t,v})^r \leq \varepsilon,$$

where the final inequality follows from our choice of $r = \left\lceil \frac{\ln \frac{1}{\varepsilon}}{2 \ln v - \ln (v+1)} \right\rceil$ and the fact that $1 - c_{t,v} \leq \frac{v+1}{v^t}$. By linearity of expectation, the expected number of $t$-sets $Q$ of columns for which the event $\mathcal{E}_Q$ occurs is at most $\varepsilon \binom{k}{t}$, and hence $A$ has at most $\varepsilon \binom{k}{t}$ uncovered $t$-sets of columns with positive probability. This proves the existence of almost-covering arrays of size $rv^t$, establishing the desired upper bound on $\text{ACAN}(t, k, v, \varepsilon)$.

We follow a similar procedure to construct small covering arrays, thereby matching the upper bound on $d(t, v)$ given by Colbourn et al. [9]. As we now have to cover all $t$-sets of columns, we shall require more copies of the optimal array, and accordingly set $r = \left\lceil \frac{(t-1) \log_2 k + \log_2 (2v)}{\log_2 \frac{v}{1-\varepsilon v}} \right\rceil$. Furthermore, rather than using a first-moment calculation as above, we will appeal to the version of the Lovász Local Lemma given by Spencer [29].

Theorem 4.1 (Lovász Local Lemma). Let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_m$ be events in some probability space. Suppose there are $p \in [0, 1]$ and $d \in \mathbb{N}$ such that for each $i \in [m]$, $\mathbb{P}(\mathcal{E}_i) \leq p$, and the event $\mathcal{E}_i$ is mutually independent of a set of all but at most $d$ of the other events. If $ep(d+1) \leq 1$, then $\mathbb{P}\left(\bigcap_{i=1}^m \mathcal{E}_i^c\right) > 0$. 

14
Our events are the same $\mathcal{E}_Q$ defined above, for which we have $\mathbb{P}(\mathcal{E}_Q) \leq (1 - c_{t,v})^r$. By our new choice of $r$, this implies we can set $p = (2ek^{t-1})^{-1}$ in the Lovász Local Lemma.

We next need to determine the parameter $d$. Our construction of the array $A$ ensures that the different columns are independent of one another. In particular, this implies that the event $\mathcal{E}_Q$ is mutually independent of the set of all events that depend on a disjoint set of columns, i.e. $\{\mathcal{E}_{Q'} : Q' \cap Q = \emptyset\}$. For a fixed set $Q$, if another set $Q''$ intersects $Q$, it contains one of the $t$ columns of $Q$, and then there are fewer than $\binom{k}{t-1}$ choices for the other columns of $Q''$. Hence we may take $d < t\binom{k}{t-1}$.

We thus have
\[
ep(d + 1) \leq \frac{t}{2k^{t-1}} \left( \frac{k}{t-1} \right) \leq \frac{t}{2(t-1)!} \leq 1.
\]

By Theorem 4.1 with positive probability none of the events $\mathcal{E}_Q$ occur, which implies the existence of some such $rv^t \times k$ array $A$ that covers all of its $\binom{k}{t}$ $t$-subsets of columns. Recalling that $1 - c_{t,v} \leq \frac{v+1}{v^2}$, this shows
\[
d(t,v) \leq \limsup_{k \to \infty} \frac{rv^t}{\log_2 k} = \frac{(t - 1)v^t}{\log_2 \frac{1}{1 - c_{t,v}}} \leq \frac{(t - 1)v^t}{2\log_2 v - \log_2(v + 1)}.
\]

Finally, we turn to the case when $v$ is not a prime power. Here we use the trivial observation that for any $v \leq v'$, $\ACAN(t, k, v, \varepsilon) \leq \ACAN(t, k, v', \varepsilon)$ and $d(t,v) \leq d(t,v')$, since projecting an array from a large set of symbols surjectively onto a smaller set cannot cause any subset of columns to become uncovered.

Given $v$, let $q$ be the smallest prime power that is at least $v$. Baker, Harman and Pintz [2] proved that, provided $v$ is sufficiently large, $v \leq q \leq v + v^{0.526}$. We thus have
\[
\ACAN(t,k,v,\varepsilon) \leq \ACAN(t,k,q,\varepsilon) \leq q^t \left[ \frac{\ln \frac{1}{\varepsilon}}{2\ln q - \ln(q + 1)} \right] \leq (1 + v^{-0.474})^t v^t \left[ \frac{\ln \frac{1}{\varepsilon}}{2\ln v - \ln(v + 1)} \right] \leq e^{tv^{-0.474}} v^t \left[ \frac{\ln \frac{1}{\varepsilon}}{2\ln v - \ln(v + 1)} \right]
\]
and, similarly,
\[
d(t,v) \leq d(t,q) \leq \frac{(t - 1)q^t}{2\log_2 q - \log_2(q + 1)} \leq \frac{(t - 1)e^{tv^{-0.474}} v^t}{2\log_2 v - \log_2(v + 1)}.
\]
the algebra with linear algebra; they first randomly constructed a covering perfect hash family — an array of vectors in $\mathbb{F}_v^t$, such that for every $t$-set of columns, there is a row where the corresponding vectors are linearly independent. They then extended this into a covering array by replacing each row of the covering perfect hash family with the $v^t$ rows obtained by taking the inner products with all vectors in $\mathbb{F}_v^t$.

Thus we not only rediscover their bound, but their construction as well. The fact that these two approaches have converged upon the same construction suggests that the goals of covering all $t$-sets of columns and having each such $t$-set witness all $v^t$ possible rows are perhaps in some sense independent of one another.

Our upper bound in Section 2 shows that the structured part of the construction, the linear algebraic $v^t \times k$ array, is the best array of its size of which one can take random copies. To further improve the upper bound, therefore, one could either find better ways to combine these copies, or instead start with a larger building block. We discuss these options further in the next section.

5 Concluding remarks

In this paper we studied the maximum coverage function $\text{cov}_{\max}(N; t, k, v)$. We showed that at the lower threshold $N = v^t$ (the minimum size of an array that permits covering a single $t$-set) one can already cover a large proportion of all $t$-sets. More precisely, for prime power $v$, we determined the value of $\text{cov}_{\max}(v^t; t, k, v)$ exactly when $\frac{v^t - 1}{v - 1}$ divides $k$, and obtained asymptotic results otherwise. Combining random copies of these small arrays, we were able to construct small covering and almost covering arrays. We close with some final remarks and possible directions for further research.

5.1 Explicit constructions of covering arrays

We obtained our upper bound on $d(t, v)$ via a random construction and an appeal to the Lovász Local Lemma, making our result an existential one. As explained by Colbourn et al. [6], one could apply the algorithmic version of the Local Lemma, due to Moser and Tardos [18], resulting in an efficient Las Vegas algorithm to produce small covering arrays. For a simpler analysis, one could take a slightly larger number of random copies of $A_{\text{opt}}$, which would then give a covering array with high probability, resulting in an efficient Monte Carlo algorithm instead.

In light of the many applications of covering arrays, there is much interest in finding explicit constructions that can be used in practice. In the case $v = 2$, there has been a series of papers aiming to provide explicit constructions of covering arrays that, while larger than the random arrays of size $O(t^2 \log_2 k)$ given by Kleitman and Spencer [14], can be efficiently constructed. The first such construction was due to Alon [1], whose arrays had $2^{O(t^4 \log_2 k)}$ rows. These explicit constructions have incrementally grown smaller; Naor, Schulman and Srinivasan [20] provided explicit arrays of size $t^{O(\log_2 t)} 2^t \log_2 k$, very close to the random constructions.
For larger values of $v$, Colbourn et al. [6] show how one can use the method of conditional expectation to derandomise their construction, thereby giving a deterministic algorithm for producing covering arrays. However, this can become computationally expensive as the parameters grow. As suggested in [6], given the inherent symmetry of the linear algebraic array $A_{opt}$, one might expect to find an explicit algebraic or geometric construction through which one could concatenate copies of this array to form a small covering array. Not only would these provide efficient deterministic covering arrays, they might even improve the upper bounds on $d(t, v)$.

5.2 Evolution of $\text{cov}_{\text{max}}(N; t, k, v)$

Another way to improve the construction in Section 4 would be to replace $A_{opt}$, and instead concatenate random copies of some other array. The upper bound proven in Section 2 shows that we cannot hope to do better with arrays of size $v^t$, but it might be beneficial to consider initial arrays with a larger number of rows. With this in mind, it would be of interest to determine how the function $\text{cov}_{\text{max}}(N; t, k, v)$ grows as $N$ increases.

On a much finer scale, what happens for $N = v^t + s$ for small values of $s$? In $A_{opt}$, if a $t$-subset $Q$ of columns is not covered, then the vectors $\vec{z}_i \in \mathbb{F}_{v^t}$ that the columns are mapped to form a matrix of rank at most $t - 1$, and hence at most $v^{t-1}$ rows appear in the subarray $(A_{opt})_Q$. This implies that to increase the number of covered $t$-subsets, we need to add at least $v^t - v^{t-1}$ new rows, almost doubling the size of $A_{opt}$.

Of course, there could be other arrays of size $v^t + s$ that do not contain $A_{opt}$ as a subarray, but cover a larger number of $t$-sets. Our proof of the optimality of $A_{opt}$ when $N = v^t$ relied heavily on the fact that if a $t$-set is covered, then each sequence in $S^t$ appears exactly once as a row in the corresponding subarray, leading to useful linear algebraic interpretations of coverage. This rigid structure is lost for larger arrays, rendering the analysis more difficult. Still, we feel that without this structure, one should not be able to cover a larger number of $t$-sets. To make this intuition more precise, we offer the following question.

Question 5.1. When $v$ is a prime power and \(\frac{v^t - 1}{v - 1}\) divides $k$, are $\text{cov}_{\text{max}}(v^t + 1; t, k, v)$ and $\text{cov}_{\text{max}}(v^t; t, k, v)$ equal? What is the smallest $s$ for which $\text{cov}_{\text{max}}(v^t + s; t, k, v) > \text{cov}_{\text{max}}(v^t; t, k, v)$?

We note that a similar phenomenon was observed by Stevens, Moura and Mendelsohn [30] when studying covering arrays of strength two. Any such covering array must have at least $v^2$ rows, and when $k \leq v + 1$, there are constructions of covering arrays with this minimum number of rows. Stevens et al. showed that increasing the number of rows by one or two does not allow for covering arrays with more columns; that is, one needs at least $v^2 + 3$ rows to cover all pairs of $v + 2$ columns.

In Question 5.1 we do not have the restriction of needing to cover all $t$-sets of columns, but we would still be surprised if the addition of an extra row allows more $t$-sets to be covered than by $A_{opt}$. 

17
5.3 Non-uniform coverage

Finally, when dealing with covering arrays, we have only focussed on the number of subsets of some fixed size $t$ that are covered. However, one could instead consider all covered sets, and look to maximise $|\text{Cov}(A)|$ instead.

In particular, when $v = 2$, similar questions have been studied in the context of set shattering. Given a family $\mathcal{F}$ of subsets of a ground set $X$, we say that $S \subseteq X$ is shattered when every possible intersection with $S$ is realised by $\mathcal{F}$; that is, $\{ F \cap S : F \in \mathcal{F} \} = 2^S$. This is a reformulation of the notion of covering: given an array $A \in \{0,1\}^{N \times k}$, we can build a set family $\mathcal{F}_A$ over the ground set $[k]$ of size $N$, by mapping the $i$th row of $A$ to the set of column indices $\{ j \in [k] : A_{i,j} = 1 \}$. A set $Q \subseteq [k]$ of columns is then covered by $A$ if and only if $Q$ is shattered by $\mathcal{F}_A$.

The famous Sauer–Shelah inequality states that a set family $\mathcal{F}$ must shatter at least $|\mathcal{F}|$ sets in total. Given its numerous applications, a pressing open problem is the classification of all families that attain this bound with equality. For details on this line of research, see, for example, [3, 17, 25].

For our problem, we instead ask which families of a given size maximise the number of shattered sets. Note that a family of size $m$ can shatter sets of size at most $\lfloor \log_2 m \rfloor$, and, if $\log_2 m$ is small compared to the size $k$ of the ground set, then almost all such sets will have size precisely $\lfloor \log_2 m \rfloor$. If $m = 2^t$, our construction maximises the number of $t$-sets shattered, and hence one might expect it also maximises the total number of shattered sets.

**Question 5.2.** Given $0 \leq m \leq 2^k$, which set family $\mathcal{F} \subseteq 2^{[k]}$ of size $|\mathcal{F}| = m$ maximises the number of shattered sets?

**Acknowledgements**  The authors would like to express their gratitude towards Noga Alon, Charlie Colbourn, Gal Kronenberg and Benny Sudakov for helpful discussions at various stages of this project. Special thanks go to Ugo Vaccaro for bringing [27] to our attention, prompting the extension of our initial results on set shattering to general covering arrays. We are also grateful to the anonymous referee for several beneficial suggestions and references.

**References**


19


