d-Galvin families

Johan Håstad[*]
School of Engineering Sciences
KTH Royal Institute of Technology
Stockholm, Sweden
johanh@kth.se

Guillaume Lagarde[*]
LaBRI
Université de Bordeaux
Bordeaux, France
guillaume.lagarde@labri.fr

Joseph Swernofsky[*]
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden
josephsw@kth.se

Submitted: Jan 6, 2019; Accepted: Jan 25, 2020; Published: Feb 7, 2020

Abstract

The Galvin problem asks for the minimum size of a family $\mathcal{F} \subseteq \binom{[n]}{n/2}$ with the property that, for any set A of size $\frac{n}{2}$, there is a set $S \in \mathcal{F}$ which is balanced on A, meaning that $|S \cap A| = |S \cap \overline{A}|$. We consider a generalization of this question that comes from a possible approach in complexity theory. In the generalization the required property is, for any A, to be able to find d sets from a family $\mathcal{F} \subseteq \binom{[n]}{n/d}$ that form a partition of $[n]$ and such that each part is balanced on A. We construct such families of size polynomial in the parameters n and d.

Mathematics Subject Classifications: 05D05, 05D40

1 Introduction

1.1 Galvin problem

The starting point of this paper is a question raised by Galvin in extremal combinatorics. Given two sets A and S, we say that S is ***balanced on A*** if $|S \cap A| = \frac{|S|}{2}$.

Definition 1 (Galvin family)**.** If $4 \mid n$, a family $\mathcal{F} \subseteq \binom{[n]}{n/2}$ is said to be ***Galvin*** if for any $A \in \binom{[n]}{n/2}$ there exists a set $S \in \mathcal{F}$ which is balanced on A (i.e., $|S \cap A| = \frac{n}{4}$).

[*] Supported by *Knut and Alice Wallenberg Foundation*

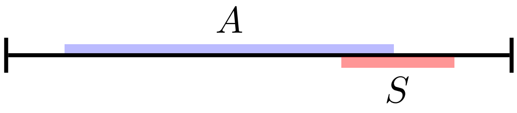

Figure 1: S balanced on A

The ***Galvin problem*** asks for the minimal size, denoted by $m(n)$, of a Galvin family. An upper bound of $m(n) \leqslant \frac{n}{2}$ follows from the family given by the sets $S_i = \{i, i+1, \ldots, i+\frac{n}{2}-1\}$ for $i \in [n/2]$. Lower bounds for the size of Galvin families are more subtle. An easy counting argument shows that $m(n) \geqslant \frac{\binom{n}{n/2}}{\binom{n/2}{n/4}^2} = \Theta(\sqrt{n})$, which is far from $n/2$. Frankl and Rödl [4] established that $m(n) \geqslant \epsilon n$ for some $\epsilon > 0$ whenever $\frac{n}{4}$ is odd, as a corollary to a strong result in extremal set theory. This linear bound was later strengthened by Enomoto, Frankl, Ito and Nomura [3] to $m(n) = n/2$, with the same parity constraint, thus showing the optimality of the construction in this special case. Later, using Gröbner basis methods and linear algebra, Hegedűs [5] obtained that $m(n) \geqslant \frac{n}{4}$ whenever $\frac{n}{4} > 3$ is a prime.

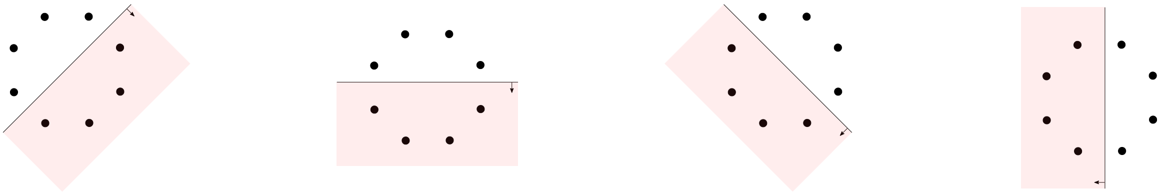

Figure 2: A Galvin family for $n = 8$ consisting of 4 sets

1.2 Generalizations and related works

Surprisingly, problems closely related to the one of Galvin proved useful in arithmetic complexity theory, in order to give lower bounds on the size of arithmetic circuits computing some target polynomials. This connection was first noticed by Jansen [8], and was recently successfully used in a paper by Alon et al. [2]. There the elements of the Galvin family $\mathcal{F}$ are allowed to be sets of size between 2τ and $n-2\tau$ (τ being an integer). Furthermore, for a given $A \in \binom{[n]}{n/2}$ instead of asking for the existence of a set $S \in \mathcal{F}$ perfectly balanced on A the authors look for a set S which is nearly balanced, i.e., $\left||S \cap A| - \frac{|S|}{2}\right| < \tau$ for the same τ. For this setting, Alon, Kumar and Volk [2] showed, using the so-called polynomial method, that $m(n) \geqslant \Omega(n/\tau)$.

Alon, Bergmann, Coppersmith, and Odlyzko [1] investigate a problem dealing with $\{-1, +1\}$ vectors which looks similar to the Galvin one. When rephrasing it as an extremal problem over sets, it reads as follows: what is the minimal number $K(n, c)$ on the size of a family $\mathcal{F} \subseteq \mathcal{P}([n])$ such that the following holds

$$\forall A \subseteq [n], \exists S \in \mathcal{F}, \left||\overline{A} \triangle S| - |A \triangle S|\right| \leqslant c,$$

where $\triangle$ denotes the symmetric difference. Setting $c = 0$ and asking all sets to be of size $n/2$ is exactly the Galvin problem.

We consider here a different type of generalization. Asking for a set $S \in \mathcal{F}$ to be balanced on $A \in \binom{[n]}{n/2}$ is equivalent (up to a factor 2 in the family size) to ask for a partition of $[n]$ into two parts, namely $(S, \overline{S})$, such that each part is balanced on A and such that S, $\overline{S}$ are elements of $\mathcal{F}$. Instead of splitting $[n]$ into two parts, we look for partitions that involve more sets. Introducing a parameter $d \in \mathbb{N}$, we want, for a given A, to be able to find d sets in $\mathcal{F}$ that form a partition of $[n]$ and such that each set is balanced on A.

The original motivation for considering this generalization stems from arithmetic circuits. There, an open question is to know whether there is a separation between two models of computation called multilinear algebraic branching programs (ml-ABPs) and multilinear circuits (ml-circuits). By "separation", we mean that there is some specific polynomial f that can be computed by a small ml-circuit but any ml-ABP for f must be of size superpolynomial in the degree and the number of variables of f. Proving that any generalized Galvin families (i.e., with d parts in the partitions – see below for a formal definition) must be of superpolynomial size (in n the size of the ground set, and d the number of parts) would imply a separation between ml-ABPs and ml-circuits. Since our main result is to prove that generalized Galvin families of polynomial size exist, this approach is unfortunately not promising. Note that it is still possible that ml-ABPs and ml-circuits can be separated, and even that the proof will involve showing that ml-ABPs cannot compute efficiently so-called "full rank polynomials". We only rule out a specific approach to showing that ml-ABPs cannot efficiently compute full rank polynomials. However, we believe that the construction is of intrinsic combinatorial interest. We present briefly at the end of the paper how bounds on d-Galvin families relate to separating ml-ABPs from ml-circuits.

2 d-Galvin families

2.1 Definition

We start with the formal definition of generalized Galvin families.

Definition 2 (***d***-Galvin families)**.** Given two integers $d, n \in \mathbb{N}$ such that $2d \mid n$, we say that a family $\mathcal{F} \subseteq \binom{[n]}{\frac{n}{d}}$ is ***d-Galvin*** if for any $A \in \binom{[n]}{n/2}$, ***A is handled by $\mathcal{F}$***, meaning that there exist d sets $S_1, \ldots, S_d \in \mathcal{F}$ such that:

- The S_i form a partition of $[n]$,
- Each S_i is balanced on A (i.e., $|S_i \cap A| = \frac{n}{2d}$).

Remark 3. Note that a 2-Galvin family is simply a Galvin family (up to adding the complements of any set in the family).

Somewhat surprisingly, small d-Galvin families exist.

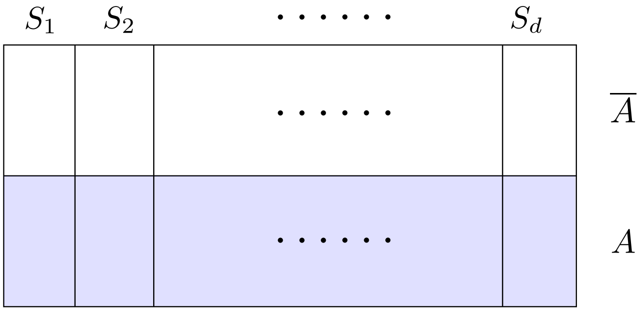

Figure 3: Set A handled by a partition $S_1, S_2, \dots S_d$

Theorem 4. *For any $d, n \in \mathbb{N}$ such that $2d \mid n$, there exists a d-Galvin family of size $\tilde{\Theta}(n^2 d^9)$.*

Here $\tilde{\Theta}(f(n,d))$ is some function g such that

$$f(n,d)(\ln f(n,d))^{c_1} \leqslant g(n,d) \leqslant f(n,d)(\ln f(n,d))^{c_2}$$

for some integers c_1, c_2. The next section is devoted to the construction of a d-Galvin family, yielding a proof of the main theorem.

2.2 Proof of Theorem 4

For technical reasons, we need to distinguish two cases in the proof of Theorem 4: we start by giving a construction when d is reasonably small, then we show how to adapt it to handle larger d.

First case: $d < \frac{n}{(\ln n)^3}$

The overall idea is to construct a family $\mathcal{F}$ of size $\tilde{\Theta}(nd^9)$ such that a random set $A \in \binom{[n]}{n/2}$ is handled by $\mathcal{F}$ with probability at least $1/2$. Taking the random family $\mathcal{G}$ which is the union of n independent such $\mathcal{F}$ increases this probability to at least $1 - 2^{-n}$. By the union bound, the probability that $\mathcal{G}$ handles all sets A is non-zero, yielding the existence of the desired family. We now focus on the construction of such a family $\mathcal{F}$.

Construction of $\mathcal{F}$

For a set X, we use the notation $A \sim X$ to denote that A is a set chosen uniformly at random from X. We let $k := \frac{n}{2d}$ for the rest of the paper.

Lemma 5. *When $d < \frac{n}{(\ln n)^3}$, there is a family $\mathcal{F} \subseteq \binom{[n]}{2k}$ of size $\tilde{\Theta}(nd^9)$ such that*

$$\Pr_{A \sim \binom{[n]}{n/2}} (A \textit{ is handled by } \mathcal{F}) \geqslant 1/2$$

Before going into the construction, let us see how we can prove the main theorem with Lemma 5 in hand.

Proof of Theorem 4, first case. Let $\sigma_1, \dots, \sigma_n$ be n permutations of $[n]$, chosen uniformly at random. For each of these, construct the family $\mathcal{F}_{\sigma_i} = \sigma_i(\mathcal{F})$, i.e., the family from Lemma 5 where any element $e \in [n]$ has been replaced by $\sigma_i(e)$. Consider the family $\mathcal{G} := \cup_{i\in[n]}\mathcal{F}_{\sigma_i}$. We aim to prove that $\mathcal{G}$ is d-Galvin with non-zero probability. Given a set A, let $\mathrm{H_i}$ be the event: "A is handled by $\mathcal{F}_{\sigma_i}$". $\mathrm{H_i}$ is equivalent to "$\sigma_i^{-1}(A)$ is handled by $\mathcal{F}$". As $\sigma_i^{-1}(A)$ is a uniformly random set independent from $\sigma_{i'}^{-1}(A)$ for $i \neq i'$, this proves the independence between the events $\mathrm{H_i}$. From this we conclude

$$\Pr_{A\sim\binom{[n]}{n/2}}\left(\forall i \in [n], A \text{ is not handled by } \mathcal{F}_{\sigma_i}\right) \leqslant 2^{-n}$$

Thus, by the union bound there is a non-zero probability that $\mathcal{G}$ handles all sets A, concluding the proof of the theorem. □

The rest of the section consists of a proof of Lemma 5. The overall strategy is to divide the elements of $[n]$ into buckets, denoted by χ_i, and build the sets S from any pair of buckets (χ_i, χ_j). Suppose the amount by which these buckets are unbalanced on A are R_i and R_j respectively. If half the elements of S are chosen from bucket χ_i and half from bucket χ_j then the amount by which S is unbalanced on A will be close to a normal distribution with expectation depending on R_i and R_j. By showing a good upper bound on the R_i, the probability that S is balanced is reasonably large, and picking only polynomially many random sets S is sufficient. In fact, we must be slightly more careful because the bucket errors accumulate as we pick many sets S. Fortunately, we can manage this by taking an ordering π of the buckets such that the error of $\cup_{j\leqslant i}\chi_{\pi(j)}$ stays small for all i.

Proof of Lemma 5. First, we divide $[n]$ into several intervals (recall that $k = \frac{n}{2d}$).

- $\chi_0 = (0, k]$,
- $\chi_i = ((2i-1)k, (2i+1)k]$ for $i \in [d-1]$,
- $\chi_d = ((2d-1)k, n]$.

For $i \in [d-1]$, let T_i be a random variable obtained by sampling uniformly at random from $\binom{\chi_i}{k}$. We create sets $G_i = \{T_i^h : h \in [1, r]\}$ by sampling independently $r = \tilde{\Theta}(n^{1/2}d^{7/2})$ subsets $T_i^h \sim \binom{\chi_i}{k}$ and adding them to G_i. For technical reasons, we let G_0 to be the singleton $\{\varnothing\}$ and $G_d = \{\chi_d\}$. Finally let $\mathcal{F} = \{\overline{T_i^h} \cup T_j^l : i, j \in [0, d], T_i^h \in G_i, T_j^l \in G_j\}$, where $\overline{T_i^h}$ denotes $\chi_i \setminus T_i^h$ (similarly, $\overline{T_i}$ denotes the random variable $\chi_i \setminus T_i$). Now, we claim that such a random $\mathcal{F}$ handles $A \sim \binom{[n]}{n/2}$ with probability at least $1/2$, giving the existence of the desired family. As there are $\Theta(d^2)$ pairs (i, j) to consider and for each one we add $\tilde{\Theta}((n^{1/2}d^{7/2})^2)$ sets S to $\mathcal{F}$, this gives a total size $|\mathcal{F}| = \tilde{\Theta}(nd^9)$.

For $I \subseteq [0, d]$ we introduce an ***error term $R(I)$*** to represent the error in balancing A. We let $\chi(I) = \cup_{i\in I}\chi_i$ and $R(I) = |A \cap \chi(I)| - \frac{|\chi(I)|}{2}$. Furthermore we write $R_i := R(\{i\})$. For reasons that will become clear later, we want to choose a permutation π of $[0, d]$ with $\pi(0) = 0$ and $\pi(d) = d$ with $\max_{i\in[0,d]} |R(\pi([0, i]))|$ small.

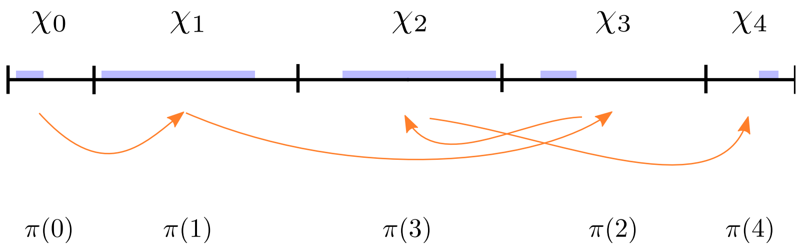

Figure 4: An ordering π

Claim 6. $\exists \pi : \max_{i\in[0,d]} |R(\pi([0,i]))| \leqslant \max_{i\in[0,d]} |R_i|$

Proof. We let $\pi(0)$ be fixed to be 0, and for each $i \geqslant 0$, pick $\pi(i+1)$ among the remaining elements such that $R_{\pi(i+1)}$ has opposite sign from $R(\pi[0,i])$. If $R(\pi[0,i]) = 0$ pick any value of $\pi(i+1)$. Note that this is always possible as $R([0,d]) = 0$. □

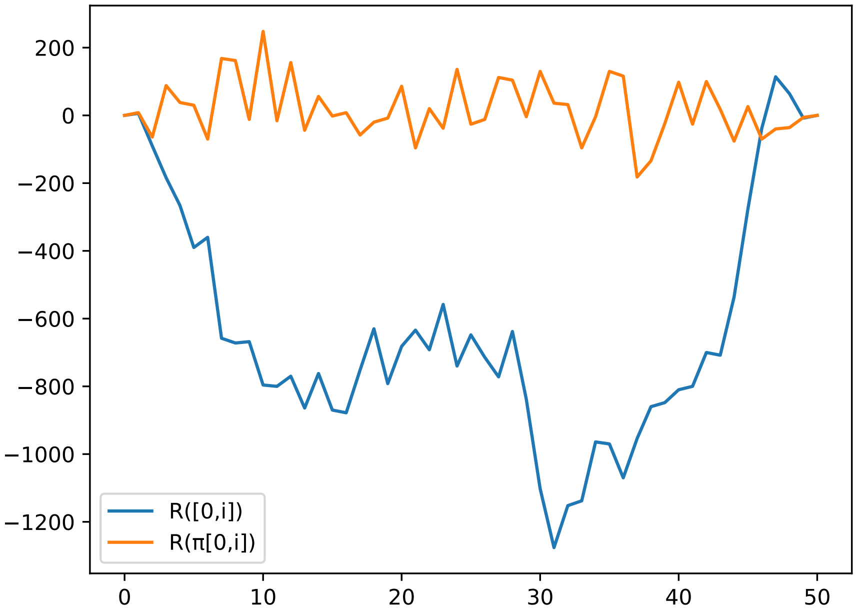

Figure 5: Cumulative R value shrinks under a good ordering. $n = 10^6$ and $d = 50$

We fix π to be a permutation that fulfills Claim 6 for the rest of the paper. Recall that the ***hypergeometric distribution*** of parameters K, N, n, written $\boldsymbol{H(K,N,n)}$, describes the following process: in a population of N balls among which exactly K are red, n balls are drawn without replacement; H is the number of red balls obtained during the process. More precisely:

Definition 7 (hypergeometric distribution)**.** A random variable X follows the hypergeometric distribution $H(K,N,n)$ if its probability mass function is given by

$$\Pr(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Claim 8. *With probability at least* $\frac{3}{4}$ *over a random* $A\sim\binom{[n]}{n/2}$, *we have* $\max_{i\in[0,d]}|R_i|\leqslant\sqrt{\ln(13d)}\sqrt{k}$.

Proof. For $i\in[1,d{-}1]$, each element R_i follows a hypergeometric distribution $H(\frac{n}{2},n,2k)$. We get the following bound, due to Hoeffding [6].

$$P(|R_i|>x)\leqslant 2\exp(-\frac{2x^2}{2k})$$

With $x=\sqrt{\ln(13d)}\sqrt{k}$ this becomes $2\exp(-\ln(13d))=\frac{2}{13}\cdot\frac{1}{d}$. R_0 and R_d follow the distribution $H(\frac{n}{2},n,k)$, which yields an even stronger bound for $i=0$ and $i=d$. Applying a union bound over all $i\in[d]$, the probability that at least one $|R_i|$ exceeds $\sqrt{\ln(13d)}\sqrt{k}$ is bounded by $\frac{2}{13}\frac{d+1}{d}<\frac{1}{4}$ (since $d\geqslant 2$). □

Claim 9. *Suppose* $d<\frac{n}{(\ln n)^3}$. *Let* $S_j:=\overline{T}_{\pi(j-1)}\cup T_{\pi(j)}$ *for* $j\in[d]$. *If* $\{S_j\}_{j<i}$ *are balanced on* A *then we have* S_i *balanced on* A *with probability at least*

$$\Theta\left(\exp(-\frac{4}{k}\max\{R(\pi[0,i-1])^2,R_{\pi(i)}^2\})\sqrt{\frac{1}{k}}\right)$$

Proof. Let $t:=-R(\pi[0,i-1])$. Since the $\{S_j\}_{j<i}$ are balanced, we have:

$$|A\cap\cup_{j=1}^{i-1}S_j|=(i-1)k \tag{1}$$

On the other hand:

$$\begin{aligned}|A\cap\chi(\pi[0,i-1])|&=|A\cap\cup_{j=1}^{i-1}S_j|+|A\cap\overline{T}_{\pi(i-1)}|\\&=(i-1)k+|A\cap\overline{T}_{\pi(i-1)}| &&\text{using (1)}\end{aligned}$$

and

$$|A\cap\chi(\pi[0,i-1])|=(2i-1)\frac{k}{2}-t \qquad\qquad \text{by definition of } R(\cdot)$$

Therefore, $|A\cap\overline{T}_{\pi(i-1)}|=\frac{k}{2}-t$. To make S_i to be balanced we must have $|A\cap T_{\pi(i)}|+|A\cap\overline{T}_{\pi(i-1)}|=k$. This means that the probability that S_i is balanced is the probability that $|A\cap T_{\pi(i)}|=\frac{k}{2}+t$. Let $x:=|A\cap T_{\pi(i)}|$ and $R:=R_{\pi(i)}$. We have that x follows a hypergeometric distribution with parameters $H(k+R,2k,k)$. Claim 11 below suffices to establish Claim 9. □

We state an easy lemma that will be helpful for Claim 11 to estimate binomial coefficients, a proof of which can be found in Spencer and Florescu [9].

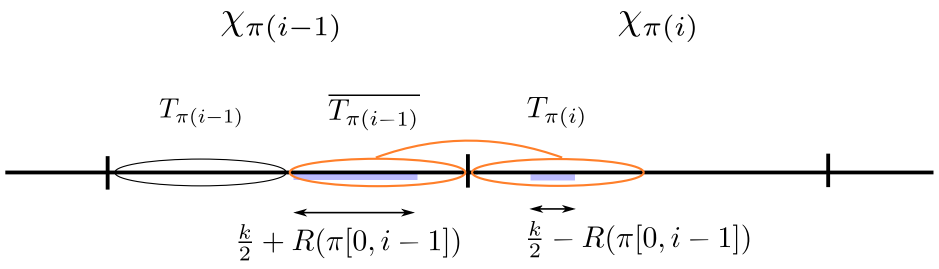
$\chi_{\pi(i-1)}$
$\chi_{\pi(i)}$
$T_{\pi(i-1)}$
$\overline{T_{\pi(i-1)}}$
$T_{\pi(i)}$
$\frac{k}{2} + R(\pi[0, i-1])$
$\frac{k}{2} - R(\pi[0, i-1])$

Figure 6: Conditions on $|A \cap T_{\pi(i)}|$

Lemma 10. $\binom{n}{\frac{n}{2}-m} = 2^n\sqrt{\frac{2}{n\pi}}\exp\left(-\frac{2m^2}{n}\right)\left(1 + O(\frac{m^3}{n^2})\right)$

Claim 11. *We have that* $x = \frac{k}{2} + t$ *with probability at least*

$$\Theta\left(\exp(-\frac{4}{k}\max\{t^2, \frac{R^2}{4}\})\sqrt{\frac{1}{k}}\right)$$

Proof. As x follows a hypergeometric distribution with parameters $H(k+R, 2k, k)$, we have that

$$P(x = \frac{k}{2} + t) = \binom{k+R}{\frac{k}{2}+t}\binom{k-R}{\frac{k}{2}-t}\binom{2k}{k}^{-1}. \tag{2}$$

By Claim 6 and Claim 8 we have $0 \leqslant t, R \leqslant \sqrt{\ln(13d)}\sqrt{k}$. Recall also that $k = \frac{n}{2d}$. Therefore, as long as $d < \frac{n}{(\ln n)^3}$, we are in the regime $(\frac{R}{2} - t)^3 = o(k^2)$; we can apply Lemma 10 and Equation (2) becomes

$$\begin{aligned}
=&2^{k+R}\sqrt{\frac{2}{(k+R)\pi}}\exp\left(-\frac{2(\frac{R}{2}-t)^2}{k+R}\right)\\
&\times 2^{k-R}\sqrt{\frac{2}{(k-R)\pi}}\exp\left(-\frac{2(\frac{R}{2}-t)^2}{k-R}\right)\\
&\times\left(2^{2k}\sqrt{\frac{2}{2k\pi}}\right)^{-1}(1+o(1))\\
=&\sqrt{\frac{4k}{(k+R)(k-R)\pi}}\exp\left(-2(\frac{R}{2}-t)^2(\frac{1}{k+R}+\frac{1}{k-R})\right)(1+o(1))\\
=&\sqrt{\frac{4k}{(k^2-R^2)\pi}}\exp\left(\frac{-4k(\frac{R}{2}-t)^2}{k^2-R^2}\right)(1+o(1))
\end{aligned}$$

Since $0 \leqslant t, R \leqslant \sqrt{\ln(13d)}\sqrt{k} = o(k)$, we finally get

$$=\sqrt{\frac{4}{k\pi}}\exp\left(-\frac{4}{k}(\frac{R}{2}-t)^2\right)(1+o(1))$$ □

Combining Claim 8 and Claim 9, we have a probability of

$$\Theta\left(\exp(-\frac{4}{k}(k\ln(13d)))\sqrt{\frac{1}{k}}\right) = \Theta\left(\exp(-4\ln(13d))\sqrt{\frac{d}{n}}\right)$$
$$= \Theta((13d)^{-7/2}n^{-1/2})$$

that S_i is balanced. Call this probability y. If $|G_i| = \frac{\ln(4d)}{y}$ then the probability that some choice of $T_{\pi(i)}$ balances S_i is at least $1 - \frac{1}{4d}$. By the union bound, the chance that $|R_i|$ is not bounded in Claim 8 or that any S_i is unbalanced is at most $\frac{1}{4} + d\frac{1}{4d} = \frac{1}{2}$. Hence the probability that we get a d-Galvin partition is at least $\frac{1}{2}$, as desired. □

In the above proof we used $d < \frac{n}{(\ln n)^3}$ to apply Lemma 10. While this could perhaps be improved to $d = \frac{n}{\ln n}$, there is a real barrier here. When d is this large we expect some buckets to be entirely empty of elements from A and the above proof does not work. We now handle the case where d is larger.

Second case: $d \geqslant \frac{n}{(\ln n)^3}$

Proof of Theorem 4, second case. First, observe that Galvin families compose nicely; if $\mathcal{F}$ is an a-Galvin family over $[n]$, and if we take a b-Galvin family $\mathcal{F}_S$ over S for each set $S \in \mathcal{F}$, then the union of all $\mathcal{F}_S$ forms an ab-Galvin family.

Set $d' = \frac{n}{(\ln n)^3}$ and assume for the moment that d' and $\frac{d}{d'}$ are valid factors of d. The idea is to start by constructing a d'-Galvin family $\mathcal{F}$ over $[n]$, using the previous construction. We then recursively apply the construction to get a $\frac{d}{d'}$-Galvin family $\mathcal{F}_S$ for any $S \in \mathcal{F}$, and the final family is the union of all $\mathcal{F}_S$. The elements of $\mathcal{F}$ are sets of size $(\ln n)^3$, therefore the families $\mathcal{F}_S$ are of size $\tilde{\Theta}(1)$, and the overall construction is of size $\tilde{\Theta}(n^2 d^9)$.

In the case that d' and $\frac{d}{d'}$ are not valid factors of d, we do the following. Let $k' = \lfloor \frac{d}{d'} \rfloor$. The idea is to construct a family $\mathcal{F}$ with sets of size $2k'k$, and $2(k'+1)k$, that behaves like a Galvin family: we ask that any set A has a partition of $[n]$ from sets in $\mathcal{F}$, where each set of the partition is balanced on A. We then apply recursively the construction to split the sets of size $2k'k$ and $2(k'+1)k$ until we get size k sets. To create the family $\mathcal{F}$, we adapt the construction of the Galvin family when $d < \frac{n}{(\ln n)^3}$, in the following way. Note that in any partition of $[n]$ into sets of these sizes, the number of sets of size $2k'k$ and $2(k'+1)k$ are fixed (given by d and n). We denote these numbers by f and c. We need to ensure that the $\overline{T_i} \cup T_j$ are of the correct sizes (i.e., $2k'k$ or $2(k'+1)k$). For that, we change the sizes of the χ_i in the following way:

- $|\chi_0| = k'k$
- For c values of $i \in [1, d-1]$, we have $|\chi_i| = 2(k'+1)k$
- For the other $i \in [1, d-1]$ we have $|\chi_i| = 2k'k$
- $|\chi_d| = k'k$.

We then choose the T_i to be of size $k'k$ except for $i = 0$ where the unique T_0 remains $\varnothing$. This gives the desired sizes for $|S_i|$ and it is not hard to see that the proof carries over to this case with some simple and obvious modifications. □

2.3 Galvin family without the divisibility condition

The previous definition of a d-Galvin family requires $2d \mid n$. Here we present a relaxed version, which can be defined without the divisibility condition, and prove that such families of polynomial size can be obtained using our previous construction.

When the divisibility condition does not hold we would like d sets to be exactly or almost exactly balanced on A and for those sets to be as close in size as possible. To be exactly balanced they must have evenly many elements, so if $[n]$ is odd then we must include a set of odd size which is imbalanced by 1 element. Of the remaining elements, the closest they can come in size is differing by 2 elements - being of size either $2\lfloor k \rfloor$ or $2\lceil k \rceil$. We are able to achieve this best possible outcome.

Definition 12 (***d***-Galvin family, second version)**.** Given two integers $d, n \in \mathbb{N}$ with $d \leqslant n$, we say that a family $\mathcal{F} \subseteq 2^{[n]}$ is ***d-Galvin*** if for any $A \in \binom{[n]}{\lceil n/2 \rceil}$, ***A is handled by*** $\boldsymbol{\mathcal{F}}$, meaning that there exist d sets $S_1, \ldots, S_d \in \mathcal{F}$ such that:

1. $\forall i < d$, $|S_i| = 2\lfloor k \rfloor$ or $|S_i| = 2\lceil k \rceil$,
2. $2\lfloor k \rfloor \leqslant |S_d| \leqslant 2\lceil k \rceil$
3. The S_i form a partition of $[n]$,
4. For $i < d$, each S_i is balanced on A.
5. $|\overline{A} \cap S_d| \leqslant |A \cap S_d| \leqslant |\overline{A} \cap S_d| + 1$.

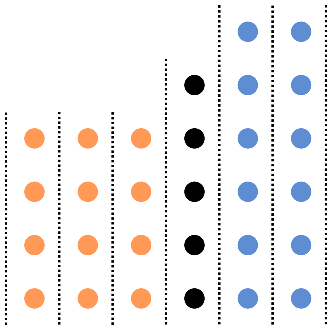

Figure 7: For $n = 29, d = 6$, we have three sets of size $2\lfloor k \rfloor$, two sets of size $2\lceil k \rceil$, and one set of size $\lfloor k \rfloor + \lceil k \rceil$.

Theorem 13. *There exists a d-Galvin family of size polynomial in d and n.*

Sketch of the proof. We modify the previous construction slightly in order to handle this more general setting. This is very similar to the proof of Theorem 4 in the case $d \geqslant \frac{n}{(\ln n)^3}$. Suppose k is not an integer and write $k' := \lfloor k \rfloor$. Furthermore, assume for the moment that $k = \omega((\ln n)^3)$ so that the construction from Claim 9 holds. Note that in any partition of $[n]$ into sets that respect properties (1) and (2) of the definition, the number of sets of size $2k'$, $2k'+1$, and $2(k'+1)$ are fixed (given by d and n). We denote these numbers by f, m and c. We need to ensure that the $\overline{T_i} \cup T_j$ are of the correct size in order to be able to fulfill our definition. For that, we change the size of the χ_i in the following way:

- $|\chi_0| = k'$ if $m = 0$ and $k'+1$ otherwise
- For c values of $i \in [1, d-1]$, we have $|\chi_i| = 2(k'+1)$
- For the other $i \in [1, d-1]$ we have $|\chi_i| = 2k'$
- $|\chi_d| = k'$.

We then choose the T_i to be of size k' except for $i = 0$ where the unique T_0 remains $\varnothing$. By doing so, the partitions from the family respect properties (1) and (2), and again the proof that this gives a valid construction is very close to the original proof and we omit the details.

Finally, if $k = O((\ln n)^3)$ then we may have to simultaneously apply the adjustments above and the ones in the proof of the second case of Theorem 4. □

3 Discussion and open questions

The actual construction is probabilistic and it could be interesting to derandomize it, without increasing too much the size of the family. A way to tackle the problem is to carefully design the sets T_i belonging to G_i instead of taking them randomly.

The given upper bound is nicely polynomial in n and d but it is unlikely to be tight. We suspect that even modifications of the current construction can yield some improvements. In particular, the family $\mathcal{F}$ from Lemma 5 is constructed by taking the union $\overline{T_i} \cup T_j$ over all possible pairs $(T_i, T_j) \in G_i \times G_j$ for $i, j \in [d]$. It might be possible to restrict (i, j) to come from the edges of a sparse graph over the vertices $[d]$, and still prove Claim 6, maybe in some slightly weaker form, possibly saving a factor close to d. Even if this is possible the resulting family is still not likely to be optimal size and hence we have not investigated this approach in detail as it would lead to considerable complications and we prefer a simple construction. A truly optimal construction is likely to require some new ideas.

There is a linear lower bound for the original Galvin problem. This is essentially tight and it would be nice to have a similar tight result for d-Galvin families for $d > 2$. The work of Hrubes et al. [7] enables us to derive an $\frac{nd}{2} - o(nd)$ lower bound by using the following theorem.

Theorem 14 (Theorem 3 from [7]). *Let n be a positive even integer and $S_1, S_2, \ldots S_k$ be proper non-empty subsets of $[n]$ such that for every $X \subset [n]$ of size $n/2$, there is an $i \in [k]$ for which $|S_i \cap X| = |S_i|/2$. Then, $k \geqslant \frac{n}{2} - o(n)$.*

We can now turn to our lower bound for d-Galvin families.

Claim 15. *A d-Galvin family must be of size at least $\frac{nd}{2} - o(nd)$.*

Proof. Consider a d-Galvin family $\mathcal{F}$, some $y \in [n]$, and the sets $S_1 \dots S_k \in \mathcal{F}$ that contain y. By definition, for $X \in \binom{[n]}{n/2}$ there must be one of these S_i such that $X \cap S_i = \frac{n}{2d} = \frac{|S_i|}{2}$. Hence sets $S_1 \dots S_k$ respect the conditions of Theorem 14 and we conclude that $k \geqslant n/2 - o(n)$.

This shows that $B := \{(S, y) | S \in \mathcal{F}, y \in S\}$ has size $|B| \geqslant \frac{n^2}{2} - o(n^2)$. Since each $S \in \mathcal{F}$ has size $|S| = \frac{n}{d}$, this shows $|\mathcal{F}| \geqslant \frac{nd}{2} - o(nd)$. □

Link with arithmetic complexity theory

In this section we present briefly how bounds on d-Galvin families relate to separating ml-ABPs from ml-circuits.

One popular measure in arithmetic complexity is based on the rank of partial derivative matrices. Given a multilinear polynomial f over a set X of n variables and a subset $A \subset X$, we can split each monomial of f into those variables in A and those in $X \setminus A$. We construct a $2^{|A|} \times 2^{n-|A|}$ matrix with respect to A, written $M_A(f)$, where rows (resp. columns) are indexed by multilinear monomials over the variables A (resp. $X \setminus A$). The entry $M_{m_1,m_2}(f)$ corresponding to monomials m_1 and m_2 is the coefficient of $m_1 m_2$ in f. Perhaps surprisingly, there is a polynomial f computable by a polynomial-sized arithmetic circuit where $M_A(f)$ is full-rank for any $A \subset X$ of size $\frac{n}{2}$. Therefore, one strategy to separate ml-circuits from another model of computation, let us say ml-ABPs in our case[1], is to prove that an ml-ABP of polynomial size cannot compute such a full-rank polynomial. The key idea to see the link with d-Galvin is the following: we show that a polynomial f computed by a small ABP can be decomposed as

$$f(X) = \sum_{i \leqslant N} f_1^i f_2^i \dots f_d^i,$$

where N is small and related to the size of the ABP and for any fixed i, the polynomials $f_1^i, \dots, f_d^i$ are over disjoint sets of variables. We write these sets as $X_1^i, \dots X_d^i$, respectively. By subadditivity and since f is full-rank, for any $A \subset X$, there is at least one $i_0 \in [d]$ for which $\mathrm{rank}(M_A(f_1^{i_0} \times \dots \cdots \times f_d^{i_0})) \geqslant \frac{2^{n/2}}{N}$. It can be shown that this can happen only if the sets $X_1^{i_0}, \dots, X_d^{i_0}$ are well balanced on A (i.e., for any $j \in [d], |X_j^{i_0} \cap A| \approx |X_j^{i_0}|/2$). In other words the dN sets of variables $\{X_j^i : i \in [N], j \in [d]\}$ behave like a d-Galvin family. Therefore, a lower bound on d-Galvin families gives a lower bound on dN which implies a lower bound on the size of the ABP.

Acknowledgements

We thank Andrew Morgan for giving helpful suggestions in the two cases of claim 4. We thank Ishay Haviv for suggestions involving a lower bound. The second author would like to thank Hervé Fournier for valuable discussions.

[1] This strategy was successfully used by Ran Raz to separate multilinear formulas and multilinear circuits.

References

[1] Noga Alon, Ernest E Bergmann, Don Coppersmith, and Andrew M Odlyzko. Balancing sets of vectors. *IEEE Transactions on Information Theory*, 34(1):128–130, 1988.

[2] Noga Alon, Mrinal Kumar, and Ben Lee Volk. Unbalancing sets and an almost quadratic lower bound for syntactically multilinear arithmetic circuits. In *33rd Computational Complexity Conference, CCC 2018, June 22-24, 2018, San Diego, CA, USA*, pages 11:1–11:16, 2018.

[3] H. Enomoto, Peter Frankl, N. Ito, and Katsuhiro Nomura. Codes with given distances. *Graphs and Combinatorics*, 3:25–38, 1987.

[4] Peter Frankl and Vojtěch Rödl. Forbidden intersections. *Transactions of the American Mathematical Society*, 300(1):259–286, 1987.

[5] Gábor Hegedűs. Balancing sets of vectors. *Studia Scientiarum Mathematicarum Hungarica*, 47(3):333–349, 2009.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[7] Pavel Hrubes, Sivaramakrishnan Natarajan Ramamoorthy, Anup Rao, and Amir Yehudayoff. Lower bounds on balancing sets and depth-2 threshold circuits. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece.*, pages 72:1–72:14, 2019.

[8] Maurice J Jansen. Lower bounds for syntactically multilinear algebraic branching programs. In *International Symposium on Mathematical Foundations of Computer Science*, pages 407–418. Springer, 2008.

[9] Joel Spencer and Laura Florescu. Asymptopia, volume 71 of student mathematical library. *American Mathematical Society, Providence, RI*, page 66, 2014.