# Query Complexity of Sampling and Small Geometric Partitions

NAVIN GOYAL, LUIS RADEMACHER and SANTOSH VEMPALA

Combinatorics, Probability and Computing / *FirstView* Article / October 2014, pp 1 - 21
DOI: 10.1017/S0963548314000704, Published online: 21 October 2014

**Link to this article:** http://journals.cambridge.org/abstract_S0963548314000704

**How to cite this article:**
NAVIN GOYAL, LUIS RADEMACHER and SANTOSH VEMPALA Query Complexity of Sampling and Small Geometric Partitions. Combinatorics, Probability and Computing, Available on CJO 2014 doi:10.1017/S0963548314000704

**Request Permissions :** Click here

*Combinatorics, Probability and Computing*: page 1 of 21. © Cambridge University Press 2014
doi:10.1017/S0963548314000704

# Query Complexity of Sampling and Small Geometric Partitions

NAVIN GOYAL[1], LUIS RADEMACHER[2]
and SANTOSH VEMPALA[3]

[1] Microsoft Research India, Vigyan, 9 Lavelle Road, Bangalore 560001, India
(e-mail: navingo@microsoft.com)

[2] Computer Science and Engineering, Ohio State University, Dreese Labs 495
2015 Neil Avenue, Columbus, OH 43210, USA
(e-mail: lrademac@cse.ohio-state.edu)

[3] College of Computing, Georgia Institute of Technology,
801 Atlantic Drive, Atlanta, GA 30332, USA
(e-mail: vempala@cc.gatech.edu)

*Received 14 October 2011; revised 13 August 2014*

In this paper we study the following problem.

*Discrete partitioning problem (DPP).* Let $\mathbb{F}_q P^n$ denote the $n$-dimensional finite projective space over $\mathbb{F}_q$. For positive integer $k \leqslant n$, let $\{A^i\}_{i=1}^N$ be a partition of $(\mathbb{F}_q P^n)^k$ such that:

(1) for all $i \leqslant N$, $A^i = \prod_{j=1}^k A^i_j$ (partition into product sets),

(2) for all $i \leqslant N$, there is a $(k-1)$-dimensional subspace $L^i \subseteq \mathbb{F}_q P^n$ such that $A^i \subseteq (L^i)^k$.

What is the minimum value of $N$ as a function of $q, n, k$? We will be mainly interested in the case $k = n$.

DPP arises in an approach that we propose for proving lower bounds for the query complexity of generating random points from convex bodies. It is also related to other partitioning problems in combinatorics and complexity theory. We conjecture an asymptotically optimal partition for DPP and show that it is optimal in two cases: when the dimension is low ($k = n = 2$) and when the factors of the parts are structured, namely factors of a part are close to being a subspace. These structured partitions arise naturally as partitions induced by query algorithms. Our problem does not seem to be directly amenable to previous techniques for partitioning lower bounds such as rank arguments, although rank arguments do lie at the core of our techniques.

2010 *Mathematics subject classification*: Primary 68Q17
Secondary 68R05, 51E20

## 1. Introduction

In this paper we study the following problem.

**Discrete partitioning problem (DPP).** Let $\mathbb{F}_q P^n$ denote the $n$-dimensional finite projective space over $\mathbb{F}_q$ (see Section 2 for a quick introduction to finite projective spaces and some related definitions). For positive integer $k \leqslant n$, let $\{A^i\}_{i=1}^N$ be a partition of $(\mathbb{F}_q P^n)^k$ such that:

(1) for all $i \leqslant N$, $A^i = \prod_{j=1}^k A^i_j$ (partition into product sets),
(2) for all $i \leqslant N$, there is a $(k-1)$-dimensional subspace $L^i \subseteq \mathbb{F}_q P^n$ such that $A^i \subseteq (L^i)^k$.

What is the minimum value of $N$ as a function of $q, n, k$? We will be mainly interested in the case $k = n$.

DPP seems interesting in its own right and several related problems have been studied in the past; we discuss these later. Before stating our results for DPP we discuss another motivation for studying it. DPP arises in our approach for proving lower bounds for the query complexity of random sampling from convex bodies. It is standard in this problem to give the convex body to an algorithm as a membership oracle, that is, a black box that when queried with a point in $\mathbb{R}^n$ answers YES if the point is in the body and answers NO if the point is outside the body (*e.g.*, [13, 16]). Given a convex body $K \in \mathbb{R}^n$ via a membership oracle, by sampling from $K$ we mean generating a random point from $K$ whose distribution is approximately uniform. Sampling is one of the most useful primitives in the algorithmic theory of convex bodies (*e.g.*, [16, 5]). The currently best known algorithm [16] for sampling makes $O(n^4)$ membership oracle queries to generate one random point. Improving this bound will directly improve the complexity of algorithms for volume computation and convex optimization. On the other hand, the best known lower bound is just $\Omega(n)$. Thus, understanding the query complexity of sampling is an important problem. Notice that we are working with oracle algorithms, and so the lower bounds are on the query complexity and not on the computational complexity of sampling.

In this paper, we propose an approach for proving an $\Omega(n^2)$ lower bound on the query complexity of sampling. The approach, discussed in the Appendix, involves proving a lower bound on the number of queries for a problem that we call SPAN: given $n-1$ vectors in $\mathbb{R}^n$ via a natural oracle, find a hyperplane close to all of them. The problem SPAN reduces efficiently to sampling from convex bodies, so that a lower bound for SPAN implies a lower bound for sampling. Randomized oracle algorithms can be interpreted as distributions over deterministic decision trees. As is standard in lower bounds for randomized decision trees, it suffices to prove a lower bound on the size of a partition of the input space induced at the leaves of any small-depth deterministic decision tree with the following property: in most parts of this partition the value of the function being computed is nearly constant. We call the problem of lower-bounding the size of this partition the *continuous partitioning problem* (CPP). While we do not solve CPP, we get insights into it by formulating an analogue of SPAN and its associated partitioning problem over finite fields and proving results in this setting. The rest of the Introduction is devoted to discussion of these discrete problems.

As the continuous problem SPAN only cares about the linear span of the input vectors, it is more conveniently stated not in a vector space but in the corresponding projective space, the space of all lines through the origin. The same can be said about the discrete analogue. Working over projective space makes counting arguments simpler.

**Discrete span problem (DSPAN).** The input consists of $n$ points $v_1, \ldots, v_n \in \mathbb{F}_q P^n$, where $\mathbb{F}_q P^n$ is the finite projective space of dimension $n$ over the finite field $\mathbb{F}_q$. The input can only be accessed via the following oracle. A query $x$ is an $(n-1)$-dimensional flat in $\mathbb{F}_qP^n$; if $x$ contains all the points then the oracle answers YES, else it gives the least index $i$ such that $v_i$ does not lie in $x$. The problem is to find an $(n-1)$-flat (this is an abbreviation for $(n-1)$-dimensional subspace) containing all $v_i$. The discrete SPAN problem is easily solved with $O(qn^2)$ queries using a deterministic algorithm.

We interpret algorithms for such a problem as randomized decision trees, namely a distribution on (deterministic) decision trees. The leaves of such a deterministic decision tree induce a partition of $(\mathbb{F}_qP^n)^n$, and the problem of lower-bounding the size of this partition is the discrete partitioning problem (DPP) stated at the beginning of this paper. The oracle described may seem a bit unnatural at first. It is motivated by the continuous problem and is chosen to be a mild strengthening of the 'membership oracle' (which in this case would just answer whether or not all $v_i$ lie in $x$). A lower bound under the stronger oracle is also a valid lower bound for the weaker membership oracle because the algorithm can always ignore the additional information provided by the stronger oracle. At the same time, the strengthening adds the property that the parts of the induced partition are product sets (see the Appendix for more details).[1] Lower bounds for partitions with product parts seem easier to prove than the general case, and the product property is crucial to our proofs. Each such product is of the form $A_1 \times A_2 \times \cdots \times A_n$, such that there is an $(n-1)$-flat $F \subset \mathbb{F}_qP^n$ with $A_1 \times A_2 \times \cdots \times A_n \subseteq F^n$. Moreover, each $A_i$ is somewhat structured: it can be represented as a flat minus a small number of other flats; thus each $A_i$ is close to a flat. See Lemma 5.1 for a precise statement.

There are a few ways of formally relating DPP and DSPAN that we will sketch now. A simple but weak way is to consider DSPAN and use Yao's minimax principle with cost giving the probability of failure of the randomized algorithm [17, Section 2.2.2], reducing the lower bound problem to proving a lower bound on the expected running time of a deterministic decision tree as in DSPAN (with the input drawn from a probability distribution) that is allowed to err with a small probability. This leads to a variation of DPP with condition (2) relaxed so that each part is not necessarily *fully* contained in the power of a $(k-1)$-flat, but only *mostly* contained in such a flat. In this paper we do not address this harder version of DPP. A stronger connection is given by first observing that the correctness of a solution to DSPAN can be verified efficiently by querying the conjectured solution: the solution is correct if and only if the oracle answers YES. Thus, the worst-case expected[2] running time of the best Las Vegas (*i.e.*, always correct) algorithm is within a constant factor of the best Monte Carlo (*i.e.*, correct with some probability) algorithm [17, Exercise 1.3]. That is, it is enough to prove a lower bound on the complexity of Las Vegas algorithms. The use of Yao's minimax principle with cost

[1] In other words, for the DSPAN problem with membership oracle, parts are not necessarily product sets. When a membership query results in NO, we learn that some input vector $v_i$ is not in the queried hyperplane and the set of tuples consistent with this is not a product: it is actually the complement of a product. After $h$ queries, the part is the intersection of some product sets, resulting from YES queries, minus the union of some other product sets, resulting from NO queries. In particular, it is a product set minus the union of at most $h$ product sets. It is easy to show that any such set can be partitioned into $n^h$ product sets. The modified oracle is one way of showing this in our case.

[2] Worst case over inputs of a given length, expected over the randomness of the algorithm.

equal to the running time reduces the lower bound for DSPAN problem to proving a lower bound on the average running time of deterministic algorithms against some input distribution (uniform, in our case), that is, a lower bound on the average depth of leaves (according to the input distribution). For clarity we focus on the number of leaves in the main statement, but we actually prove that most (all but nearly a $1/q$ fraction) leaves are small (according to the input distribution). See Lemma 6.3 and the proof of Theorem 1.3 for a precise statement, as well as Section 1.3 for an overview of the argument.

Let us make some easy observations about DPP. The kind of partitions we are looking for always exist. Take any element $(p_1, \ldots, p_k)$ in $(\mathbb{F}_q P^n)^k$, where each $p_i$ is a point in $\mathbb{F}_q P^n$. Together $p_1, \ldots, p_k$ span a $(k-1)$-flat. Thus the trivial partition in which each part is a singleton is a valid partition, giving an upper bound on $N$ of size $((q^{n+1}-1)/(q-1))^k$, the total number of elements in $(\mathbb{F}_q P^n)^k$. For $q > k$, this is at most $eq^{kn}$.

A lower bound of $\Omega(q^{k(n-k+1)})$ (again assuming $q > k$) is obtained by a volume argument. The number of elements in $(\mathbb{F}_q P^n)^k$, as we noted, is $((q^{n+1}-1)/(q-1))^k$. The maximum number of elements in a part is $((q^k-1)/(q-1))^k$. This is because each factor is contained in a $(k-1)$-flat which has $(q^k-1)/(q-1)$ points. Thus $N$ must be at least

$$\left(\frac{q^{n+1}-1}{q-1}\right)^k \cdot \left(\frac{q-1}{q^k-1}\right)^k.$$

For $q > k$ this is at least $q^{kn}/eq^{k(k-1)} \geqslant q^{k(n-k+1)}/e$. Note that if we just wanted to cover instead of partition, then $\Theta(q^{k(n-k+1)})$ is the tight upper and lower bound (when $q > k$). The covering given by the $k$th powers of all $(k-1)$-flats achieves the lower bound; it is well known that the number of $(k-1)$-flats in $\mathbb{F}_q P^n$ is

$$\frac{(q^{n+1}-1)(q^{n+1}-q)\cdots(q^{n+1}-q^{k-1})}{(q^k-1)(q^k-q)\cdots(q^k-q^{k-1})} \leqslant eq^{k(n-k+1)}.$$

For the case $k = n$, the upper and lower bounds above become $O(q^{n^2})$ and $\Omega(q^n)$.

### 1.1. Related work

Problems with similar flavour, namely finding a small partition of a product set into product sets with certain properties, abound in communication complexity, and are also studied in combinatorics. Many techniques used to prove such lower bounds actually prove lower bounds on the covering number, with a few exceptions, such as the rank method [15] and certain lower bounds on the non-negative rank [11, 6]; see also [14] for some more recent work on partition lower bounds. The covering problem is easy in our setting, but the smallest covering seems to be much smaller than the smallest partition and thus does not provide insight into the size of the smallest partition. Our problem does not seem to be directly amenable to rank arguments or other techniques, although rank arguments do lie at the core of our techniques. We now discuss some specific results related to our topic.

Alon, Bohman, Holzman and Kleitman [2] consider the problem of partitioning a finite set $A = A_1 \times \cdots \times A_n$ (where $|A_i| \geqslant 2$ for all $i$) into parts of the form $B_1 \times \cdots \times B_n$, where $\emptyset \neq B_i \subsetneq A_i$ for $i = 1, \ldots, n$. They show that any such partition has size at least $2^n$. Our problem (DPP) is essentially a $q$-analogue of their problem.

Razborov [19] considers a more general partitioning problem in the context of formula complexity, albeit only for $k = 2$. Briefly, suppose we have a covering of a set $U \times V = \cup_i C_1^i \times C_2^i$. We say that a partition $\{A_1^i \times A_2^i\}$ of $U \times V$ (so $\dot\cup_i A_1^i \times A_2^i = U \times V$, where $\dot\cup$ denotes disjoint union) is a refinement of the covering $\{C_1^i \times C_2^i\}$ if, for each part $A_1^i \times A_2^i$, there is a $j$ such that $A_1^i \times A_2^i \subseteq C_1^j \times C_2^j$. Razborov considers the problem of proving a lower bound on the size of partitions refining certain coverings. Clearly, our problem for $k = 2$ is such a problem, as our partitions refine the covering of the $k$th powers of $(k-1)$-flats. Razborov gives a method of proving lower bounds for the size of such partitions. This method seems to be specific to the $k = 2$ case; for $k = 2$, specialized to our problem, this method does not seem to give a bound better than $\Omega(q^2)$.

A lower bound for DPP would imply a lower bound for a deterministic *number in hand* multiparty communication complexity problem (see [15] for an account of communication complexity). There are $k$ players. Each player is given a private (unknown to other players) point from $\mathbb{F}_q P^n$. The players want to determine a $(k-1)$-flat containing the points of all the players. Notice that the output of the communication problem is not unique, and thus here we are interested in the communication complexity of a relation rather than that of a function.

Our problem fits into the category of problems where one obtains a discrete model of a problem over the real field by changing the real field to a finite field. There are many examples of this interaction between the continuous and the discrete. The Kakeya problem over finite fields is one recent example with connections to the theory of computing; see, *e.g.*, [9]. Here too the problem becomes more tractable in the finite field setting.

### 1.2. Our results

For $k = n = 2$, the upper and lower bounds in Section 1 for the general problem become $O(q^4)$ and $\Omega(q^2)$. The truth turns out to be $\Theta(q^3)$.

**Theorem 1.1.** *In the discrete partitioning problem for $k = n = 2$, the size of the smallest partition satisfies $N = \Theta(q^3)$.*

For the general problem, we get an upper bound improving the trivial upper bound from Section 1, and generalizing the upper bound in Theorem 1.1.

**Theorem 1.2.** *The discrete partitioning problem for $k = n$ and $q \geqslant 2n$ has a partition of size $q^{\binom{n+1}{2}}(1 + O(n/q))$.*

In the previous theorem, the partition is made of parts whose factors are either a flat or a flat minus a lower-dimensional flat, which we call an *almost-flat*. For partitions of this kind we have a lower bound that matches our upper bound up to a multiplicative constant for $q \geqslant n$, and the constant approaches 1 for large $q$.

**Theorem 1.3 (partitioning lower bound for almost-flats).** *For the discrete partitioning problem, if $k = n$ and each factor of every part is an almost-flat, then the partition size satisfies*

$$N \geqslant q^{n(n+1)/2}\left(1 - \frac{1}{q}\left(\frac{q+1}{q-2}\right)^n\right).$$

Another motivation for studying such structured partitions comes from the fact that the partitions induced by decision trees for the DSPAN problem involve parts whose factors are flats minus a small number of flats. This is shown in Section 5. Our proof of Theorem 1.3 does not seem to immediately generalize to this case.

Our approach for DPP, namely the idea of using the fraction of dependent tuples as a parameter of a part to lower-bound the size of the partition in Theorem 1.3, suggests using a similar idea for CPP, perhaps the density of 'approximately dependent' tuples. While there remain technical difficulties in carrying out this approach in the continuous setting, it appears promising, and is the direct result of considering DPP.

### 1.3. Techniques

In the proof of Theorem 1.1 the key idea is that the partitioning problem can be decomposed into smaller instances of simpler partitioning problems (Lemma 3.1). These smaller problems admit rank arguments for their lower bounds and are thus easy. Our decomposition shows that on average each of these smaller problems requires a large partition via a rank argument, giving us a good overall bound. While the rank lower bounds are fairly standard, the decomposition idea seems to be new.

The high level idea of the proof of Theorem 1.3 is as follows. We classify parts into two types, large and small (defined according to the dimensions of its factors, later called 'non-dominated' and 'dominated' parts), where small parts contain at most about $q^{n^2/2}$ tuples each, while the total number of tuples is about $q^{n^2}$. On the other hand, each large part contains at least roughly a $1/q$ fraction of dependent tuples (meaning that their span has dimension less than $n-1$: see Lemma 6.3), while the set to be partitioned, $(\mathbb{F}_q P^n)^n$, contains only about a $1/q^2$ fraction of dependent tuples, which implies that large parts can only cover about a $1/q$ fraction of all tuples. The rest must be covered by small parts, which by the previous discussion needs about $q^{n^2/2}$ parts (proof of Theorem 1.3). We remark that this high level idea has the flavour of the so-called corruption bound in communication complexity (see [4]) and its subsequent generalizations (*e.g.*, [7, 14]). Most of the work in our proof is in the lower bound for the fraction of dependent tuples in large parts (Lemma 6.3), which is done by first partitioning any such part into parts having only one-dimensional factors, and then handling this case by induction (Lemma 6.2) with the aid of a Sylvester–Gallai-type property (Lemma 6.1).

### 1.4. Organization

The rest of the paper is organized as follows. Section 2 contains relevant definitions. Section 3 shows an optimal lower bound (up to constant factors) for DPP when $k = n = 2$. In Section 4 we present a non-trivial partition construction with structured parts; Section 5 shows that this construction is essentially optimal for structured partitions. The Appendix gives more details about how a solution to CPP would lead to a lower bound for sampling from convex bodies.

## 2. Preliminaries

For $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$. We will work with projective spaces over finite fields. Projective spaces over finite fields are basic and extensively studied objects; see, *e.g.*, [3] for an introduction. Here we define projective spaces and note their relevant properties. In this paragraph, we follow the exposition of [3] closely. Consider the $(n+1)$-dimensional linear space $W := \mathbb{F}_q^{n+1}$ (where $\mathbb{F}_q$ is the finite field of cardinality $q$ and $q$ is a prime power), and set $W^\times := W \setminus \{0\}$. Points in the $n$-dimensional projective space $\mathbb{F}_q P^n$ over $\mathbb{F}_q$ correspond to lines in $W$ through the origin. More precisely, for $p \in W^\times$, consider the sets $\{ap \mid a \in \mathbb{F}_q \setminus \{0\}\}$. Clearly, two such distinct sets are disjoint. These sets together give a partition of $W^\times$. The projective space $\overline{W}$ consists of these sets as points. We define the dimension of $\overline{W}$ to be $n$ and denote this projective space by $\mathbb{F}_q P^n$. It is easy to see that $|\mathbb{F}_q P^n| = (q^{n+1}-1)/(q-1)$; in particular, the cardinality of the *projective plane* $\mathbb{F}_q P^2$ is $(q^3-1)/(q-1) = q^2+q+1$. A *flat* or *subspace* of $\overline{W}$ is a set of the form $\overline{U}$ for a subspace $U$ of $W$. The dimension of $\overline{U}$ is defined to be $\dim(U)-1$; thus $\dim(\emptyset) = -1$. We will often use the term $k$-flat for a $k$-dimensional flat. For $S \subseteq \mathbb{F}_q P^n$, denote by $\operatorname{span}(S)$ the intersection of all flats containing $S$. For a tuple $(p_1, \ldots, p_k)$ of $k$ points in $\mathbb{F}_q P^n$, clearly $\dim \operatorname{span}\{p_1, \ldots, p_k\} \leqslant k-1$. We say that $(p_1, \ldots, p_k)$ is *dependent* if $\dim \operatorname{span}\{p_1, \ldots, p_k\} < k-1$. Clearly, if a sub-tuple of a tuple is dependent then the whole tuple is dependent. A projective space of dimension 2 is called a projective plane, and flats of dimension 1 are called (projective) lines. Projective planes have nice combinatorial properties; *e.g.*, each point lies in exactly $q+1$ lines, each line contains $q+1$ points, every pair of points lies on a unique line, and every pair of lines intersects in a unique point. Higher-dimensional spaces also have similar regularity properties.

**Definition 1.** We say that a subset of $\mathbb{F}_q P^n$ is an *almost-flat* if it is either a flat or a $k$-flat minus a flat of dimension at most $k-1$. Let the *dimension* of an almost-flat be the dimension of the minimal flat containing it. In particular, an almost-line is a line or a line minus a point.

We will need an appropriate counterpart for our setting (projective spaces over finite fields) of the familiar notion of orthogonal projection in projective spaces over the reals. This requires care because the notion of orthogonality can behave very differently over finite fields. In particular, a point can be orthogonal to itself.

We define the projection using *quotient by a flat*. We will only use elementary properties of quotients and our discussion here is mostly self-contained; see, *e.g.*, [10] for a detailed treatment of quotients. Let $F$ and $S$ be two flats in $\mathbb{F}_q P^n$. An equivalence relation on $F \setminus S$ (an almost-flat) is given by $p \sim q$ if and only if $\operatorname{span}((F \cap S) \cup \{p\}) = \operatorname{span}((F \cap S) \cup \{q\})$. The equivalence classes of $\sim$ are of the form $\operatorname{span}((F \cap S) \cup \{p\}) \setminus S =: [p]$ for $p \in F \setminus S$. The set of equivalence classes of $F \setminus S$ given by $\sim$ is called the quotient set and is denoted by $F/S$. Note that in our definition we did not require that $S \subseteq F$. Quotient set $F/S$ inherits the projective structure from $F$ in the natural way: for $p, q \in F \setminus S$ with $[p] \neq [q]$, the points are given by $[p]$, the lines are given by

$$\{[r] : r \in \operatorname{span}((F \cap S) \cup \{p\} \cup \{q\}) \setminus S\},$$

and so on. Thus $F/S$ is a projective space of dimension $\dim(F) - \dim(F \cap S) - 1$ living in $\mathbb{F}_q P^{n-\dim(S)-1} = \mathbb{F}_q P^n / S$. Notice that when $F \cap S = \emptyset$, then $\dim(F/S) = \dim(F)$, as $\dim(\emptyset) = -1$ according to our convention.

For a flat $F' \subseteq F$, define

$$F'|_{F/S} := \{x \in F \setminus S : [x] \in [F']\},$$

where $[F'] := \{[x] : x \in F' \setminus S\}$. In words, $F'|_{F/S}$ is the union of equivalence classes in $F \setminus S$ that intersect $F'$.

We will use the following easy facts, which we state without proof. The first claim deals with invariance of dependence under quotient.

**Claim 2.1.** *Consider $t = (p_1, \dots, p_k)$, $p_i \in \mathbb{F}_q P^n$, $p_1 \notin \{p_2, \dots, p_k\}$, and let $[p_2], \dots, [p_k]$ be the images of $p_2, \dots, p_k$ in the quotient of the space by $p_1$. Then $t$ is dependent if and only if $([p_2], \dots, [p_k])$ is dependent.*

The next claim shows that intersection of sub-flats with equivalence classes behaves nicely.

**Claim 2.2.** *For all equivalence classes $C \in F/S$ with non-empty intersection with a given flat $F'$, the intersection size $|C \cap F'|$ is the same.*

The last claim shows that dependence is a property of the equivalence classes.

**Claim 2.3.** *Let $t = (p_1, \dots, p_k, q_{k+1}, \dots, q_j, \dots q_m)$, where $p_i$ and $q_j$ are points in $\mathbb{F}_q P^n$. Let $t'$ be obtained from $t$ by replacing $q_j$ by $q'_j$. Also assume that $q_j, q'_j$ are in the same equivalence class in the quotient of $\mathbb{F}_q P^n$ by $S = \operatorname{span}(p_1, \dots, p_k)$, i.e., $\operatorname{span}(S \cup \{q_i\}) = \operatorname{span}(S \cup \{q'_i\})$. Then either both $t$ and $t'$ are dependent or both are independent.*

## 3. The discrete partitioning problem for $n = 2$

In this section, instead of the projective space $\mathbb{F}_q P^n$, we restrict ourselves to the projective plane $\mathbb{F}_q P^2$. Let us restate the problem for the projective plane. We want a partition of $(\mathbb{F}_q P^2)^2$ of the form

$$(\mathbb{F}_q P^2)^2 = \dot{\bigcup}_{i=1}^{N} A_1^i \times A_2^i, \tag{3.1}$$

such that for all $i$ we have $A_1^i \times A_2^i \subseteq (L^i)^2$, where $L^i$ is a line in $\mathbb{F}_q P^2$.

We have $|(\mathbb{F}_q P^2)^2| = (q^2 + q + 1)^2 \approx q^4$. The upper and lower bounds we discussed in Section 1 for the general problem now become $O(q^4)$ and $\Omega(q^2)$. However, it turns out that $N = \Theta(q^3)$.

**The upper bound.** First, for any point $p \in \mathbb{F}_q P^2$ there are $q+1$ lines $L_1^p, L_2^p, \dots, L_{q+1}^p$ through $p$. These lines only intersect in $p$ and together they cover all of $\mathbb{F}_q P^2$. Thus $L_1^p$ and $L_2^p \setminus \{p\}, L_3^p \setminus \{p\}, \dots, L_{q+1}^p \setminus \{p\}$ partition $\mathbb{F}_q P^2$. Now we can state our $O(q^3)$ size partition of $(\mathbb{F}_q P^2)^2$. Each part is of the form $p \times L_1^p$ or $p \times (L_i^p \setminus \{p\})$ for $i \in \{2, \dots q+1\}$ and $p \in \mathbb{F}_q P^2$. Clearly these parts are mutually disjoint. For any two parts, either the first factors are different and disjoint, or if they are the same, then the second factors are disjoint by our construction of the partition of $\mathbb{F}_qP^2$. It is also clear that we cover all of $(\mathbb{F}_qP^2)^2$ in this way. The size of this partition is $(q^2+q+1)(q+1) = O(q^3)$.

We now show that the above upper bound is the best possible up to a constant: $N = \Omega(q^3)$.

**Theorem 1.1 restated.** *In the discrete partitioning problem for $k = 2 = n$ the partition size satisfies $N = \Theta(q^3)$.*

**Proof.** The key idea of the proof is that the partitioning problem can be decomposed into smaller instances of simpler partitioning problems (Lemma 3.1 below). These smaller problems admit rank arguments (similar to the one used in some proofs of a theorem by Graham and Pollak [12]) for their lower bounds. Our decomposition shows that on average each of these smaller problems requires a large partition, giving us a good overall bound.

It will be useful to work without loss of generality with what we will call *canonical* partitions, as it is easier to prove a lower bound for this restricted kind of partition. We say that a partition of $(\mathbb{F}_qP^2)^2$ as in (3.1) is canonical if each of its parts is canonical. We say that a part $A_1 \times A_2$ is canonical if either $A_1 = A_2$ (*square parts*) or $A_1 \cap A_2 = \emptyset$ (*non-square parts*). In other words, either the two factors are equal or they are disjoint.

Given any partition $\{A_1^i \times A_2^i\}$, we can construct a canonical partition with at most four times as many parts, as follows. For each part, decompose it into four canonical parts:

$$\begin{aligned} A_1^i \times A_2^i = {} & [(A_1^i \cap A_2^i) \times (A_1^i \cap A_2^i)] \dot{\cup} [(A_1^i \setminus A_2^i) \times (A_1^i \cap A_2^i)] \\ & \dot{\cup} [(A_1^i \cap A_2^i) \times (A_2^i \setminus A_1^i)] \dot{\cup} [(A_1^i \setminus A_2^i) \times (A_2^i \setminus A_1^i)]. \end{aligned}$$

Henceforth we assume that our partitions are canonical.

It will be helpful to think of $(\mathbb{F}_qP^2)^2$ as a complete bipartite graph, with one copy of $\mathbb{F}_qP^2$ in the product representing one side of vertices and the other copy representing the other side. Edges in this graph are then the elements of $(\mathbb{F}_qP^2)^2$. Each canonical part can be thought of as an induced complete bipartite subgraph.

Clearly, the number of square parts in any canonical partition is at most $q^2+q+1 = O(q^2)$. We will show that the number of non-square parts is $\Omega(q^3)$.

Notice that if $\{S^i \times S^i \mid i \in [N]\}$ is the set of square parts, then $\{S^i\}$ form a partition of $\mathbb{F}_qP^2$. Thus, $\{S^i \mid i \in [N]\}$ also induce a partition of each line $L$; let $\phi(L)$ be the number of parts in such a partition of $L$. Clearly $\phi(L) \leqslant q+1$. The following lemma shows that on average $\phi(L)$ is almost as large as $q+1$.

**Lemma 3.1.** $\sum_L \phi(L) \geqslant q(q^2+q+1)$*, where the summation is over all lines.*

**Proof.** For any point $a$ there is some square part $S^i \times S^i$ such that $a \in S^i$. Now $a$ lies in $q+1$ lines, say, $L_1, \ldots, L_{q+1}$. Since our requirement on the partition is that $S^i$ should be completely in some line, we have that for all but at most 1 of the $q+1$ lines $L \in \{L_1, \ldots, L_{q+1}\}$ we have $|L \cap S^i| = 1$. Thus $a$ appears as a singleton in the partitions (induced by the square parts) for at least $q$ lines. So each of the $q^2+q+1$ points contributes at least $q$ to the sum, which gives the bound in the lemma. □

Remove the edges covered by square parts, then we are left with a bipartite graph whose edge set is partitioned by non-square parts. In this graph, each line $L$ induces a bipartite subgraph $G(L)$ defined as follows: $G(L)$ is the bipartite subgraph induced by a copy of $L$ in the left vertices and a copy of $L$ in the right vertices. In other words, the edges of $G(L)$ are the edges in $L \times L$ not covered by square parts. This implies that the edge set of each $G(L)$ is covered by non-square parts. Also, the edge sets of graphs $\{G(L)\}_L$ are disjoint by our construction. But a stronger property holds: each non-square part completely lies in one of the $G(L)$s. More precisely, if $R_1^i \times R_2^i$ is a non-square part such that $R_1^i \subseteq L$ and $R_2^i \subseteq L$ for some line $L$, then $(R_1^i \times R_2^i) \cap (L' \times L') = \emptyset$ for all lines $L' \neq L$.

We know that $G(L)$ looks like this. Let $L = S^1 \dot\cup \cdots \dot\cup S^{\phi(L)}$ be the partition of $L$ induced by square parts as above. Then $G(L)$ has all the edges in sets $S^i \times S^j$ for $i, j \in [\phi(L)], i \neq j$. Now an easy adaptation of the matrix proof of the Graham–Pollak theorem [12] (see Lemma 3.2 below) gives that $G(L)$ needs $\phi(L)$ non-square parts. To see this, choose one point $p^i$ from each $S^i$, and consider the subgraph of $G(L)$ induced by the vertices in both colour classes of $G(L)$ corresponding to points $\{p^1, \ldots, p^{\phi(L)}\}$. Applying Lemma 3.2 to this subgraph gives the required bound on the number of non-square parts. Thus the total number of parts we need is $\sum_L \phi(L) \geqslant q(q^2+q+1)$ by the lemma above. □

We note that the proof did not make use of the algebraic structure of the projective plane, and it holds for combinatorial projective planes as well.

**Lemma 3.2.** *Let $B = ((U,V),E)$ be a bipartite graph with $|U| = |V| = n$, and let*

$$E = \{(u_i, v_j) \mid i, j \in [n] \text{ and } i \neq j\}.$$

(*In other words, $B$ is a complete $n \times n$ bipartite graph minus a perfect matching.*) *Any partition of $E$ into complete bipartite graphs requires at least $n$ graphs.*

**Proof.** Consider the bipartite adjacency matrix $A(B)$ of $B$ (rows indexed by $U$ and columns by $V$, and $A(B)_{(u,v)} = 1$ if $(u,v) \in E$ else $A(B)_{(u,v)} = 0$). Let $B_1, \ldots, B_r$ be complete bipartite subgraphs whose edges sets partition $E$. Then we can write

$$A(B) = \sum_{i \in [r]} A(B_i). \tag{3.2}$$

The algebra in the rest of the proof is over $\mathbb{R}$. Now, notice that $\operatorname{rank} A(B) = n$ (this is because $A(B) = J - I$, where $J$ is the all ones matrix and $I$ is the identity matrix, after a suitable reordering of the vertices), but $\operatorname{rank} A(B_i) = 1$ for $i \in [r]$. The subadditivity of rank implies that $r \geqslant n$. □

We remark that there are generalizations of the Graham–Pollak theorem for hypergraphs [1, 8] and it is natural to try to use these to solve the partitioning problem for higher $k$. However, we have not succeeded in this.

## 4. A small size partition

We construct a partition of $(\mathbb{F}_qP^n)^n$ with size $O(q^{\binom{n+1}{2}})$, by generalizing our partition construction for the product of two projective planes (Section 3). More generally, the same ideas give a partition of $(\mathbb{F}_qP^n)^k$ with size $O(q^{\binom{k+1}{2}})$ (independent of $n$). Informally, for the product of two projective planes the parts were of type (point × almost-line). For $(\mathbb{F}_qP^n)^n$, parts are of type (point × almost-line × almost-2-flat × $\cdots$ × almost-$(n-1)$-flat), where an almost-$r$-flat is either an $r$-flat or an $r$-flat minus an $(r-1)$-subflat. We now describe our construction in detail.

**Proof of Theorem 1.2.** Let $1 \leqslant r < n$. For an $(r-1)$-flat $F$ consider $r$-flats $F_1, F_2, \ldots$ containing $F$. There are $(q^{n+1}-q^r)/(q^{k+1}-q^r)$ such flats and any two of them intersect precisely in $F$. This provides a partition of $\mathbb{F}_qP^n$ into almost-$r$-flats with size $(q^{n+1}-q^r)(q^{k+1}-q^r)$. The first part is $F_1$ and other parts are $F_2 \setminus F, F_3 \setminus F, \ldots$. We call this partition a *partition around $F$*.

Now to construct a partition of $(\mathbb{F}_qP^n)^n$, it will be convenient to index the $n$ copies of $\mathbb{F}_qP^n$ as $P_1, \ldots, P_n$. So we are considering a partition of $P_1 \times P_2 \times \cdots \times P_n$. We start by partitioning $P_1$. Let $\mathcal{P}_1$ be the partition of $P_1$ into singletons. For each $S_1 \in \mathcal{P}_1$, consider a partition of $P_2$ around $\operatorname{span}(S_1) = S_1$. Denote this by $\mathcal{P}_2(S_1)$. For $S_2 \in \mathcal{P}_2(S_1)$, consider a partition of $P_3$ around $\operatorname{span}(S_2)$, and so on.

Our partition of $(\mathbb{F}_qP^n)^n$ is then made up of all the parts of the form $S_1 \times \cdots \times S_n$. The number of choices for the first factor is $|\mathcal{P}_1| = (q^{n+1}-1)/(q-1)$. Having fixed the first factor $S_1$, the number of choices for the second factor is $|\mathcal{P}_2(S_1)| = (q^{n+1}-q)/(q^2-q)$, and so on. So the total number of choices is

$$\begin{aligned}\frac{q^{n+1}-1}{q-1}\cdot\frac{q^{n+1}-q}{q^2-q}\cdot\frac{q^{n+1}-q^2}{q^3-q^2}\cdots\frac{q^{n+1}-q^{n-1}}{q^n-q^{n-1}} &= \frac{q^{n+1}-1}{q-1}\cdot\frac{q^n-1}{q-1}\cdot\frac{q^{n-1}-1}{q-1}\cdots\frac{q^2-1}{q-1}\\ &\leqslant \frac{q^{\binom{n+1}{2}}}{(1-1/q)^n}\\ &\leqslant \frac{q^{\binom{n+1}{2}}}{1-n/q}.\end{aligned}$$

For $q \geqslant 2n$ we have $1/(1-n/q) \leqslant 1 + 2n/q$. The claim follows. □

## 5. The structure of decision trees for DSPAN

In this section we prove the claim from the Introduction on the structure of the partition induced by a decision tree for the DSPAN problem. Each part is a product set, where each factor is a flat minus a few flats.

**Lemma 5.1.** *Consider a deterministic decision tree for DSPAN making at most $h$ queries. Let $A$ be a part of the partition of $(\mathbb{F}_qP^n)^n$ induced by the leaves of the tree. Then:*

(1) *there is an $(n-1)$-flat $F \subset \mathbb{F}_qP^n$ with $A \subseteq F^n$,*

(2) *we can write $A = A_1 \times A_2 \times \cdots \times A_n$, where each $A_i$ is of the form $G \setminus (G_1 \cup G_2 \cup \cdots \cup G_h)$, where $G, G_1, \ldots, G_h$ are flats.*

**Proof.** Part (1) must hold because the output of the tree is correct for DSPAN.

We prove the stronger version of (2) that it holds for the set of tuples $A$ associated to *any* node of the decision tree with $h$ equal to the depth of the node. We use induction on $h$. It is clearly true for $h = 0$ (no queries, the root) as in this case $A = (\mathbb{F}_q P^n)^n$. For the inductive step, let $A \subseteq (\mathbb{F}_q P^n)^n$ be the part associated to a node of depth $h$. By the inductive hypothesis, its parent part $A'$ is of the form $A_1 \times A_2 \times \cdots \times A_n$, where each $A_i$ is of the form $G \setminus G_1 \cup G_2 \cup \cdots \cup G_{h-1}$, where $G, G_1, \dots, G_{h-1}$ are (possibly empty) flats. The query that restricts $A'$ to get $A$ is some $(n-1)$-flat $p \subseteq \mathbb{F}_q P^n$. If the result of the query is YES, the interpretation of the query means that the restriction is to intersect each $A_1, \dots, A_n$ with $p$. If the result of the query is NO and index $i \in [n]$, the interpretation of the query means that the restriction is to intersect each $A_1, \dots, A_{i-1}$ with $p$, subtract $p$ from $A_i$ and leave $A_{i+1}, \dots, A_n$ unchanged. The claimed structure holds in both cases. □

## 6. Lower bound for structured partitions

In this section we show a lower bound for the discrete partitioning problem when factors of each part are almost-flats (Theorem 1.3, Definition 1). The outline of the proof in Section 1.3 will be useful for reading the proof below.

**Definition 2 (projective lines in general position).** We say that a set of at most $n + 1$ projective lines in $\mathbb{F}_q P^n$ is in *general position* if, for any $k \in [n-1]$, no $k + 1$ of them are contained in a $k$-flat.

**Lemma 6.1 (Sylvester–Gallai-type property).** *Let $L$ be a set of at most $n + 1$ projective lines in $\mathbb{F}_q P^n$ in general position (Definition 2). Then there exists a projective line $l \in L$ that intersects the other projective lines in $L$ in at most two points, i.e., there are (at most) two points $p, q \in l$ such that $l \cap l' \in \{p, q\}$ for all $l' \in L \setminus \{l\}$.*

**Proof.** We use induction on $n$. It is true for $n = 1$. For general $n$, we will define a sequence $l_1, l_2, \dots$ of lines in $L$. We will add lines incrementally preserving $\dim \operatorname{span}\{l_1, \dots, l_i\} = i$. Start by picking any line $l_1 \in L$. Pick a line $l_2 \in L \setminus \{l_1\}$ that intersects $l_1$ (if there is no such line then $l_1$ is the desired line). In general, if there exists $l_i \in L \setminus \{l_1, \dots, l_{i-1}\}$ that intersects at least one of $l_1, \dots, l_{i-1}$, then we have $\dim \operatorname{span}\{l_1, \dots, l_i\} = \dim \operatorname{span}\{l_1, \dots, l_{i-1}\} + 1 = i$ ($l_i$ cannot be contained in $\operatorname{span}\{l_1, \dots, l_{i-1}\}$ if $L$ is in general position). If no such $l_i$ exists, then the inductive hypothesis applied to $\{l_1, \dots, l_{i-1}\}$ gives the line desired in the statement. Suppose we pick all lines in $L$ in this way and the last line is $l_k$. If $k < n + 1$, then $l_k$ intersects the others in one point. If $k = n + 1$, then the fact that $L$ is in general position implies that $l_k$ intersects at most one of $l_1, \dots, l_{n-1}$, and it can possibly intersect $l_n$. Thus, $l_k$ is the desired line. □

The previous lemma is tight in the following sense: for $n = 2$, the case of the projective plane, any three lines in general position intersect pairwise.

**Lemma 6.2 (fraction of dependent tuples in products of almost-lines).** *Let $L = (l_i)_{i=1}^{n+1}$ be a family of almost-lines in $\mathbb{F}_q P^n$ with $q \geqslant 3$. Then the number of dependent tuples in $T = \prod_{i=1}^{n+1} l_i$ is at least $(q-2)^{n-1}(q-1)$.*

**Proof.** We use induction on $n$. For $n = 1$, we are in the projective line of cardinality $q+1$, the two lines in $L$ coincide except for the missing points, and the dependent tuples are pairs of equal points. Thus, there are at least $q-1$ dependent tuples.

For general $n$, if $L$ is not in general position, use the inductive hypothesis on the subfamily of $k$ lines not in general position. The number of dependent tuples in that subset is at least $(q-2)^{k-2}(q-1)$, any completion of such a dependent tuple to an $(n+1)$-tuple is also dependent, and each can be completed in at least $q^{n+1-k}$ ways. Thus the number of dependent tuples in $L$ is at least $q^{n-k+1}(q-2)^{k-2}(q-1)$.

Otherwise, consider the line in $L$ given by Lemma 6.1 (applied to the completion of each almost-line to a line), say this line is $l_1$, and let $p$ be a point in this line that is not missing from it and such that no other line in $L$ goes through it. Consider the quotient of the whole space by $p$. In the quotient, the image of a point $p' \neq p$ is $[p']$, and the image of a line $l$ not containing $p$ is the union of the images of the points in $l$. As the almost-lines in $(l_i)_{i=2}^{n+1}$ do not contain $p$, their images in the quotient are also almost-lines. Thus the inductive hypothesis can be used on the quotient space of dimension $n-1$ and the $n$ quotient lines to conclude that the product of the quotient lines contains at least $(q-2)^{n-2}(q-1)$ dependent tuples.

Now, by the invariance of dependence (Claim 2.1), there are at least $(q-2)^{n-2}(q-1)$ dependent tuples in $T$ whose first coordinate is $p$. Also, there are at least $q-2$ choices of $p$, so there are at least $(q-2)^{n-1}(q-1)$ dependent tuples overall. □

**Definition 3.** For $Q = \prod_{i=1}^{k} Q_i$, a product of subsets of $\mathbb{F}_q P^n$, where each $Q_i$ is an almost-flat, the *dimension pattern of* $Q$, denoted by $\dim Q$, is the $k$-tuple of dimensions of the $Q_i$ sorted in non-decreasing order. We will consider the partial order on dimension patterns defined by $(s_1, \ldots, s_k) \preceq (t_1, \ldots, t_k)$ if and only if for all $i$ we have $s_i \leqslant t_i$.

**Lemma 6.3 (dependence of non-dominated almost-flats).** *Let $Q = \prod_{i=1}^{n} Q_i$ be a product of subsets of $\mathbb{F}_q P^{n-1}$, where each $Q_i$ is an almost-flat. Assume*

$$\dim Q \npreceq (0, 1, \ldots, n-1). \tag{6.1}$$

*Then the fraction of dependent tuples in $Q$ is at least*

$$\frac{1}{q+1}\left(\frac{q-2}{q+1}\right)^{n-1}.$$

**Proof.** The proof will reduce estimating the fraction in the general case to the case of lines, given by Lemma 6.2. We will do this by first reducing to the case of partitions consisting of parts with *minimal* dimension patterns satisfying (6.1) and then reducing to the case of products of lines.

The minimal dimension patterns satisfying (6.1) are the following $n-1$ patterns:

$$(1,1,\dots,1),\ (0,2,2,\dots,2),\ (0,0,3,\dots,3),\dots,(0,\dots,0,n-1,n-1).$$

Formally, they are given by $(s_1,\dots,s_n)$ for $j=1,\dots,n-1$, where

$$s_i = \begin{cases} j & i \geqslant j, \\ 0 & i < j. \end{cases}$$

It suffices to prove the lemma for $Q$ with minimal dimension patterns satisfying (6.1), because of the following two facts.

- A $Q$ with a non-minimal dimension pattern can be partitioned into parts with minimal dimension patterns. This is shown in the next claim.
- The fraction of dependent tuples in $Q$ is at least the minimum of such fractions for the parts in a partition of $Q$.

**Claim 6.4.** *Let $Q=\prod_{i=1}^n Q_i$ be a product of subsets of $\mathbb{F}_qP^{n-1}$, where each $Q_i$ is an almost-flat, and let $Q$ satisfy* (6.1). *Then $Q$ can be partitioned into parts with minimal dimension patterns and satisfying the assumptions of Lemma 6.3.*

**Proof.** Let $k+1$ be the least index $i$ such that $(\dim Q)_i \geqslant i$; such an $i$ exists because of our assumption that (6.1) is satisfied. Then we claim that we can partition $Q$ into parts of the form

$$p_1 \times \cdots \times p_k \times R_{k+1} \times \cdots \times R_n, \tag{6.2}$$

where $p_i \in Q_i$, for $i \leqslant k$, are points, and $R_i \subseteq Q_i$ is an almost-flat of dimension $k+1$ for $i > k$. We construct this partition by first partitioning individual factors in $Q$, and then the resulting (refined) product partition of $Q$ will be our desired partition.

Partitioning into flats of dimension 0 (points) is straightforward. For partitioning into higher-dimensional parts there are three cases depending on the factor being partitioned and the dimension of the target parts. We will also assume that when we need to partition an almost flat it's of type $F_d \setminus F_{d''}$ with $F_{d''} \subseteq F_d$. We have the following three cases.

**Case 1.** We want to partition a $d$-flat $F_d$ into almost-flats of dimension $d'$ for some $0 < d' \leqslant d$. Fix a $(d'-1)$-flat $F_{d'} \subseteq F_d$ arbitrarily, and consider the $d'$-dimensional flats

$$F(p) := \operatorname{span}(\{p\} \cup F_{d'}) \quad \text{for } p \in F_d \setminus F_{d'}.$$

For two such points $p, p'$ we either have $F(p) = F(p')$ or $F(p) \cap F(p') = F_{d'}$. Thus we can construct a partition of $F_d$ with one flat of the form $F(p)$ and almost-flats of the form $F(p) \setminus F_{d'}$. More precisely, fix any point $p^* \in F_d \setminus F_{d'}$. Then the partition is

$$\{F(p^*)\} \cup \{F(p) \setminus F_{d'} : p \in F_d \setminus F(p^*)\}.$$

**Case 2.** We need to partition $F_d \setminus F_{d''}$, a $d$-flat minus a $d''$-flat, into almost-flats of dimension $d'$ for $d > d'' \geqslant d' > 0$. This is a slight modification of the previous argument. We fix a $(d'-1)$-flat $F_{d'} \subseteq F_{d''}$ arbitrarily and we can construct a partition of $F_d \setminus F_{d''}$

with almost-flats of the form $F(p) \setminus F_{d'}$. The partition is

$$\{F(p) \setminus F_{d'} : p \in F_d \setminus F_{d''}\}.$$

**Case 3.** We need to partition $F_d \setminus F_{d''}$, a $d$-flat minus a $d''$-flat, into almost-flats of dimension $d'$ for $d > d' > d'' > 0$. This is again a slight modification of the previous argument. We arbitrarily fix a $(d'-1)$-dimensional flat $F_{d'} \subseteq F_d$ containing $F_{d''}$ and we construct a partition of $F_d \setminus F_{d''}$ with one almost-flat of the form $F(p) \setminus F_{d''}$ and almost-flats of the form $F(p) \setminus F_{d'}$. More precisely, fix any point $p^* \in F_d \setminus F_{d'}$. The partition is

$$\{F(p^*) \setminus F_{d''}\} \cup \{F(p) \setminus F_{d'} : p \in F_d \setminus F(p^*)\}.$$

Applying the above procedure to each factor $Q_i$ for $i > k$ with $d' = k$, we get the desired partition, completing the proof of the claim. □

To complete the proof of the lemma, we now reduce the case of minimal dimension patterns to the case of lines, which is handled by Lemma 6.2. That lemma gives a lower bound for the fraction of dependent tuples for the product of $n+1$ lines in $\mathbb{F}_qP^n$. At this point in the proof we are dealing with parts as in (6.2), which have as factors $k$ points and $n-k$ almost-flats of dimension $k+1$. We could partition the almost-flats into lines to apply Lemma 6.2 and ignore the first $k$ points of each tuple, but then the lines would be living in $\mathbb{F}_qP^{n-1}$ with only $n-k$ lines, and Lemma 6.2 would not apply for $k \geqslant 1$. To fix this, we confine the almost-flats into a common $(n-k-1)$-dimensional space by 'projecting them orthogonal to $p_1, \ldots, p_k$', or more precisely by taking the quotient by $S = \operatorname{span}(p_1, \ldots, p_k)$ and then appropriately modifying the $R_i$. We now describe this procedure.

Let $Q$ be as in (6.2). If $(p_1, \ldots, p_k)$ is dependent, there is nothing more to prove for this part as the fraction of dependent tuples is 1.

Otherwise, we sequentially go over $R_{k+1}, \ldots, R_n$ and replace them by $P_{k+1}, \ldots, P_n$, as described below. The new product set $Q' = p_1 \times \cdots \times p_k \times P_{k+1} \times \cdots \times P_n$ has the following properties: (1) $f(Q') \leqslant f(Q)$, where $f(Q)$ is defined to be the fraction of dependent tuples in $Q$; (2) each $P_i$ is an almost-flat; (3) each $P_i$ is the union of some of the equivalence classes induced by the quotient of $R_i$ by $S$. Thus, if we take the quotient, then each $P_i$ can be identified with an almost-flat living in a space isomorphic to $\mathbb{F}_qP^{n-k-1}$, and hence Lemma 6.2 is applicable after partitioning each $P_i$ into lines.

Now we explain the construction of the $P_i$, which depends on two cases. (1) If $R_i$ is a flat, then set $P_i := R_i \setminus S$. (2) Suppose $R_i$ is an almost-flat, *i.e.*, $R_i = F_i \setminus F_i'$, where $F_i$ is a $(k+1)$-flat and $F_i'$ is a sub-flat of dimension at most $k$. Then set $P_i$ to either $F_i \setminus S$ or $(F_i \setminus S) \setminus (F_i'|_{F_i/S})$, whichever makes the current density of dependent tuples smaller. If the second option is empty, pick the first, which is never empty. (By the current density of dependent tuples we mean the density of dependent tuples in $p_1 \times \cdots \times p_k \times P_{k+1} \times \cdots \times P_i \times R_{i+1} \times \cdots \times R_n$.)

We do not use the more natural choice of a straightforward quotient in the case of almost-flats (that is, $P_i = R_i/S$), as in that case the fraction of dependent tuples may increase or decrease. With our choice we will now show that the fraction of dependent tuples never increases.

**Claim 6.5.** *The fraction of dependent tuples in $Q$ is at least that in $Q'$.*

**Proof.** We will see the effect on the fraction of dependent tuples in each step of our procedure of replacing $R_i$ by $P_i$ as defined above. There will be several cases.

**Case 1.** If $R_i$ is a $(k+1)$-flat, then we set $P_i := R_i \setminus S$. This cannot increase the fraction of dependent tuples because we removed $S$ and the tuples involving points of $S$ in the $i$th position are all dependent.

**Case 2.** If $R_i = F_i \setminus F_i'$, with $F_i \supset F_i'$, is an almost-flat, then we use a refinement of the previous argument. First, we replace $R_i$ by $R_i' = (F_i \setminus F_i') \setminus S = (F_i \setminus S) \setminus (F_i' \setminus S)$; as before, this cannot increase the fraction of dependent tuples. Now we have two cases depending on the intersection pattern of $F_i'$ with the equivalence classes in $F_i/S$.

**Case 2(a).** $F_i/S = [F_i']$, that is, $F_i'$ intersects all equivalence classes of the quotient $F_i/S$, and in this case each intersection is of the same cardinality by Claim 2.2. Therefore, by Claim 2.3 the fraction of dependent tuples does not change when we replace $R_i' = (F_i \setminus S) \setminus (F_i' \setminus S)$ by $P_i = F_i \setminus S$.

**Case 2(b).** $F_i/S \supsetneq [F_i']$, that is, $F_i'$ does not intersect all equivalence classes of $F_i/S$. For $U \subseteq R_i'$, define $f(U)$ to be the fraction of dependent tuples in[3]

$$p_1 \times \cdots \times p_k \times P_{k+1} \times \cdots \times P_{i-1} \times U \times R_{i+1} \times \cdots \times R_n.$$

Informally, we will either 'remove the equivalence classes intersected by $F_i'$' or 'complete them', whichever does not increase $f(\cdot)$. More precisely, we will show that for one of the following choices of $P_i$ we have $f(P_i) \leqslant f(R_i')$: set $P_i = (F_i \setminus S) \setminus (F_i'|_{F_i/S})$ ('remove'), or set $P_i = F_i \setminus S$ ('complete').

It remains to prove that one of these choices will not increase $f(\cdot)$. We need some notation. Denote the equivalence classes in $F_i/S$ by

$$C_1, \ldots, C_r, C_{r+1}, \ldots, C_{r+s}.$$

Let $c := |C_1| = |C_2| = \cdots = |C_{r+s}|$. Of these, $C_1, \ldots, C_r$ have non-empty intersection with $F_i'$. Let $c := |C_1| = |C_2| = \cdots = |C_{r+s}|$. By Claim 2.2, $|C_1 \cap F_i'| = \cdots = |C_r \cap F_i'|$ and let us denote this common intersection size by $c'$. Let $\alpha$ be the fraction of dependent tuples induced by $C_1 \cup \cdots \cup C_r = F_i'|_{F_i/S}$, and let $\beta$ be the fraction of dependent tuples induced by $C_{r+1} \cup \cdots \cup C_{r+s}$. Then we have

$$\begin{aligned} f(R_i') &= \frac{\alpha(c-c')r + \beta cs}{(c-c')r + cs}, \\ f(F_i \setminus S) &= \frac{\alpha cr + \beta cs}{cr + cs}, \\ f((F_i \setminus S) \setminus (F_i'|_{F_i/S})) &= \frac{\beta cs}{cs} = \beta. \end{aligned}$$

[3] We are overloading the function $f$ as it was used with a different type of argument ($Q$) earlier, but this should not cause confusion.

From the above expressions we see that if $\beta > \alpha$ then $f(F_i \setminus S) < f(R_i')$, if $\beta < \alpha$ then $f((F_i \setminus S) \setminus (F_i'|_{F_i/S})) < f(R_i')$, and if $\alpha = \beta$ then either choice works.

This completes the proof of the claim. □

Let $P = P_{k+1} \times \cdots \times P_n$. By our construction, the fraction of dependent tuples in $Q'$ is no less than that in $P$. Define $P/S := (P_{k+1}/S) \times \cdots \times (P_n/S)$, the result of taking the quotient with respect to $S$, where $(P_j/S) \subseteq \mathbb{F}_q P^{n-k-1}$ for $k < j \leqslant n$. Note that $P_j/S$ is an almost-flat. We have $f(P/S) = f(p_1 \times p_2 \times \cdots \times p_k \times P)$.

Claim 6.5 with the fact just noted implies that a lower bounding of the fraction of dependent tuples of $Q$ is given by a lower bound of the fraction of dependent tuples of a part having all factors of dimension 1 or more. Applying the partitioning argument from the first half of the proof once more to such a part, it is enough to lower-bound the fraction of dependent tuples for a part having factors of dimension exactly 1 (minimal dimension pattern). The estimate in Lemma 6.2 gives that each such part with $n-k$ factors has at least $(q-2)^{n-k-2}(q-1)$ dependent tuples. A part like that also has at most $(q+1)^{n-k}$ tuples and therefore a fraction of at least

$$\frac{(q-2)^{n-k-2}(q-1)}{(q+1)^{n-k}}$$

dependent tuples. As a function of $k$ only, this fraction is smallest when $k = 0$, and thus it is at least

$$\frac{(q-2)^{n-1}}{(q+1)^n}.$$

We showed that this is a lower bound on the fraction of dependent tuples in $Q$. This completes the proof of the lemma. □

**Proof of Theorem 1.3.** We will first estimate the fraction of dependent tuples in $(\mathbb{F}_q P^n)^n$. Probabilistic language is helpful here. We consider a random tuple $T = (t_1, \ldots, t_n)$ and we want an upper bound on the probability that it is dependent. Recall that the cardinality of an $i$-dimensional flat is $1 + q + \cdots + q^i$. Then

$$\begin{aligned}
\Pr(T \text{ is dependent}) &= \sum_{i=2}^{n} \Pr((t_1, \ldots, t_{i-1}) \text{ is independent and } (t_1, \ldots, t_i) \text{ is dependent}) \\
&\leqslant \sum_{i=2}^{n} \Pr((t_1, \ldots, t_i) \text{ is dependent} \mid (t_1, \ldots, t_{i-1}) \text{ is independent}) \\
&= \sum_{i=2}^{n} \frac{1 + q + \cdots + q^{i-2}}{1 + q + \cdots + q^n} \\
&\leqslant \sum_{i=2}^{n} \frac{1}{q^{n-i+2}} \leqslant \sum_{i=2}^{\infty} \frac{1}{q^i} = \frac{1}{q(q-1)}.
\end{aligned}$$

Thus the fraction of dependent tuples in $(\mathbb{F}_q P^n)^n$ is at most $1/(q(q-1))$.[4] This and Lemma 6.3 imply that parts whose dimension pattern is not less than or equal to $(0,1,\ldots,n-1)$ (non-dominated) can cover at most a

$$\frac{1}{q(q-1)}\left(\frac{1}{q+1}\left(\frac{q-2}{q+1}\right)^{n-1}\right)^{-1} \leqslant \frac{1}{q}\left(\frac{q+1}{q-2}\right)^{n}$$

fraction of $(\mathbb{F}_q P^n)^n$. The rest has to be covered with 'dominated' parts, that is, parts whose dimension pattern is less than or equal to $(0,1,\ldots,n-1)$. Any such part has cardinality at most $1(q+1)\cdots(q^{n-1}+\cdots+1)$. The total number of tuples to be covered by these parts is at least

$$\left(1-\frac{1}{q}\left(\frac{q+1}{q-2}\right)^{n}\right)(q^n+\cdots+1)^n.$$

This needs at least

$$q^{n(n+1)/2}\left(1-\frac{1}{q}\left(\frac{q+1}{q-2}\right)^{n}\right)$$

parts. □

## Acknowledgements

We thank László Lovász, Alexander Razborov, Michael Saks, Miklos Santha and David Xiao for useful discussions and pointers to the literature. We also thank an anonymous reviewer for careful reading and useful comments.

## Appendix: From sampling lower bound to the continuous partitioning problem

A question that immediately arises when trying to prove a lower bound on sampling is that sampling is not a computational task in the usual sense of having a definite output. A way to get around this problem is to prove lower bounds for a problem that can be solved using sampling. An $\Omega(n)$ lower bound is easy. Consider the following set of $n$ bodies in $\mathbb{R}^n$. For $i \in [n]$, define body $B_i = [0,1]^{i-1} \times [0,2] \times [0,1]^{n-i}$. In other words, $B_i$ is an axis-parallel cuboid with length 1 along all but the $i$th axis. Now consider a randomized algorithm that gets as input (via membership oracle) a uniformly randomly chosen body from the set of bodies just defined and its output is the index of the input body. A straightforward application of Yao's minimax principle shows that any such algorithm must make $\Omega(n)$ membership queries to achieve a constant probability of success. On the other hand, if sampling can be done with $q$ queries, then the body can be identified in

[4] This estimate is not too far from the true value. By picking the points in the tuple in sequence and considering the chance that the last point makes the tuple dependent (*i.e.*, lies in a certain $(n-2)$-dimensional flat), we have that the fraction of dependent tuples is at least

$$\frac{q^{n-1}-1}{q^{n+1}-1} \geqslant \frac{1}{q^2} - \frac{1}{q^{n+1}}.$$

$O(q)$ queries with constant probability of success. Suppose that the input body was $B_i$. Sample a point by making $q$ queries. With probability about $1/2$ its $i$th coordinate is greater than 1, thus telling us that the body is $B_i$. We can improve the probability of success by repeating this. This gives $q = \Omega(n)$.

For a quadratic lower bound as a function of the dimension, our candidate hard algorithmic problem is the following. We are given a membership oracle for a convex body given by $\{x \in \mathbb{R}^n \mid \langle x, v_i \rangle \leqslant 1 \text{ for } i \in [n-1], \langle x, v \rangle \leqslant p(n)\}$, for $n-1$ unit vectors $v_1, \ldots, v_{n-1} \in S^{n-1}$ (the unit sphere in $\mathbb{R}^n$) spanning a hyperplane, $v$ a normal to that hyperplane and $p(n)$ some fixed polynomial in $n$. The problem is to find $v$ approximately, or more precisely, a vector whose direction makes an angle with $v$ that is at most $1/\operatorname{poly}(n)$. As usual in algorithmic convexity, the oracle complexity of problems of this kind depends on the roundness of the input body [13], and our problem as stated can have very high complexity as there is no *a priori* bound on the roundness of the input. For a meaningful worst-case lower bound for randomized algorithms one needs to restrict the input body so that it contains $rB_n$ and is contained in $RB_n$ for $R/r = \operatorname{poly}(n)$ (where $B_n$ is the unit ball in $\mathbb{R}^n$). It is easy to show algorithms solving the problem in this case with essentially quadratic number of queries. Yao's lemma implies that the probability of success of any randomized algorithm against the worst such input is at least the probability of success of the best deterministic algorithm against a distribution on inputs of our choice. Choosing $v_i$ uniformly and independently at random in $S^{n-1}$ and restricting the distribution to bodies satisfying the roundness condition is a natural option. But it seems cleaner simply to choose $v_i \in S^{n-1}$ uniformly at random without any additional constraint, prove a lower bound for deterministic algorithms against this distribution (say, an algorithm that fails with probability at most $p$ needs to make $q$ queries), and then argue that for a suitable choice of $r$ and $R$ the fraction of the distribution that is not well-rounded is at most $p/2$. So any algorithm when running on a distribution of well-rounded bodies needs to make at least $q$ queries to fail with probability at most

$$\frac{p/2}{1-(p/2)} = p/(2-p).$$

As before, it is easy to see that if we can sample with $O(q)$ queries then we can find a vector whose direction is within a $1/\operatorname{poly}(n)$ angle of $v$ in $O(q \operatorname{polylog}(n))$ queries with constant probability.

The next observation is that any deterministic algorithm against our distribution can be thought of as a decision tree (if we only care about the number of queries and not the computational complexity). Every node represents a query, the children of a node represent different choices depending on the result of a query, and on leaves the algorithm stops and has to output a candidate vector. The leaves induce a partition of the support of the input distribution, which can be thought to be $(S^{n-1})^{n-1}$. The algorithm succeeds with high probability if, for most parts, most tuples of $n-1$ vectors in the part have their normal direction near a fixed vector that depends on the part ('most' here according to the input distribution). It simplifies the problem somewhat to assume that the oracle gives a bit more information than just YES or NO. Instead, the modified oracle answers YES when the query point is in the body (as usual), but when the query point $x \in \mathbb{R}^n$

is not in the body it answers NO and gives the least index among violated constraints (that is, $\min\{i : |x \cdot v_i| > 1\}$). This idea was introduced in [18] and it has the following consequence (as shown there). The partition induced by the corresponding decision tree is made of product sets, namely, every part is of the form $P_1 \times \cdots \times P_{n-1}$, where $P_i \subseteq S^{n-1}$. Clearly a lower bound on the number of queries for algorithms with the modified oracle is a valid lower bound for the original oracle.

For any given part that is a product set, it can be shown that if the angle of localization of the normals to its tuples is forced to be small enough (say, normal directions are within an angle $1/n^{O(1)}$ of a given direction for a $1-\alpha$ fraction of the part, for a small constant $\alpha$), then most of the part lies in a narrow 'band', that is, it satisfies the following 'band condition'. For a set of the form $P_1 \times \cdots \times P_{n-1} \subseteq (S^{n-1})^{n-1}$, there is a vector $v \in S^{n-1}$ such that

$$\mu(P_i \cap \{x : |v \cdot x| \leqslant 1/n^{O(1)}\}) \geqslant (1-\alpha)\mu(P_i) \quad \text{for all } i$$

(where $\mu$ denotes surface area).

The previous discussion reduces the problem of proving a lower bound $\Omega(n^2/\log n)$ for sampling to the following partitioning problem.

**Continuous partitioning problem (informal).** Suppose $Q_1, \ldots, Q_k$ is a partition of $(S^{n-1})^{n-1}$ where each part is a product set and satisfies the above 'band condition'. A lower bound of $k \geqslant 2^{\Omega(n^2)}$ would translate to a quadratic query lower bound for the sampling problem. (The loss of a log factor is explained by the fact that the decision tree associated to the modified oracle has fan-out $\Theta(n)$.)

A natural approach to solving the partitioning problem is to try to discretize the problem perhaps by subdividing the sphere into sufficiently small cells, and then working with these cells as atoms. However, we found the discretization considered in this paper cleaner and more useful to work with. Although we do not have a formal connection between the two problems, they have a very similar flavour, and insights from the discrete version can be directly useful for the continuous version. For example, the partition in the proof of Theorem 1.2 translates into a non-trivial partition of $(S^{n-1})^{n-1}$ satisfying the band condition above. We now briefly describe the construction of this partition. We first give an infinite size partition, which is essentially the one in the proof of Theorem 1.2 except that now we are working over the real field; the parts are of the form $P_1 \times P_2 \times \cdots \times P_{n-1}$. The first factor $P_1 \subset S^{n-1}$ is a point and its antipode in $S^{n-1}$ (this corresponds to a single point in projective space). The second factor $P_2 \subset S^{n-1}$ is obtained from a great circle $C$ in $S^{n-1}$ containing $P_1$; factor $P_2$ is either $C$ itself or $C \setminus P_1$ (this corresponds to a line in the projective space). Factor $P_3$ is obtained from the intersection of $S^{n-1}$ with a three-dimensional subspace of $\mathbb{R}^n$ containing $P_2$ (this corresponds to a plane in the projective space), and so on. To turn this into a finite partition, we 'fatten' each factor by $1/p'(n)$, where the polynomial $p'(n)$ is related to the precision with which we have to determine $v$, the normal to $v_1, \ldots, v_{n-1}$. For points, this fattening is achieved by subdividing $S^{n-1}$ into regions of diameter at most $1/p'(n)$. For a given $P_1$, the second factor $P_2$ is obtained by similarly partitioning $S^{n-1}$ into a finite number of regions such that for each region there is a great circle with every point in the region within distance $1/p'(n)$ from the great circle, and one of these regions contains $P_1$ and the others are disjoint from $P_1$. We proceed similarly for higher-dimensional factors.

## References

[1] Alon, N. (1986) Decomposition of the complete $r$-graph into complete $r$-partite $r$-graphs. *Graphs Combin.* **2** 95–100.

[2] Alon, N., Bohman, T., Holzman, R. and Kleitman, D. J. (2002) On partitions of discrete boxes. *Discrete Math.* **257** 255–258.

[3] Babai, L. and Frankl, P. (1992) *Linear Algebra Methods in Combinatorics*, Department of Computer Science, University of Chicago.

[4] Beame, P., Pitassi, T., Segerlind, N. and Wigderson, A. (2006) A strong direct product theorem for corruption and the multiparty communication complexity of disjointness. *Comput. Complex.* **15** 391–432.

[5] Bertsimas, D. and Vempala, S. (2004) Solving convex programs by random walks. *J. Assoc. Comput. Mach.* **51** 540–556.

[6] Braun, G., Jain, R., Lee, T. and Pokutta, S. (2013) Information-theoretic approximations of the nonnegative rank. *Electronic Colloquium on Computational Complexity: ECCC* **20** 158.

[7] Chakrabarti, A. and Regev, O. (2012) An optimal lower bound on the communication complexity of gap-Hamming-distance. *SIAM J. Comput.* **41** 1299–1317.

[8] Cioaba, S. M., Kündgen, A. and Verstraëte, J. (2009) On decompositions of complete hypergraphs. *J. Combin. Theory Ser. A* **116** 1232–1234.

[9] Dvir, Z. (2009) On the size of Kakeya sets in finite fields. *J. Amer. Math. Soc.* **22** 1093–1097.

[10] Faure, C.-A. and Frölicher, A. (2000) *Modern Projective Geometry*, Kluwer.

[11] Fawzi, H. and Parrilo, P. A. (2012) New lower bounds on nonnegative rank using conic programming, http://arxiv.org/abs/1210.6970.

[12] Graham, R. and Pollak, H. (1971) On the addressing problem for loop switching. *Bell Syst. Tech. J.* **50** 2495–2519.

[13] Grötschel, M., Lovász, L. and Schrijver, A. (1988) *Geometric Algorithms and Combinatorial Optimization*, Springer.

[14] Jain, R. and Klauck, H. (2010) The partition bound for classical communication complexity and query complexity. In *IEEE Conference on Computational Complexity*, pp. 247–258.

[15] Kushilevitz, E. and Nisan, N. (1996) *Communication Complexity*, Cambridge University Press.

[16] Lovász, L. and Vempala, S. (2006) Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *IEEE Symposium on Foundations of Computer Science*, pp. 57–68.

[17] Motwani, R. and Raghavan, P. (1995) *Randomized Algorithms*, Cambridge University Press.

[18] Rademacher, L. and Vempala, S. (2008) Dispersion of mass and the complexity of randomized geometric algorithms. *Adv. Math.* **219** 1037–1069.

[19] Razborov, A. A. (1990) Applications of matrix methods to the theory of lower bounds in computational complexity. *Combinatorica* **10** 81–93.