On the Beck-Fiala conjecture for random set systems

Esther Ezra$^{1,2}$ | Shachar Lovett$^3$

1 Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel
2 School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia
3 Computer Science and Engineering, University of California, San Diego, California

Correspondence
Esther Ezra, School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160.
Email: eezra3@math.gatech.edu

Funding information
This research was supported by the NSF CAREER; CCF-15-53354, Israel Science Foundation; 824/17.

Abstract
Motivated by the Beck-Fiala conjecture, we study discrepancy bounds for random sparse set systems. Concretely, these are set systems $(X, \Sigma)$, where each element $x \in X$ lies in $t$ randomly selected sets of $\Sigma$, where $t$ is an integer parameter. We provide new bounds in two regimes of parameters. We show that when $|\Sigma| \geq |X|$ the hereditary discrepancy of $(X, \Sigma)$ is with high probability $O(\sqrt{t \log t})$; and when $|X| \gg |\Sigma|^t$ the hereditary discrepancy of $(X, \Sigma)$ is with high probability $O(1)$. The first bound combines the Lovász Local Lemma with a new argument based on partial matchings; the second follows from an analysis of the lattice spanned by sparse vectors.

KEYWORDS
beck-fiala conjecture, discrepancy theory, random set systems

1 INTRODUCTION

Let $(X, \Sigma)$ be a finite set system, with $X$ a finite set and $\Sigma$ a collection of subsets of $X$. A two-coloring of $X$ is a mapping $\chi : X \rightarrow \{-1, +1\}$. For a subset $S \in \Sigma$ we define $\chi(S) := \sum_{x \in S} \chi(x)$. The discrepancy of $\Sigma$ is defined as

$$\text{disc}(\Sigma) := \min_{\chi} \max_{S \in \Sigma} |\chi(S)|.$$ 

In other words, the discrepancy of the set system $(X, \Sigma)$ is the minimum over all colorings $\chi$ of the largest deviation from an even split, over all subsets in $\Sigma$. For background on discrepancy theory, we refer the reader to the books of Chazelle [8] and Matoušek [12].

In this paper, our interest is in the discrepancy of sparse set systems. The set system $(X, \Sigma)$ is said to be $t$-sparse if any element $x \in X$ belongs to at most $t$ sets $S \in \Sigma$. A well-known result of Beck and Fiala [4] is that sparse set systems have discrepancy bounded only in terms of their sparsity.

Theorem 1.1 ([4]) If $(X, \Sigma)$ is $t$-sparse then $\text{disc}(\Sigma) \leq 2t - 2$. 

Random Struct Alg. 2019;54:665–675. wileyonlinelibrary.com/journal/rsa © 2018 Wiley Periodicals, Inc.
The bound was improved to $2t – 3$ by Bednarchak and Helm [5], to $2t – 4$ by Helm [11], and to $2t – \log^* t$ by Bukh [7]. However, Beck and Fiala conjectured that the actual bound should be $O(\sqrt{t})$, analogous to Spencer’s theorem for nonsparse set systems [16]. This is a long standing open problem in discrepancy theory. The best result to date (which allows dependency on the size of the set system) is by Banaszczyk [2], and very recently Bansal, Dadush, and Garg [3] showed a matching constructive bound.

**Theorem 1.2** ([2, 3])  
If $(X, \Sigma)$ is $t$-sparse with $|X| = n$ then $\text{disc}(\Sigma) \leq O(\sqrt{t \log n})$.

## 1.1 Our results

In this paper, we study random sparse set systems. To sample a random $t$-sparse set system $(X, \Sigma)$ with $|X| = n, |\Sigma| = m$, for each $x \in X$ choose uniformly and independently a subset $T_x \subset [m]$ of size $|T_x| = t$. Then set $S_i = \{x \in X : i \in T_x\}$ and $\Sigma = \{S_1, \ldots, S_m\}$. Letting $E[\cdot]$ denote expectation, our main quantity of interest is $E[\text{disc}(\Sigma)]$. We show that when $m \geq n$, this is close to the conjectured bound of Beck and Fiala. Specifically, we show $E[\text{disc}(\Sigma)] = O(\sqrt{t \log t})$. In particular, the bound does not depend on $n$.

In fact, we obtain such bound for the hereditary discrepancy of the set system. For $Y \subset X$ let $\Sigma|_Y = \{S \cap Y : S \in \Sigma\}$ be the set system restricted to $Y$. The hereditary discrepancy of a set system $(X, \Sigma)$ is defined as

$$\text{herdisc}(\Sigma) = \max_{Y \subset X} \text{disc}(\Sigma|_Y).$$

Our main result is the following.

**Theorem 1.3**  
Assume $m \geq n \geq t$. Let $(X, \Sigma)$ be a random $t$-sparse set system with $|X| = n, |\Sigma| = m$. Then

$$E[\text{disc}(\Sigma)] \leq E[\text{herdisc}(\Sigma)] \leq O(\sqrt{t \log t}).$$

In fact, the bound holds with probability $1 - \exp(-\Omega(t))$.

We note that our technique can be extended to the case where $m \geq cn$ for any absolute constant $c > 0$, but fails whenever $m \ll n$. The main reason is that in this regime, most sets are large (formally, larger than $\Omega(t)$). Nevertheless, when $n$ is considerably larger than $m$, we use a different approach and show that the discrepancy is small in this case as well. Specifically, when $n$ is somewhat larger than $\binom{m}{t}$ we show that the discrepancy is only $O(1)$.

**Theorem 1.4**  
Fix $m \geq t$ and let $N = \binom{m}{t}$. Assume that $n \geq \Omega(N \log N)$. Let $(X, \Sigma)$ be a random $t$-sparse set system with $|X| = n, |\Sigma| = m$. Then

$$E[\text{disc}(\Sigma)] = O(1).$$

In fact, the bound holds with probability $1 - N^{-\Omega(1)}$.

To summarize, the work in this paper was motivated by the elusive Beck-Fiala conjecture. We consider a natural setting of random $t$-sparse set systems, and show that in this case, under some regimes of parameters, the conjecture holds (with the bound of $O(\sqrt{t})$ replaced by the slightly weaker
bound of $O(\sqrt{t \log t})$ in our first result). We hope that the techniques developed in this work will be useful for the study of random sparse set systems in the full spectrum of parameters, as well as for the original Beck-Fiala conjecture.

1.2 Further research

We conclude the introduction with open problems regarding the discrepancy of random set systems. The first is the natural extension of this work to all ranges of parameters.

Open problem 1.5 Is true that for any $m, n \geq t$, a random $t$-sparse set system with $|X| = n, |\Sigma| = m$ has expected discrepancy $O(\sqrt{t})$?

The next question is about set systems induced by permutations. Let $\pi$ be a permutation of $[n]$. It defines the set system given by the intervals $\Sigma_\pi = \{ \{ j : \pi(j) \leq i \} : i \in [n] \}$. Given a collection of permutations $\Pi = \{ \pi_1, \ldots, \pi_k \}$, define $\Sigma_\Pi = \bigcup_{\pi \in \Pi} \Sigma_\pi$ to be the set system obtained by their union. It is known that if $k = 1$ or $k = 2$ then $\text{disc}(\Pi) = O(1)$, see for example [17]. It was conjectured by Beck that the same holds for any constant $k$, however this was refuted by Newman et al. [15] who constructed an example of $k = 3$ permutations for which the discrepancy is $\Omega(\log n)$. This is tight as any family of $k$ permutations has discrepancy $O(\sqrt{k \log n})$ [6, 18].

Here, we ask about the case of random permutations. Does a “typical” instance have constant or super-constant discrepancy? for concreteness, we post the question for $k = 3$.

Open problem 1.6 Let $\pi_1, \pi_2, \pi_3$ be uniformly chosen permutations of $[n]$, and let $\Pi = \{ \pi_1, \pi_2, \pi_3 \}$. Is the expected discrepancy of $\Sigma_\Pi$ an absolute constant, or does it grow with $n$?

2 Preliminaries and proof overview

The Lovász Local Lemma [9] is a powerful probabilistic tool. In this paper we only need its symmetric version.

Theorem 2.1 Let $E_1, E_2, \ldots, E_k$ be a series of events such that each event occurs with probability at most $p$ and such that each event is independent of all the other events except for at most $d$ of them. If $ep(d + 1) \leq 1$ then $\Pr[\wedge_{i=1}^{k} \overline{E_i}] > 0$, where $e$ is the base of the natural logarithm.

In our analysis we exploit several standard tail bounds for the sum of independent random variables (ie, Chernoff-Hoeffding bounds, see eg, [1]).

Lemma 2.2 (Tail bounds for additive error) Let $Z_1, \ldots, Z_k \in \{-1, 1\}$ be independent random variables and let $Z = Z_1 + \ldots + Z_k$. Then for any $\lambda > 0$

$$\Pr \left[ |Z - \mathbb{E}[Z]| \geq \lambda \sqrt{k} \right] \leq 2 \exp(-2\lambda^2).$$

Lemma 2.3 (Tail bounds for multiplicative errors) Let $Z_1, \ldots, Z_k \in \{0, 1\}$ be independent random variables and let $Z = Z_1 + \ldots + Z_k$. Then for any $\lambda > 0$

$$\Pr \left[ Z \geq (1 + \lambda)\mathbb{E}[Z] \right] \leq \exp(-\lambda^2/3 \cdot \mathbb{E}[Z]).$$
2.1 Proof overview for Theorem 1.3

We next present an overview of our proof for Theorem 1.3. For simplicity of exposition, we present the overview only for the derivation of the discrepancy bound. In Section 3 we present the actual analysis and show a bound on the hereditary discrepancy.

First, we classify each set as being either “small” if its cardinality is $O(t)$, or “large” otherwise. Then we proceed in several steps:

(i) **Making large sets pairwise disjoint:** Initially, we show that with high probability over the choice of the set system, it is possible to delete at most one element from each large set, such that they become pairwise disjoint after the deletion. This property is proved in Lemma 3.1.

(ii) **Partial matching:** For each large set resulting after step (i), we pair its elements, leaving at most, say, two unpaired elements. Since each pair appears in a unique set, this process results in a partial matching $M = \{(a_1, b_1), \ldots, (a_k, b_k)\}$ on $X$. We observe that as soon as we have such a matching, we can restrict the two-coloring function $\chi$ on $X$ to assign alternating signs on each pair of $M$. Since each large set $S$ has at most two unpaired elements, we immediately conclude that $|\chi(S)| \leq 2$.

(iii) **Applying the Lovász local lemma on the small sets:** We are thus left to handle the small sets. In this case, we observe that a random coloring $\chi$, with alternating signs on $M$ as above\(^1\), satisfies with positive probability that $|\chi(S)| \leq O(\sqrt{t \log t})$ for all small sets $S \in \Sigma$. This is a consequence of the Lovász Local Lemma, as each small set $S$ contains only $O(t)$ elements, and each of these elements participates in $t$ sets of $\Sigma$. The fact that some of these elements appear in the partial matching implies that $S$ can “influence” (w.r.t. the random coloring $\chi$) at most $2|S|t = O(t^2)$ other small sets; see Section 3 for the details.

We point out that as soon as we have a partial matching $M$ as above, we can “neutralize” the deviation that might be caused by the large sets, and only need to keep the deviation, caused by the small sets, small. The latter is fairly standard to do, and so the main effort in the analysis is to show that we can indeed make large sets disjoint as in step (i).

We note that our proof technique is constructive. Our arguments for steps (i) and (ii) (see Lemma 3.1 and our charging scheme in Claim 4.2) give an efficient algorithm to find an element to delete in each large set, thereby making large sets disjoint, as well as build the partial matching, or, alternatively, report (with small probability) that a partial matching of the above kind does not exist and halt (in case the random set system does not behave as “expected”). In step (iii) we can apply the algorithmic Lovász Local Lemma of Moser and Tardos [13, 14], since the colors are assigned independently among the pairs in $M$ as well as the unpaired elements. Thus, we obtain an expected polynomial time algorithm, which, with high probability over the choice of the set system, constructs a coloring with discrepancy $O(\sqrt{t \log t})$.

3 A LOW HEREDITARY DISCREPANCY BOUND: THE ANALYSIS

We now proceed with the proof of Theorem 1.3. We classify the sets in $\Sigma$ based on their size. A set $S \in \Sigma$ is said to be large if $|S| \geq 6t$ and small otherwise. Note that since $m \geq n$, most sets in $\Sigma$ are small. Let $I = \{i : S_i \text{ is large}\}$ be a random variable capturing the indices of the large sets. To

---
\(^1\)That is, each pair in $M$ is assigned $(+1, -1)$ or $(-1, +1)$ independently with probability $1/2$, and any remaining (unmatched) element of $X$ is assigned $+1$ independently with probability $1/2$. 
construct a coloring, we proceed in several steps. First, we show that with high probability the large sets are nearly disjoint. We will assume throughout that $t$ is sufficiently large (concretely $t \geq 55$).

**Lemma 3.1** Fix $t \geq 55$. Let $E$ denote the following event: “there exists a choice of $x_i \in S_i$ for $i \in I$ such that the sets $\{S_i \setminus \{x_i\} : i \in I\}$ are pairwise disjoint”. Then $\Pr[E] \geq 1 - 2^{-t}$.

We defer the proof of Lemma 3.1 to Section 4 and prove Theorem 1.3 based on it, in the remainder of this section. Observe that

$$
\mathbb{E}[\text{herdisc} (\Sigma)] = \mathbb{E}[\text{herdisc}(\Sigma)|E] \Pr[E] + \mathbb{E}[\text{herdisc}(\Sigma)|\bar{E}] \Pr[\bar{E}]
$$

$$
\leq \mathbb{E}[\text{herdisc}(\Sigma)|E] + (2t - 1) \Pr[\bar{E}]
$$

$$
\leq \mathbb{E}[\text{herdisc}(\Sigma)|E] + 1
$$

where we bounded $\mathbb{E}[\text{herdisc}(\Sigma)|\bar{E}]$ by the Beck-Fiala theorem (Theorem 1.1) which holds for any $t$-sparse set system, and bounded $\Pr[\bar{E}]$ by $2^{-t}$ according to Lemma 3.1. To conclude the proof we will show that when $E$ holds then $\text{herdisc}(\Sigma) \leq O(\sqrt{t \log t})$. Thus, we assume from now on that the event $E$ holds. Fix a subset $Y \subset X$, where we will construct a two-coloring for $\Sigma' = \Sigma|_Y$ of low discrepancy.

Partition each $S_i \cap Y = A_i \cup B_i$ for $i \in I$, where $|A_i|$ is even, $|B_i| \leq 2$ and the sets $\{A_i : i \in I\}$ are pairwise disjoint. Partition each $A_i$ arbitrarily into $|A_i|/2$ pairs, and let $M$ be the union of these pairs, over all $i \in I$. That is, $M$ is a partial matching on $Y$ given by $M = \{(a_1, b_1), \ldots, (a_k, b_k)\}$ where $a_1, b_1, \ldots, a_k, b_k \in Y$ are distinct, each $A_i$ is a union of a subset of $M$, and each pair $a_j, b_j$ appears in a unique set $A_i$ due to the fact that these sets are pairwise disjoint (they thus form a partition of $M$). We say that a coloring $\chi : Y \to \{-1, +1\}$ is consistent with $M$ if $\chi(a_j) = -\chi(b_j)$ for all $j \in [k]$. Note that if $S_i$ is a large set, then for any coloring $\chi$ consistent with $M$,

$$
|\chi(S_i \cap Y)| = |\chi(A_i) + \chi(B_i)| = |0 + \chi(B_i)| \leq |B_i| \leq 2.
$$

Thus, we only need to minimize the discrepancy of $\chi$ over the small sets in $\Sigma$. To do so, we choose $\chi$ uniformly from all two-colorings consistent with $M$. These are given by choosing uniformly and independently $\chi(a_i) \in \{-1, +1\}$ for $i \in [k]$, setting $\chi(b_i) = -\chi(a_i)$ and choosing $\chi(x) \in \{-1, +1\}$ uniformly and independently for all $x \not\in \{a_1, b_1, \ldots, a_k, b_k\}$.

Let $S_i$ be a small set, that is $|S_i| \leq 6t$. Let $E_i$ denote the event

$$
E_i := \left[|\chi(S_i \cap Y)| \geq c\sqrt{t \log t}\right],
$$

where $c > 0$ is a sufficiently large constant. Each pair $\{a_j, b_j\}$ contained in $S_i$ contributes 0 to the discrepancy, and all other elements obtain independent colors. Hence $\chi(S_i)$ is the sum of $t' \leq 6t$ independent signs. By Lemma 2.2, for an appropriate choice of $c$ we have

$$
\Pr[E_i] \leq 1/100t^2.
$$

We next claim that each event $E_i$ depends on at most $d = 12t^2$ other events $\{E_j : j \neq i\}$. Indeed, let $S'_i = S_i \cup \{a_j : b_j \not\in S_i\} \cup \{b_j : a_j \not\in S_i\}$. Then $|S'_i| \leq 2|S_i| \leq 12t$ and $\chi(S_i)$ is independent of $\chi(x)$ for all $x \not\in S'_i$. So, if $E_i$ depends on $E_j$, it must be the case that $S_j$ intersects $S'_i$. However, as each $x \in S'_i$ is contained in $t$ sets, there are at most $12t^2$ such events $E_j$. 
We are now in a position to apply the Lovász Local Lemma (Theorem 2.1). Its conditions are satisfied as we have \( p = 1/100r^2 \) and \( d = 12r^2 \). Hence \( \Pr[\land E_i] > 0 \), that is, there exists a coloring \( \chi \) consistent with \( M \) for which \( |\chi(S_i)| \leq c\sqrt{t}\log t \) for all small sets \( S_i \). This coloring shows that \( \text{disc}(\Sigma') \leq \max(c\sqrt{t}\log t, 2) \) as claimed.

4 | PROOF OF LEMMA 3.1

Let \((X, \Sigma)\) be a \( t \)-sparse set system with \(|X| = n, |\Sigma| = m\). It will be convenient to identify it with a bi-partite graph \( G = (X, V, E) \) where \(|V| = m\) and \( E = \{(x, i) : x \in S_i\} \). Then, a random \( t \)-sparse set system is the same as a random left \( t \)-regular bi-partite graph. That is, a uniform graph satisfying \( \text{deg}(x) = t \) for all \( x \in X \).

Large sets in \( \Sigma \) correspond to the subset of the vertices \( V' = \{v \in V : \text{deg}(v) \geq 6t\} \). For a vertex \( v \in V \) let \( \Gamma(v) \subseteq X \) denote its neighbors. Lemma 3.1 is equivalent to the following lemma, which we prove in this section.

Lemma 4.1 Fix \( t \geq 55 \). With probability at least \( 1 - 2^{-t} \) over the choice of \( G \), there exists a choice of \( x_v \in \Gamma(v) \) such that the sets \( \{\Gamma(v) \setminus \{x_v\} : v \in V'\} \) are pairwise disjoint.

Let \( G' \) be the induced (bi-partite) sub-graph on \((X, V')\). We will show that with high probability \( G' \) has no cycles. In such a case Lemma 4.1 follows from the straightforward scheme described below:

Claim 4.2 Assume that \( G' \) has no cycles. Then there exists a choice of \( x_v \in \Gamma(v) \) such that the sets \( \{\Gamma(v) \setminus \{x_v\} : v \in V'\} \) are pairwise disjoint.

Proof We present a charging scheme of the vertices \( x_v \in \Gamma(v), \) for each \( v \in V' \). If \( G' \) has no cycles then it is a forest, that is, a disjoint union of trees. For any maximal tree \( T \) in \( G' \) fix an arbitrary root \( v_T \in V' \cap T \), and orient the edges of \( T \) from \( v_T \) to the leaves. For each \( v \in V' \) other than the root, choose \( x_v \) to be the parent of \( v \) in the tree \( T \), and choose \( x_{v_T} \) arbitrarily. Let \( A_v = \Gamma(v) \setminus \{x_v\} \) for \( v \in V' \). We claim that \( \{A_v : v \in V'\} \) are pairwise disjoint. To see that, assume for the sake of contradiction that \( x \in A_{v_1} \cap A_{v_2} \) for some \( x \in X, v_1, v_2 \in V' \). Then \( v_1, x, v_2 \) is a path in \( G' \) and hence \( v_1, v_2 \) must belong to the same tree \( T \). However, the only case where this can happen (as \( T \) is a tree) is that \( x \) is the parent of either \( v_1 \) or \( v_2 \) in \( T \). However, by construction in this case \( x = x_{v_1} = x_{v_2} \) and hence \( x \notin A_{v_1} \cup A_{v_2} \), from which we conclude that \( \{A_v : v \in V'\} \) are pairwise disjoint, as claimed.

In the remainder of the proof we show that with high probability \( G' \) has no cycles. The girth of \( G' \), denoted girth\((G')\), is the minimal length of a cycle in \( G' \) if such exists, and otherwise it is \( \infty \). Note that since \( G' \) is bipartite, then girth\((G')\) is (if finite) the minimal \( 2\varepsilon \) such that there exists a cycle \( x_1, v_1, x_2, v_2, \ldots, x_\varepsilon, v_\varepsilon, x_1 \) in \( G' \) with \( x_i \in X \) and \( v_i \in V' \).

Claim 4.3 \( \Pr[\text{girth}(G') = 4] \leq t^4 \exp(-t) \).

Proof Fix distinct \( x_1, x_2 \in X \) and distinct \( v_1, v_2 \in V \). They form a cycle of length 4 if \( v_1, v_2 \in \Gamma(x_1) \cap \Gamma(x_2) \). As each \( \Gamma(x_i) \) is a uniformly chosen set of size \( t \) we have that

\[
\Pr[v_1, v_2 \in \Gamma(x_1) \cap \Gamma(x_2)] = \left( \frac{\binom{t}{2}}{\binom{m}{2}} \right)^2 \leq (t/m)^4.
\]
Next, conditioned on the event that \( v_1, v_2 \in \Gamma(x_1) \cap \Gamma(x_2) \), we still need to have \( v_1, v_2 \in V' \) (that is \( v_1, v_2 \) represent large sets of \( \Sigma \)). We will only require that \( v_1 \in V' \) for the bound. Note that so far we only fixed \( \Gamma(x_1), \Gamma(x_2) \), and hence the neighbors of \( \Gamma(x) \) for \( x \neq x_1, x_2 \) are still uniform. Then \( v_1 \in V' \) if at least \( 6t - 2 \) other nodes \( x \in X \) have \( v_1 \) as their neighbor. We will apply Lemma 2.3 to bound the probability of this event.

To that end, for each \( x \in X \setminus \{x_1, x_2\} \) define the indicator random variable \( Z_x = 1_{v_1 \in \Gamma(x)} \). Observe that \( \{Z_x\} \) are independent and \( \Pr[Z_x = 1] = t/m \). Let \( Z = \sum Z_x \), where \( \mathbb{E}[Z] = t(n - 2)/m \leq t \) by our assumption that \( n \leq m \). Thus by Lemma 2.3 we have

\[
\Pr[v_1 \in V'|v_1, v_2 \in \Gamma(x_1) \cap \Gamma(x_2)] \leq \Pr[Z \geq 6t - 2] \leq \exp(-((5t - 2)/t)^2/3 \cdot t) \leq \exp(-t).
\]

So,

\[
\Pr[v_1, v_2 \in \Gamma(x_1) \cap \Gamma(x_2) \land v_1 \in V'] \leq (t/m)^4 \cdot \exp(-t).
\]

To bound \( \Pr[\text{girth}(G') = 4] \) we union bound over all \( \binom{n}{3} \binom{m}{2} \) choices of \( x_1, x_2, v_1, v_2 \). Using our assumption that \( m \geq n \) we get

\[
\Pr[\text{girth}(G') = 4] \leq m^4(t/m)^4 \exp(-t) \leq t^4 \exp(-t).
\]

**Claim 4.4** For any \( \ell \geq 3 \), \( \Pr[\text{girth}(G') = 2\ell] \leq \exp(-t\ell). \)

**Proof** Let \( x_1, v_1, \ldots, x_{\ell}, v_{\ell} \) denote a potential minimal cycle of length \( 2\ell \). As it is a minimal cycle and \( \ell \geq 3 \), the vertices \( v_i, v_j \) have no common neighbors, unless \( j = i + 1 \) in which case \( x_i \) is the only common neighbor of \( v_i, v_{i+1} \) (where indices are taken modulo \( \ell \)). Thus there exist sets \( X_i \subset X \) of size \( |X_i| = 6t - 2 \) such that \( X_i \subset \Gamma(v_i) \) and \( X_1, \ldots, X_{\ell}, \{x_1, \ldots, x_{\ell}\} \) are pairwise disjoint.

Let \( E(x_1, v_1, \ldots, x_{\ell}, v_{\ell}, x_1, \ldots, X_{\ell}) \) denote the event described above, for a fixed choice of \( x_1, v_1, \ldots, x_{\ell}, v_{\ell}, X_1, \ldots, X_{\ell} \). The event holds if

1. \( v_i, v_{i+1} \) are neighbors of \( x_i \).
2. \( v_i \) is a neighbor of all \( x \in X_i \).

These are independent events, as \( \Gamma(x) \) is independently chosen for each \( x \in X \). So

\[
\Pr[E(x_1, v_1, \ldots, x_{\ell}, v_{\ell}, X_1, \ldots, X_{\ell})] = \prod_{i=1}^{\ell} \Pr[v_i, v_{i+1} \in \Gamma(x_i)] \cdot \prod_{i=1}^{\ell} \prod_{x \in X_i} \Pr[v_i \in \Gamma(x)] = \left( \frac{\binom{t}{2}}{\binom{m}{2}} \right)^{\ell} \cdot \left( \frac{t}{m} \right)^{(6t-2)\ell} \leq \left( \frac{t}{m} \right)^{6t\ell}.
\]

To bound \( \Pr[\text{girth}(G') = 2\ell] \) we union bound over all choices of \( x_1, v_1, \ldots, x_{\ell}, v_{\ell}, X_1, \ldots, X_{\ell} \). The number of choices is bounded by (recall that we assume \( n \leq m \)):

\[
n^{\ell} m^{\ell} \left( \frac{n}{6t-2} \right)^{\ell} \leq \left( \frac{nm \cdot e^{6t-2} \cdot n^{6t-2}}{(6t-2)^{6t-2}} \right)^{\ell} \leq \left( \frac{(em)^{6t}}{(6t - 2)^{6t-2}} \right)^{\ell}.
\]
Thus,
\[
\Pr[girth(G') = 2\epsilon] \leq \left( \frac{(em)^{6t}}{(6t - 2)^{6t - 2}} \right)^\epsilon \cdot \left( \frac{t}{m} \right)^{6t\epsilon} = \left( (6t - 2)^2 \left( \frac{et}{6t - 2} \right)^{6t} \right)^\epsilon \leq \exp(-t\epsilon).
\]

**Proof of Lemma 4.1** Using Claims 4.3 and 4.4, the probability that girth(G') < \infty is bounded by:
\[
\Pr[girth(G') < \infty] = \sum_{\ell=2}^{\infty} \Pr[girth(G') = \ell] \leq t^\ell \exp(-t) + \sum_{\ell=3}^{\infty} \exp(-t\ell) \leq 2t^\ell \exp(-t).
\]
For \(t \geq 55\), we have that \(\Pr[girth(G') < \infty] \leq 2^{-t}\).

## 5 | THE REGIME OF LARGE SETS

We next prove Theorem 1.4. Let \((X, \Sigma)\) be a \(t\)-sparse set system with \(|X| = n, |\Sigma| = m\). In this setting, we consider the case of fixed \(m, t\) and \(n \to \infty\). Consider its \(m \times n\) incidence matrix. The columns are \(t\)-sparse vectors in \(\{0, 1\}^m\), and hence have \(N = \binom{m}{t}\) possible values. When \(n \gg N\), there will be many repeated columns. We show that in this case, the discrepancy of the set system is low. Setting notations, let \(v_1, \ldots, v_N \in \{0, 1\}^m\) be all the possible \(t\)-sparse vectors, and let \(r_1, \ldots, r_N\) denote their multiplicity in the set system. Note that they define the set system uniquely (up to permutation of the columns, which does not effect the discrepancy).

In what follows we will assume that \(m\) is large enough and that \(4 \leq t \leq m - 4\). We note that if \(t \leq 3\) or \(t \geq m - 3\) then result immediately follows from the Beck-Fiala theorem (Theorem 1.1), for any set systems. The first case follows by a direct application, and the second case by first partitioning the columns to pairs and subtracting one vector from the next in each pair, which gives a 6-sparse \([-1, 0, 1]\) matrix, to which we apply the Beck-Fiala theorem, and then assign a two-coloring function on the original columns in the obvious way. Our main result in this section is the following:

**Theorem 5.1** Let \((X, \Sigma)\) be a \(t\)-sparse set system with \(4 \leq t \leq m - 4\) and \(m\) large enough. Assume that \(\min(r_1, \ldots, r_N) \geq 7\). Then \(\text{disc}(\Sigma) \leq 2\).

Note that the statement in Theorem 5.1 is somewhat stronger than that in Theorem 1.4, as it only assumes that all possible \(t\)-sparse column vectors comprise the incidence matrix of \((X, \Sigma)\), and their multiplicity is 7 or higher. In fact, Theorem 1.4 follows from Theorem 5.1 using a straightforward coupon-collector argument [10]. In this regime, with high probability (say, with probability at least \(1 - 1/N\), a random sample of \(\Theta(N \log N)\) columns guarantees that each \(t\)-sparse column appears with multiplicity 7 (or higher). Therefore, we obtain:
\[
\mathbb{E}[\text{disc}(\Sigma)] \leq 2 \left( 1 - \frac{1}{N} \right) + \frac{2t - 1}{N} = O(1).
\]

We are thus left to prove Theorem 5.1. First, we present an overview of the proof.

### 5.1 | Proof Overview

Every column \(v_i\) is repeated \(r_i\) times. As we may choose arbitrary signs for each occurrence of a vector, the aggregate total would be \(c_i v_i\), where \(c_i \in \mathbb{Z}, |c_i| \leq r_i\) and \(c_i \equiv r_i \mod 2\). Our goal is to show that such a solution \(c_i\) always exists, for which \(\| \sum c_i v_i \|_\infty\) is bounded, for any initial settings of \(r_1, \ldots, r_N\), as long as they are all large enough.
We show that such a solution always exists, with \(|c_i| \leq 7\). In order to show it, we first fix some solution with the correct parity, and then correct it to a low discrepancy solution, by adding an even number of copies of each vector. In order to do that, we study the integer lattice \(L\) spanned by the vectors \(v_1, \ldots, v_N\), as our correction comes from \(2L\). We show that \(L = \{x \in \mathbb{Z}^m : \sum x_i = 0 \mod t\}\), which was already proved by Wilson [19] in a more general scenario. However, we need an additional property: vectors in \(L\) are efficiently spanned by \(v_1, \ldots, v_N\). This allows us to perform the above correction efficiently, keeping the number of times that each \(v_i\) is repeated bounded. Putting that together, we obtain the result.

5.2 Proof of Theorem 5.1

Initially, we investigate the lattice spanned by the vectors \(v_1, \ldots, v_N\). As the sum of the coordinates of each of them is \(t\), they sit within the lattice

\[
L = \left\{ x \in \mathbb{Z}^m : \sum x_i \equiv 0 \mod t \right\}.
\]

We first show that they span this lattice, and moreover, they do so effectively.

**Lemma 5.2** For any \(w \in L\) there exist \(a_1, \ldots, a_N \in \mathbb{Z}\) such that \(\sum a_i v_i = w\). Moreover, \(|a_i| \leq A\) for all \(i \in [N]\) where

\[
A = \frac{2\|w\|_1}{\binom{m}{\lfloor m/2 \rfloor}} + 2.
\]

**Proof** Assume first that we have \(\sum w_i = 0\). We will later show how to reduce to this case. Pair the positive and negative coordinates of \(w\). For \(L = \|w\|_1/2\) let \((i_1, j_1), \ldots, (i_L, j_L)\) be pairs of elements of \([m]\) such that: if \((i, j)\) is a pair then \(w_i > 0, w_j < 0\); each \(i \in [m]\) with \(w_i > 0\) appears \(w_i\) times as the first element in a pair; and each \(j \in [m]\) with \(w_j < 0\) appears \(-w_j\) times as the second element in a pair. For any \(\ell \in [L]\) we will choose a set \(S_{\ell} \subseteq [m]\) of size \(t - 1\), disjoint from \(i_{\ell}, j_{\ell}\). Define \(I_\ell = S_{\ell} \cup \{i_{\ell}\}\) and \(J_\ell = S_{\ell} \cup \{j_{\ell}\}\). Note that \(I_\ell, J_\ell\) are both subsets of \([m]\) of size \(t\). To simplify notations, for any set \(I \subseteq [m]\) of size \(t\), we denote by \(v_I \in \{0, 1\}^m\) its indicator vector, which is one of \(v_1, \ldots, v_N\). We thus have

\[
w = \sum_{\ell=1}^{L} v_{I_\ell} - v_{J_\ell}.
\]

We choose the sets \(S_1, \ldots, S_L\) to minimize the maximum number of times that each vector from \([v_1, \ldots, v_N]\) is repeated in the decomposition. When we choose \(S_\ell\), we can choose one of \(M = \binom{m-2}{t-1}\) many choices. There is a choice for \(S_\ell\) such that both \(I_\ell\) and \(J_\ell\) appeared thus far less than \(2\ell/M\) times. Choosing such a set, we maintain the invariant that after choosing \(S_1, \ldots, S_\ell\), each vector is repeated at most \(2\ell/M + 1\) times. Thus, at the end each vector is repeated at most \(2L/M + 1\) times.

In the general case, we have \(\sum w_i = st\), where we may assume \(s > 0\). We apply the previous argument to \(w - (v_{i_1} + \ldots + v_{i_s})\), whose coordinates sum to zero. We choose \(i_1, \ldots, i_s \in [N]\) (potentially with repetitions) so as to minimize the maximum number of times that each vector participates; this number is \(\lfloor s/N \rfloor \leq \|w\|_1/M + 1\). Combining the two estimates, we obtain that at the end each vector is repeated at most \(4L/M + 2 = 2\|w\|_1/M + 2\) times.

**Lemma 5.3** For any \(b_1, \ldots, b_N \in \{0, 1\}\) there exist \(c_1, \ldots, c_N \in \mathbb{Z}\) such that

\[
(i) \quad c_i \equiv b_i \mod 2.
\]
\[(ii) \| \sum c_i v_i \|_\infty \leq 2.\]
\[(iii) |c_i| \leq 7 \text{ for all } i \in [N].\]

**Proof** As a first step, choose \(z_i \in \{-1, 0, 1\}\) such that \(z_i = 0\) if \(b_i = 0\), and \(z_i \in \{-1, 1\}\) chosen uniformly if \(b_i = 1\). Let \(u = \sum z_i v_i\). Note that for \(j \in [m]\), if there are \(k_j\) indices \(i \in [N]\) for which \((v_i)_j = 1\) and \(b_i = 1\), then \(E_c[u_j^2] = k_j\). Thus,
\[E_c[\|u\|_2^2] = \sum k_j \leq Nt.\]

Thus, with probability at least \(1/2\), \(\|u\|_2 \leq \sqrt{2Nt}\) and hence \(\|u\|_1 \leq \sqrt{2Ntm}\). Fix such a \(u\).

Next, we choose \(w \in \mathcal{L}\) such that \(\|u - 2w\|_\infty\) is bounded. If we only wanted that \(w \in \mathbb{Z}^m\) we could simply choose \(q \in \{0, 1\}^m\) with \(q_i \equiv u_i \mod 2\) and take \(w = (u - q)/2\). In order to guarantee that \(w \in \mathcal{L}\), namely that \(\sum w_i \equiv 0 \mod 2\), we change at most \(t\) coordinates in \(q\) by adding or subtracting 2. Thus, we obtain \(q \in \{-2, -1, 0, 1, 2\}^m\) where \(q_i \equiv u_i \mod 2\) and set \(w = (u - q)/2 \in \mathcal{L}\). We have \(\|u - 2w\|_\infty \leq 2\).

Next, we apply Lemma 5.2 to \(w\). Using the notation there, we obtain a decomposition \(w = \sum q_i v_i\). This implies that if we set \(c_i = z_i - 2a_i\) then indeed \(c_i \equiv b_i \mod 2\) and \(\| \sum c_i v_i \|_\infty = \|u - 2w\|_\infty \leq 2\). To bound \(|c_i|\), note that \(\|w\|_1 \leq \|u\|_1/2 + m\). We have by Lemma 5.2 that \(|a_i| \leq A\) for
\[A = \eta + 2 \leq 3,\]
where
\[\eta = 2 \|w\|_1 \leq O \left( \frac{\sqrt{mt \binom{m}{t}}}{\binom{m-2}{t-1}} \right) \leq \left( \frac{m^{3/2}}{\binom{m}{t}} \right) \leq 1,\]
whenever \(4 \leq t \leq m - 4\) and \(m\) is large enough, as is easily verified by the fact that the last term is a decreasing function of \(m\). Thus \(|c_i| \leq 7\) as asserted.

**Proof of Theorem 5.1** Assume that \(r_1, \ldots, r_N \geq 7\). By Lemma 5.3, there exists \(c_i \in \mathbb{Z}\) such that \(c_i \equiv r_i \mod 2\), \(|c_i| \leq 7\) and \(\| \sum c_i v_i \|_\infty \leq 2\). For each \(i \in [N]\), we color \(|c_i|\) of the vectors \(v_i\) with \(\operatorname{sign}(c_i) \in \{-1, +1\}\) and the remaining \(r_i - |c_i|\) vectors with alternating +1 and −1 colors (so that their contribution cancels, since \(r_i - |c_i|\) is even). The total coloring produces exactly the vector \(\sum c_i v_i\), which as guaranteed has discrepancy bounded by 2.

**ACKNOWLEDGMENTS**

The authors wish to thank Aravind Srinivasan for presenting this problem during a discussion at the IMA Workshop on the Power of Randomness in Computation. Esther Ezra is supported by an NSF CAREER award CCF-15-53354 and by Grant 824/17 from the Israel Science Foundation. Shachar Lovett is supported by an NSF CAREER award 1350481, CCF award 1614023 and a Sloan fellowship. A preliminary version of this paper appeared in the 20th International Workshop on Randomization and Computation (RANDOM’2016).

**REFERENCES**

1. N. Alon and J. H. Spencer, *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York, Chichester, Weinheim, 2000.
2. W. Banaszczyk, *Balancing vectors and Gaussian measures of n-dimensional convex bodies*, Rand. Struct. Algor. **12** (1998), no. 4, 351–360.

3. N. Bansal, D. Dadush, and S. Garg, *An algorithm for Komlós conjecture matching Banaszczyk’s bound*, FOCS (2016).

4. J. Beck and T. Fiala, *Integer-making theorems*, Discr. Appl. Math. **3** (1981), no. 1, 1–8.

5. D. Bednarchak and M. Helm, *A note on the Beck-Fiala theorem*, Combinatorica. **17** (1997), no. 1, 147–149.

6. G. Bohus, *On the discrepancy of 3 permutations*, Rand. Struct. Algor. **1** (1990), no. 2, 215–220.

7. B. Bukh, *An improvement of the Beck-Fiala theorem*. arXiv: 1306.6081, 2013.

8. B. Chazelle, *The Discrepancy Method: Randomness and Complexity*, Cambridge University Press, Cambridge, New York, 2000.

9. P. Erdős and L. Lovász, *Problems and results on 3-chromatic hypergraphs and some related questions*, Inf. Finite Sets (to Paul Erdős on his 60th birthday). **II** (1975), 609–627.

10. P. Erdős and A. Rényi, *On a classical problem of probability theory*, Publ. Math. Inst. Hung. Acad. Sci. **6** (1961), 215–220.

11. M. Helm, *On the Beck-Fiala theorem*, Discr. Math. **207** (1999), no. 1, 73–87.

12. J. Matoušek, *Geometric Discrepancy: An Illustrated Guide*. Vol. 18, Springer Science & Business Media, Berlin/Heidelberg, Germany, 2009.

13. R. A. Moser, *A constructive proof of the Lovász local lemma*. Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, ACM, New York, pp. 343–350, 2009.

14. R. A. Moser and G. Tardos, *A constructive proof of the general Lovász local lemma*, J. ACM. **57** (2010), no. 2, 11.

15. A. Newman, O. Neiman, and A. Nikolov, *Beck’s three permutations conjecture: A counterexample and some consequences*. 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, Washington, DC, 2012, pp. 253–262.

16. J. Spencer, *Six standard deviations suffice*, Trans. Am. Math. Soc. **289** (1985), no. 2, 679–706.

17. J. H. Spencer, *Ten Lectures on the Probabilistic Method*. Vol. 52, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

18. J. H. Spencer, A. Srinivasan, and P. Tetali. The discrepancy of permutation families. Unpublished manuscript, 2001.

19. R. M. Wilson, *A diagonal form for the incidence matrices of t-subsets vs k-subsets*, Eur. J. Combinatorics. **11** (1990), no. 6, 609–615.

---

**How to cite this article**: Ezra E., Lovett S. On the Beck-Fiala conjecture for random set systems. *Random Struct Alg.* 2019;54:665–675. https://doi.org/10.1002_rsa.20810