On the Beck-Fiala Conjecture for Random Set Systems

Esther Ezra [*]   Shachar Lovett [†]

November 1, 2015

Abstract

Motivated by the Beck-Fiala conjecture, we study discrepancy bounds for random sparse set systems. Concretely, these are set systems (X, Σ), where each element $x \in X$ lies in t randomly selected sets of Σ, 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, Σ) is with high probability $O(\sqrt{t \log t})$; and when $|X| \gg |\Sigma|^t$ the hereditary discrepancy of (X, Σ) 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.

1 Introduction

Let (X, Σ) be a finite set system, with X a finite set and Σ 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 Σ is defined as

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

In other words, the discrepancy of the set system (X, Σ) is the minimum over all colorings χ of the largest deviation from an even split, over all subsets in Σ. For background on discrepancy theory, we refer the reader to the books of Chazelle [Cha00] and Matoušek [Mat09].

In this paper, our interest is in the discrepancy of sparse set systems. The set system (X, Σ) 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 [BF81] is that sparse set systems have discrepancy bounded only in terms of their sparsity.

Theorem 1.1 ([BF81]). *If (X, Σ) is t-sparse then* $\mathrm{disc}(\Sigma) \leq 2t - 1$.

[*]School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. eezra3@math.gatech.edu.

[†]Computer Science and Engineering, University of California, San Diego. slovett@cse.ucsd.edu. Research supported by an NSF CAREER award and a Sloan fellowship.

Beck and Fiala conjectured that in fact, the bound can be improved to $O(\sqrt{t})$, analogous to Spencer's theorem for non-sparse set systems [Spe85]. This is a long standing open problem in discrepancy theory. The best result to date is by Banaszczyk [Ban98].

Theorem 1.2 ([Ban98])**.** *If* (X, Σ) *is* t-*sparse with* $|X| = n$ *then* $\mathrm{disc}(\Sigma) \leq O(\sqrt{t \log n})$.

Our results. In this paper, we study random sparse set systems. To sample a random t-sparse set system (X, Σ) 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 $\mathbb{E}[\cdot]$ denote expectation, our main quantity of interest is $\mathbb{E}[\mathrm{disc}(\Sigma)]$. We show that when $m \geq n$, this is close to the conjectured bound of Beck and Fiala. Specifically, we show $\mathbb{E}[\mathrm{disc}(\Sigma)] = O(\sqrt{t \log t})$. 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, Σ) is defined as

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

Our main result is the following.

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

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

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. 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, Σ) *be a random* t-*sparse set system with* $|X| = n, |\Sigma| = m$. *Then*

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

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 considered a natural setting of random t-sparse set systems, and showed that in this case, in 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.

2 Preliminaries and Proof Overview

The Lovász Local Lemma. The Lovász Local Lemma [EL75] is a powerful probabilistic tool. In this paper we only need its symmetric version.

Theorem 2.1. *Let $E_1, E_2, ..., E_k$ 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 $\Pr[\wedge_{i=1}^m \overline{E_i}] > 0$.*

Tail bounds. In our analysis we exploit a few standard tail bounds for the sum of independent random variables (Chernoff-Hoeffding bounds, see, e.g., [AS00]).

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 χ 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 χ, with

3

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 Σ. The fact that some of these elements appear in the partial matching implies that S can "influence" (w.r.t. the random coloring χ) 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 step (iii) we can apply the algorithmic Lovász Local Lemma of Moser and Tardos [Mos09, MT10], 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 Σ based on their size. A set $S \in \Sigma$ is said to be *large* if $|S| \geq 6t$ and *small* otherwise. Note that as $m \geq n$, most sets in Σ are small. Let $I = \{i : S_i \text{ is large}\}$ be a random variable capturing the indices of the large sets. To 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. Decompose

$$
\begin{aligned}
\mathbb{E}[\mathrm{herdisc}(\Sigma)] &= \mathbb{E}[\mathrm{herdisc}(\Sigma)|E] \Pr[E] + \mathbb{E}[\mathrm{herdisc}(\Sigma)|\overline{E}] \Pr[\overline{E}] \\
&\leq \mathbb{E}[\mathrm{herdisc}(\Sigma)|E] + (2t - 1) \Pr[\overline{E}] \\
&\leq \mathbb{E}[\mathrm{herdisc}(\Sigma)|E] + 1
\end{aligned}
$$

where we bounded $\mathbb{E}[\mathrm{herdisc}(\Sigma)|\overline{E}]$ by the Beck-Fiala theorem (Theorem 1.1) which holds for any t-sparse set system, and bounded $\Pr[\overline{E}]$ by 2^{-t} according to Lemma 3.1. To conclude the

---

[1]That is, each pair in M is assigned $(+1, -1)$ or $(-1, +1)$ independently with probability $1/2$.

proof we will show that when E holds then $\mathrm{herdisc}(\Sigma) \leq O(\sqrt{t \log t})$. 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. 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, and 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 χ 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 χ over the small sets in Σ. To do so, we choose χ 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 $x \notin \{a_1, b_1, \ldots, a_k, b_k\}$.

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

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

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 constant 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 $S_i' = S_i \cup \{a_j : b_j \in S_i\} \cup \{b_j : a_j \in S_i\}$. Then $|S_i'| \leq 2|S_i| \leq 12t$ and $\chi(S_i)$ is independent of $\chi(x)$ for all $x \notin S_i'$. 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 condition are satisfied as we have $p = 1/100t^2$ and $d = 12t^2$. Hence $\Pr[\wedge \overline{E_i}] > 0$, that is, there exists a coloring χ consistent with M for which $|\chi(S_i)| \leq c\sqrt{t \log t}$ for all small sets S_i. This coloring shows that $\mathrm{disc}(\Sigma') \leq \max(c\sqrt{t \log t}, 2)$ as claimed.

4 Proof of Lemma 3.1

Let (X, Σ) 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 $\deg(x) = t$ for all $x \in X$.

5

Large sets in Σ correspond to the subset of the vertices $V' = \{v \in V : \deg(v) \geq 6t\}$. For a vertex $v \in V$ let $\Gamma(v) \subset X$ 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. Fix a tree T in G' and an arbitrary root $v_T \in V'$ of T. Orient the edges of T from v_T to the leaves. For each $v \in T$ other than the root, choose x_v to be the parent of v in the tree, 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 towards 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 both v_1, v_2 in T. However, by construction in this case $x = x_{v_1} = x_{v_2}$ and hence $x \notin A_{v_1}, A_{v_2}$, from which we conclude that $\{A_v : v \in V'\}$ are pairwise disjoint, as claimed. $\qquad\square$

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

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

Proof. Fix $x_1, x_2 \in X$ and $v_1, v_2 \in V$. 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 Σ). 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. By Lemma 2.3, the probability for this is bounded by

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

6

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

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

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

$\square$

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

Proof. Let $x_1, v_1, \ldots, x_\ell, v_\ell$ denote a potential cycle of length 2ℓ. 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 ℓ). 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 $x \in X_i$.

There 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 $\Pr[\mathrm{girth}(G') = 2\ell]$ 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

$$n^\ell m^\ell \binom{n}{6t-2}^\ell \leq \left(\frac{nm \cdot e^{6t-2} \cdot n^{6t-2}}{(6t-2)^{6t-2}} \right)^\ell \leq \left(\frac{(em)^{6t}}{(6t-2)^{6t-2}} \right)^\ell.$$

Thus,

$$\Pr[\mathrm{girth}(G') = 2\ell] \leq \left(\frac{(em)^{6t}}{(6t-2)^{6t-2}} \right)^\ell \cdot \left(\frac{t}{m} \right)^{6t\ell} = \left((6t-2)^2 \left(\frac{et}{6t-2} \right)^{6t} \right)^\ell \leq \exp(-t\ell).$$

$\square$

Proof of Lemma 4.1. Using Claims 4.3 and 4.4, the probability that $\text{girth}(G') < \infty$ is bounded by:

$$\Pr[\text{girth}(G') < \infty] = \sum_{\ell=2}^{\infty} \Pr[\text{girth}(G') = \ell] \leq t^4 \exp(-t) + \sum_{\ell=3}^{\infty} \exp(-t\ell) \leq 2t^4 \exp(-t).$$

For $t \geq 55$, we have that $\Pr[\text{girth}(G') < \infty] \leq 2^{-t}$. $\qquad\qquad\qquad\qquad\qquad\qquad\square$

5 The regime of large sets

We next prove Theorem 1.4. Let (X, Σ) 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).

Our main result in this section is the following. We will assume throughout 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.

Theorem 5.1. *Let (X, Σ) 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, Σ), and their multiplicity is 7 or higher. In fact, Theorem 1.4 follows from Theorem 5.1 using a straightforward coupon-collector argument [EA61]. 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.

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 $|c_i| \le 7$. 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 $\mathcal{L}$ spanned by the vectors $v_1, \ldots, v_N$, as our correction comes from $2\mathcal{L}$. We show that $\mathcal{L} = \{x \in \mathbb{Z}^m : \sum x_i = 0 \mod t\}$, which was already proved by Wilson [Wil90] in a more general scenario. However, we need an additional property: vectors in $\mathcal{L}$ 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.

5.1   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

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

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

Lemma 5.2. *For any $w \in \mathcal{L}$ there exist $a_1, \ldots, a_N \in \mathbb{Z}$ such that $\sum a_i v_i = w$. Moreover, $|a_i| \le A$ for all $i \in [N]$ where $A = \frac{2\|w\|_1}{\binom{m-2}{t-1}} + 2$.*

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 $[N]$ 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 $\ell \in [L]$ choose $S_\ell \subset [m]$ of size $t-1$. Set $I_\ell = S_\ell \cup \{i_\ell\}$ and $J_\ell = S_\ell \cup \{j_\ell\}$. Identifying $[N]$ with subsets of $[m]$ of size t, we 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_ℓ, we can choose one of $M = \binom{m-2}{t-1}$ many choices. There is a choice for S_ℓ such that both I_ℓ and J_ℓ 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

9

times that each vector participates; this number is $\lceil s/N \rceil \leq \|w\|_1/M + 1$. 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. $\qquad\square$

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) $c_i \equiv b_i \mod 2$.

(ii) $\|\sum c_i v_i\|_\infty \leq 2$.

(iii) $|c_i| \leq 7$ 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 $\mathbb{E}_z[u_j^2] = k_j$. Thus,

$$\mathbb{E}_z[\|u\|_2^2] = \sum k_j \leq Nt.$$

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 $q_i = u_i \mod 2$ and take $w = (u - q)/2$. In order to guarantee that $w \in \mathcal{L}$, namely that $\sum w_i = 0 \mod t$, 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. We obtain a decomposition $w = \sum a_i v_i$. 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

$$A = 2 + \eta \leq 3,$$

where

$$\eta = 2\frac{\|w\|_1}{\binom{m-2}{t-1}} \leq O\left(\frac{\sqrt{mt\binom{m}{t}}}{\binom{m-2}{t-1}}\right) \leq O\left(\frac{m^{3/2}}{\binom{m}{t}^{1/2}}\right) \leq 1,$$

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. $\qquad\square$

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 $\text{sign}(c_i) \in \{-1, +1\}$ 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. $\qquad\square$

References

[AS00]   Noga Alon and Joel H. Spencer. *The probabilistic method.* Wiley-Interscience series in discrete mathematics and optimization. Wiley, New York, Chichester, Weinheim, 2000.

[Ban98] Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

[BF81]   József Beck and Tibor Fiala. "integer-making theorems". *Discrete Applied Mathematics*, 3(1):1–8, 1981.

[Cha00] Bernard Chazelle. *The discrepancy method : randomness and complexity.* Cambridge University Press, Cambridge, New York, 2000.

[EA61]   Paul Erdös and Rnyi Alfrd. "on a classical problem of probability theory". *Magyar Tudomnyos Akadmia Matematikai Kutat Intzetnek Kzlemnyei*, 6:215–220, 1961.

[EL75]   Paul Erdös and Laszlo Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. *Infinite and Finite Sets (to Paul Erdös on his 60th birthday)*, II:609–627, 1975.

[Mat09] Jiri Matoušek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer Science & Business Media, 2009.

[Mos09] Robin A Moser. A constructive proof of the lovász local lemma. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 343–350. ACM, 2009.

[MT10]   Robin A Moser and Gábor Tardos. A constructive proof of the general lovász local lemma. *Journal of the ACM (JACM)*, 57(2):11, 2010.

[Spe85] Joel Spencer. Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289(2):679–706, 1985.

[Wil90] Richard M. Wilson. A diagonal form for the incidence matrices of t-subsets vs. k-subsets. *European Journal of Combinatorics*, 11(6):609–615, 1990.