

# DECOMPOSITIONS OF FINITE HIGH-DIMENSIONAL RANDOM ARRAYS

PANDELIS DODOS, KONSTANTINOS TYROS AND PETROS VALETTAS

ABSTRACT. A  $d$ -dimensional random array on a nonempty set  $I$  is a stochastic process  $\mathbf{X} = \langle X_s : s \in \binom{I}{d} \rangle$  indexed by the set  $\binom{I}{d}$  of all  $d$ -element subsets of  $I$ . We obtain structural decompositions of finite, high-dimensional random arrays whose distribution is invariant under certain symmetries.

Our first main result is a distributional decomposition of finite, (approximately) spreadable, high-dimensional random arrays whose entries take values in a finite set; the two-dimensional case of this result is the finite version of an infinitary decomposition due to Fremlin and Talagrand. Our second main result is a physical decomposition of finite, spreadable, high-dimensional random arrays with square-integrable entries which is the analogue of the Hoeffding/Efron–Stein decomposition. All proofs are effective.

We also present applications of these decompositions in the study of concentration of functions of finite, high-dimensional random arrays.

## CONTENTS

1. Introduction	2
2. Approximation by a random array of lower complexity	8
3. A coding for distributions	15
4. Proofs of Theorems 1.4 and 1.5	18
5. Orbits	28
6. Comparing two-point correlations of spreadable random arrays	29
7. Proof of Theorem 1.6	31
8. Connection with concentration	39
Appendix A. Proof of Lemma 3.4	44
References	45

---

2010 *Mathematics Subject Classification*: 60G07, 60G09, 60G42, 60E15.

*Key words*: exchangeable random arrays, spreadable random arrays, decompositions, martingales.

P.V. is supported by Simons Foundation grant 638224.

## 1. INTRODUCTION

1.1. **Overview.** Our topic is *probability with symmetries*, a classical theme in probability theory which originates in the work of de Finetti and whose basic objects of study are the following classes of stochastic processes.

**Definition 1.1** (Random arrays, and their subarrays). *Let  $d$  be a positive integer, and let  $I$  be a (possibly infinite) set with  $|I| \geq d$ . A  $d$ -dimensional random array on  $I$  is a stochastic process  $\mathbf{X} = \langle X_s : s \in \binom{I}{d} \rangle$  indexed by the set  $\binom{I}{d}$  of all  $d$ -element subsets of  $I$ . If  $J$  is a subset of  $I$  with  $|J| \geq d$ , then the subarray of  $\mathbf{X}$  determined by  $J$  is the  $d$ -dimensional random array  $\mathbf{X}_J := \langle X_s : s \in \binom{J}{d} \rangle$ .*

The infinitary branch of the theory was developed in a series of foundational papers by Aldous [Ald81], Hoover [Hoo79] and Kallenberg [Kal92], with important earlier contributions by Fremlin and Talagrand [FT85]. The subject is presented in the monographs of Aldous [Ald85] and Kallenberg [Kal05]; more recent expositions, which also discuss several applications, are given in [Ald10, Au08, Au13, DJ08].

However, the focus of this paper is on the finitary case which is significantly less developed (see, e.g., [Au13, page 16] for a discussion on this issue). Our motivation stems from certain applications, in particular, from the concentration results obtained in the companion paper [DTV20] which are inherently finitary; we shall comment further on these connections in Section 8.

1.2. **Notions of symmetry.** Arguably, the most well-known notion of symmetry of random arrays is exchangeability. Let  $d$  be a positive integer, and recall that a  $d$ -dimensional random array  $\mathbf{X} = \langle X_s : s \in \binom{I}{d} \rangle$  on a (possibly infinite) set  $I$  is called *exchangeable*<sup>1</sup> if for every (finite) permutation  $\pi$  of  $I$ , the random arrays  $\mathbf{X}$  and  $\mathbf{X}_\pi := \langle X_{\pi(s)} : s \in \binom{I}{d} \rangle$  have the same distribution. Another well-known notion of symmetry, which is weaker<sup>2</sup> than exchangeability, is spreadability: a  $d$ -dimensional random array  $\mathbf{X}$  on  $I$  is called *spreadable*<sup>3</sup> if for every pair  $J, K$  of finite subsets of  $I$  with  $|J| = |K| \geq d$ , the subarrays  $\mathbf{X}_J$  and  $\mathbf{X}_K$  have the same distribution.

Beyond these two notions, in this paper we will consider yet another notion of symmetry which is a natural weakening of spreadability (see also [DTV20, Definition 1.3]).

**Definition 1.2** (Approximate spreadability). *Let  $\mathbf{X}$  be a  $d$ -dimensional random array on a (possibly infinite) set  $I$ , and let  $\eta \geq 0$ . We say that  $\mathbf{X}$  is  $\eta$ -spreadable (or approximately spreadable if  $\eta$  is understood), provided that for every pair  $J, K$  of finite subsets of  $I$  with*

<sup>1</sup>Some authors refer to this notion as *joint exchangeability*.

<sup>2</sup>Actually, the relation between these two notions is more subtle. For infinite sequences of random variables, spreadability coincides with exchangeability (see [Kal05]), but it is a weaker notion for higher-dimensional random arrays.

<sup>3</sup>We note that this is not standard terminology. In particular, in [FT85] spreadable random arrays are referred to as *deletion invariant*, while in [Kal05] they are called *contractable*.

$|J| = |K| \geq d$  we have

$$(1.1) \quad \rho_{\text{TV}}(P_J, P_K) \leq \eta$$

where  $P_J$  and  $P_K$  denote the laws of the random subarrays  $\mathbf{X}_J$  and  $\mathbf{X}_K$  respectively, and  $\rho_{\text{TV}}$  stands for the total variation distance.

The following proposition—whose proof is a fairly straightforward application of Ramsey’s theorem [Ra30]—justifies Definition 1.2 and shows that approximately spreadable random arrays are ubiquitous.

**Proposition 1.3.** *For every triple  $m, n, d$  of positive integers with  $n \geq d$ , and every  $\eta > 0$ , there exists an integer  $N \geq n$  with the following property. If  $\mathcal{X}$  is a set with  $|\mathcal{X}| = m$  and  $\mathbf{X}$  is an  $\mathcal{X}$ -valued,  $d$ -dimensional random array on a set  $I$  with  $|I| \geq N$ , then there exists a subset  $J$  of  $I$  with  $|J| = n$  such that the random array  $\mathbf{X}_J$  is  $\eta$ -spreadable.*

**1.3. Random arrays with finite-valued entries.** Our first main result is a distributional decomposition of finite, approximately spreadable, high-dimensional random arrays whose entries take values in a finite set. In order to state this decomposition we need to recall a canonical way for defining finite-valued spreadable random arrays. In what follows, by  $\mathbb{N} = \{1, 2, \dots\}$  we denote the set of positive integers, and for every positive integer  $n$  we set  $[n] := \{1, \dots, n\}$ .

1.3.1. Let  $\mathcal{X}$  be a finite set; to avoid degenerate cases, we will assume that  $|\mathcal{X}| \geq 2$ . Also let  $d$  be a positive integer, let  $(\Omega, \Sigma, \mu)$  be a probability space, and let  $\Omega^d$  be equipped with the product measure. We say that a collection  $\mathcal{H} = \langle h^a : a \in \mathcal{X} \rangle$  of  $[0, 1]$ -valued random variables on  $\Omega^d$  is an  $\mathcal{X}$ -partition of unity if  $\mathbf{1}_{\Omega^d} = \sum_{a \in \mathcal{X}} h^a$  almost surely. With every  $\mathcal{X}$ -partition of unity  $\mathcal{H}$  we associate an  $\mathcal{X}$ -valued, spreadable,  $d$ -dimensional random array  $\mathbf{X}_{\mathcal{H}} = \langle X_s^{\mathcal{H}} : s \in \binom{\mathbb{N}}{d} \rangle$  on  $\mathbb{N}$  whose distribution<sup>4</sup> satisfies the following: for every nonempty finite subset  $\mathcal{F}$  of  $\binom{\mathbb{N}}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$ , we have

$$(1.2) \quad \mathbb{P}\left(\bigcap_{s \in \mathcal{F}} [X_s^{\mathcal{H}} = a_s]\right) = \int \prod_{s \in \mathcal{F}} h^{a_s}(\boldsymbol{\omega}_s) d\boldsymbol{\mu}(\boldsymbol{\omega})$$

where  $\boldsymbol{\mu}$  stands for the product measure on  $\Omega^{\mathbb{N}}$  and, for every  $s = \{i_1 < \dots < i_d\} \in \binom{\mathbb{N}}{d}$  and every  $\boldsymbol{\omega} = (\omega_i) \in \Omega^{\mathbb{N}}$ , by  $\boldsymbol{\omega}_s = (\omega_{i_1}, \dots, \omega_{i_d}) \in \Omega^d$  we denote the restriction of  $\boldsymbol{\omega}$  on the coordinates determined by  $s$ .

These distributions were considered by Fremlin and Talagrand who showed that if “ $d = 2$ ” and “ $\mathcal{X} = \{0, 1\}$ ”, then they are precisely the extreme points of the compact convex set of all distributions of boolean, spreadable, two-dimensional random arrays on  $\mathbb{N}$ ; see [FT85, Theorem 5H]. This striking probabilistic/geometric fact together with Choquet’s representation theorem yield that the distribution of an arbitrary boolean, spreadable, two-dimensional random array on  $\mathbb{N}$  is a mixture of distributions of the form (1.2).

<sup>4</sup>See [FT85, Section 1G] for a justification of the existence of this random array.

1.3.2. The decomposition alluded to earlier—which applies to any dimension  $d$  and any finite set  $\mathcal{X}$ —is the finite analogue of the Fremlin–Talagrand decomposition. Of course, instead of mixtures, we will consider finite convex combinations. Specifically, let  $J$  be a nonempty finite index set, let  $\boldsymbol{\lambda} = \langle \lambda_j : j \in J \rangle$  be convex coefficients (that is, positive coefficients which sum-up to 1) and let  $\boldsymbol{\mathcal{H}} = \langle \mathcal{H}_j : j \in J \rangle$  be  $\mathcal{X}$ -partitions of unity such that each  $\mathcal{H}_j = \langle h_j^a : a \in \mathcal{X} \rangle$  is defined on  $\Omega_j^d$  where  $\Omega_j$  is the sample space of a probability space  $(\Omega_j, \Sigma_j, \mu_j)$ . Given these data, we define an  $\mathcal{X}$ -valued, spreadable,  $d$ -dimensional random array  $\mathbf{X}_{\boldsymbol{\lambda}, \boldsymbol{\mathcal{H}}} = \langle X_s^{\boldsymbol{\lambda}, \boldsymbol{\mathcal{H}}} : s \in \binom{\mathbb{N}}{d} \rangle$  on  $\mathbb{N}$  whose distribution satisfies

$$(1.3) \quad \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s^{\boldsymbol{\lambda}, \boldsymbol{\mathcal{H}}} = a_s] \right) = \sum_{j \in J} \lambda_j \int \prod_{s \in \mathcal{F}} h_j^{a_s}(\boldsymbol{\omega}_s) d\boldsymbol{\mu}_j(\boldsymbol{\omega})$$

for every nonempty finite subset  $\mathcal{F}$  of  $\binom{\mathbb{N}}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$ .

1.3.3. We are now in a position to state the first main result of this paper.

**Theorem 1.4** (Distributional decomposition). *Let  $d, m, k$  be positive integers with  $m \geq 2$  and  $k \geq d$ , let  $0 < \varepsilon \leq 1$ , and set*

$$(1.4) \quad C = C(d, m, k, \varepsilon) := \exp^{(2d)} \left( \frac{2^8 m^{7k^d}}{\varepsilon^2} \right)$$

where for every positive integer  $\ell$  by  $\exp^{(\ell)}(\cdot)$  we denote the  $\ell$ -th iterated exponential<sup>5</sup>. Also let  $n \geq C$  be an integer, let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $(1/C)$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Then there exist

- two nonempty finite sets  $J$  and  $\Omega$  with  $|J|, |\Omega| \leq C$ ,
- convex coefficients  $\boldsymbol{\lambda} = \langle \lambda_j : j \in J \rangle$ , and
- for every  $j \in J$  a probability measure  $\mu_j$  on the set  $\Omega$  and an  $\mathcal{X}$ -partition of unity  $\mathcal{H}_j = \langle h_j^a : a \in \mathcal{X} \rangle$  defined on  $\Omega^d$

such that, setting  $\boldsymbol{\mathcal{H}} := \langle \mathcal{H}_j : j \in J \rangle$  and letting  $\mathbf{X}_{\boldsymbol{\lambda}, \boldsymbol{\mathcal{H}}}$  be as in (1.3), the following holds. If  $L$  is a subset of  $[n]$  with  $|L| = k$ , and  $P_L$  and  $Q_L$  denote the laws of the subarrays of  $\mathbf{X}$  and  $\mathbf{X}_{\boldsymbol{\lambda}, \boldsymbol{\mathcal{H}}}$  determined by  $L$  respectively, then we have

$$(1.5) \quad \rho_{\text{TV}}(P_L, Q_L) \leq \varepsilon.$$

An immediate consequence<sup>6</sup> of Theorem 1.4 is that every, not too large, subarray of a finite, finite-valued, approximately spreadable random array is “almost extendable” to an infinite spreadable random array.

Closely related to Theorem 1.4 is the following theorem.

**Theorem 1.5.** *Let the parameters  $d, m, k, \varepsilon$  be as in Theorem 1.4, and let the constant  $C = C(d, m, k, \varepsilon)$  be as in (1.4). Also let  $n \geq C$  be an integer, let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ ,*

<sup>5</sup>Thus, we have  $\exp^{(1)}(x) = \exp(x)$ ,  $\exp^{(2)}(x) = \exp(\exp(x))$ ,  $\exp^{(3)}(x) = \exp(\exp(\exp(x)))$ , etc.

<sup>6</sup>This fact can also be proved using an ultraproduct argument but, of course, this sort of reasoning is not effective.

and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $(1/C)$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Then there exists a Borel measurable function  $f: [0, 1]^{d+1} \rightarrow \mathcal{X}$  with the following property. Let  $\mathbf{X}_f = \langle X_s^f : s \in \binom{\mathbb{N}}{d} \rangle$  be the  $\mathcal{X}$ -valued, spreadable,  $d$ -dimensional random array on  $\mathbb{N}$  defined by setting for every  $s = \{i_1 < \dots < i_d\} \in \binom{\mathbb{N}}{d}$ ,

$$(1.6) \quad X_s^f = f(\zeta, \xi_{i_1}, \dots, \xi_{i_d})$$

where  $(\zeta, \xi_1, \dots)$  are i.i.d. random variables uniformly distributed in  $[0, 1]$ . Then, for every subset  $L$  of  $[n]$  with  $|L| = k$ , denoting by  $P_L$  and  $Q_L$  the laws of the subarrays of  $\mathbf{X}$  and  $\mathbf{X}_f$  determined by  $L$  respectively, we have  $\rho_{\text{TV}}(P_L, Q_L) \leq \varepsilon$ .

Theorem 1.5 is akin to the Aldous–Hoover–Kallenberg representation theorem. The main difference is that in Theorem 1.5 the number of variables which are needed in order to represent the random array  $\mathbf{X}$  is  $d + 1$ , while the corresponding number of variables required by the Aldous–Hoover–Kallenberg theorem is  $2^d$ . This particular information is a genuinely finitary phenomenon, and it is important for the results related to concentration which are presented in Section 8.

**1.4. Random arrays with square-integrable entries.** Our second main result is a physical decomposition of finite, spreadable, high-dimensional random arrays with square integrable entries which is in the spirit of the classical Hoeffding/Efron–Stein decomposition [Hoe48, ES81]. It is less informative than Theorem 1.4, but this is offset by the fact that it applies to a fairly large class of distributions (including bounded, gaussian, subgaussian, etc.).

1.4.1. At this point it is appropriate to introduce some terminology and notation which will be used throughout the paper. Given two subsets  $F, L$  of  $\mathbb{N}$ , by  $\text{PartIncr}(F, L)$  we denote the set of strictly increasing partial maps  $p$  whose domain,  $\text{dom}(p)$ , is contained in  $F$  and whose image,  $\text{Im}(p)$ , is contained in  $L$ . (The empty partial map is included in  $\text{PartIncr}(F, L)$ , and it is denoted by  $\emptyset$ .) For every  $p \in \text{PartIncr}(F, L)$  and every subset  $G$  of  $\text{dom}(p)$  by  $p \upharpoonright G \in \text{PartIncr}(F, L)$  we denote the restriction of  $p$  on  $G$ .

Next, let  $p_1, p_2 \in \text{PartIncr}(F, L)$  be distinct partial maps. We say that the pair  $\{p_1, p_2\}$  is *aligned* if there exists a (possibly empty) subset  $G$  of  $\text{dom}(p_1) \cap \text{dom}(p_2)$  such that: (i)  $p_1 \upharpoonright G = p_2 \upharpoonright G$ , and (ii)  $p_1(\text{dom}(p_1) \setminus G) \cap p_2(\text{dom}(p_2) \setminus G) = \emptyset$ . We shall refer to the (necessarily unique) set  $G$  as the *root* of  $\{p_1, p_2\}$  and we shall denote it by  $r(p_1, p_2)$ ; moreover, we set  $p_1 \wedge p_2 := p_1 \upharpoonright r(p_1, p_2) \in \text{PartIncr}(F, L)$ .

1.4.2. Whenever necessary, we identify subsets of  $\mathbb{N}$  with strictly increasing partial maps as follows. Let  $L$  be a nonempty finite subset of  $\mathbb{N}$ , set  $\ell := |L|$ , and let  $\{i_1 < \dots < i_\ell\}$  denote the increasing enumeration of  $L$ . We define the *canonical isomorphism*  $\mathbb{I}_L: [\ell] \rightarrow L$  associated with  $L$  by setting  $\mathbb{I}_L(j) := i_j$  for every  $j \in [\ell]$ . Note that  $\mathbb{I}_L \in \text{PartIncr}([\ell], L)$ .

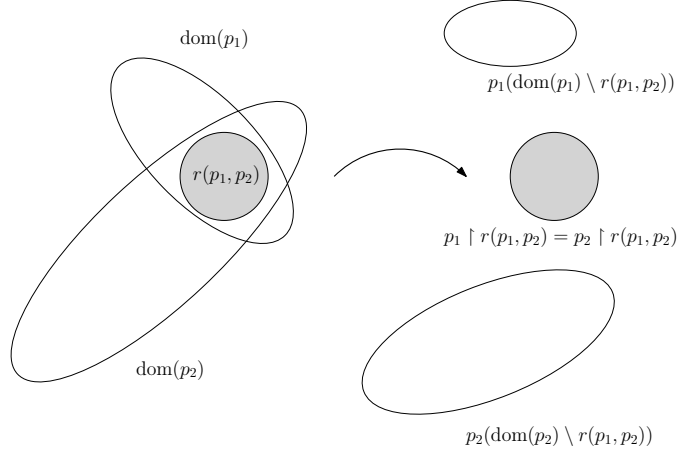


FIGURE 1. Aligned pairs of partial maps.

1.4.3. After this preliminary discussion, and in order to motivate our second decomposition, let us consider the model case of a spreadable,  $d$ -dimensional random array  $\mathbf{X}$  on  $\mathbb{N}$  whose entries are of the form  $X_s = h(\xi_{i_1}, \dots, \xi_{i_d})$  for every  $s = \{i_1 < \dots < i_d\} \in \binom{\mathbb{N}}{d}$ , where  $h: [0, 1]^d \rightarrow [0, 1]$  is Borel measurable and  $(\xi_i)$  are i.i.d. random variables uniformly distributed in  $[0, 1]$ . For every subset  $F$  of  $\mathbb{N}$  let  $\mathcal{A}_F$  denote the  $\sigma$ -algebra generated by the random variables  $\langle \xi_i : i \in F \rangle$ . (In particular,  $\mathcal{A}_\emptyset$  is the trivial  $\sigma$ -algebra.) Since the random variables  $(\xi_i)$  are independent, the  $\sigma$ -algebras  $\langle \mathcal{A}_F : F \subseteq \mathbb{N} \rangle$  generate a lattice of projections: for every pair  $F, G$  of subsets of  $\mathbb{N}$  and every random variable  $Z$  we have  $\mathbb{E}[\mathbb{E}[Z | \mathcal{A}_F] | \mathcal{A}_G] = \mathbb{E}[Z | \mathcal{A}_{F \cap G}]$ . This lattice of projections can be used, in turn, to decompose the random array  $\mathbf{X}$  in a natural (and standard) way.

Specifically, for every  $p \in \text{PartIncr}([d], \mathbb{N})$  we select<sup>7</sup>  $s \in \binom{\mathbb{N}}{d}$  such that  $\mathbf{I}_s \upharpoonright \text{dom}(p) = p$ , and we set  $Y_p := \mathbb{E}[X_s | \mathcal{A}_{\text{dom}(p)}]$ . (Notice that  $Y_p$  is independent of the choice of  $s$ .) Via inclusion-exclusion, the process  $\mathbf{Y} = \langle Y_p : p \in \text{PartIncr}([d], \mathbb{N}) \rangle$  induces the “increments”  $\Delta = \langle \Delta_p : p \in \text{PartIncr}([d], \mathbb{N}) \rangle$  defined by

$$\Delta_p := \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} Y_{p \upharpoonright G}.$$

Then, for every  $s \in \binom{\mathbb{N}}{d}$ , we have

$$X_s = \sum_{F \subseteq [d]} \Delta_{\mathbf{I}_s \upharpoonright F}.$$

More importantly, the fact that the random variables  $(\xi_i)$  are independent yields that if  $p_1, p_2 \in \text{PartIncr}([d], \mathbb{N})$  are distinct and the pair  $\{p_1, p_2\}$  is aligned, then the random variables  $\Delta_{p_1}$  and  $\Delta_{p_2}$  are orthogonal; in particular, we have  $\|X_s\|_{L_2} = \sum_{F \subseteq [d]} \|\Delta_{\mathbf{I}_s \upharpoonright F}\|_{L_2}$ .

<sup>7</sup>Note that this selection is not always possible, but it is certainly possible if  $\text{Im}(p) \subseteq \{d, d+1, \dots\}$ . Here, we ignore this minor technical issue for the sake of exposition.

1.4.4. The following theorem—which is our second main result—shows that an approximate version of the decomposition described above can be obtained in full generality.

**Theorem 1.6** (Physical decomposition). *Let  $d$  be a positive integer, let  $\varepsilon > 0$ , and set*

$$(1.7) \quad c = c(d, \varepsilon) := 2^{-16} \varepsilon^{4/(d+1)}$$

$$(1.8) \quad n_0 = n_0(d, \varepsilon) := 2^{20(d+1)^2} \varepsilon^{-(d+5)}.$$

*Then for every integer  $n \geq n_0$  there exists a subset  $N$  of  $[n]$  with  $|N| \geq c^{d+1} \sqrt{n}$  and satisfying the following property. If  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  is a real-valued, spreadable,  $d$ -dimensional random array on  $[n]$  such that  $\|X_s\|_{L_2} = 1$  for all  $s \in \binom{[n]}{d}$ , then there exists a real-valued stochastic process  $\Delta = \langle \Delta_p : p \in \text{PartIncr}([d], N) \rangle$  such that the following hold true.*

(i) (Decomposition) *For every  $s \in \binom{N}{d}$  we have*

$$(1.9) \quad X_s = \sum_{F \subseteq [d]} \Delta_{I_s \upharpoonright F}.$$

(ii) (Approximate zero mean) *If  $p \in \text{PartIncr}([d], N)$  with  $p \neq \emptyset$ , then*

$$(1.10) \quad |\mathbb{E}[\Delta_p]| \leq \varepsilon.$$

(iii) (Approximate orthogonality) *If  $p_1, p_2 \in \text{PartIncr}([d], N)$  are distinct and the pair  $\{p_1, p_2\}$  is aligned, then*

$$(1.11) \quad |\mathbb{E}[\Delta_{p_1} \Delta_{p_2}]| \leq \varepsilon.$$

(iv) (Uniqueness) *The process  $\Delta$  is unique in the following sense. There exists a subset  $L$  of  $N$  with  $|L| \geq ((\varepsilon^{-1} + 2^{2d})d)^{-1} |N|$  such that for every real-valued stochastic process  $\mathbf{Z} = \langle Z_p : p \in \text{PartIncr}([d], N) \rangle$  which satisfies (i) and (iii) above, we have  $\|\Delta_p - Z_p\|_{L_2} \leq 2^{\binom{d+2}{2}} \sqrt{2\varepsilon}$  for all  $p \in \text{PartIncr}([d], L)$ .*

**1.5. Outline of the proofs/Structure of the paper.** The proofs of Theorems 1.4, 1.5 and 1.6 are a blend of analytic, probabilistic and combinatorial ideas.

1.5.1. The proofs of Theorems 1.4 and 1.5 rely on two preparatory steps which are largely independent of each other and can be read separately.

The first step is to approximate, in distribution, any finite-valued, approximately spreadable random array by a random array of “lower-complexity”. We note that a similar approximation is used in the proof of the Aldous–Hoover theorem; see, *e.g.*, [Au13, Section 5]. However, our argument is technically different since we work with approximately spreadable, instead of exchangeable, random arrays. The details of this approximation are given in Section 2.

The second step, which is presented in Section 3, is a coding lemma for distributions of the form (1.2). It asserts that the laws of their finite subarrays can be approximated, with arbitrary accuracy, by the laws of subarrays of distributions of the form (1.2) which are

generated by genuine partitions instead of partitions of unity. The proof of this coding is based on a random selection of uniform hypergraphs.

Section 4 is devoted to the proofs of Theorems 1.4 and 1.5. We actually prove a slightly stronger result—Theorem 4.1 in the main text—which encompasses both Theorem 1.4 and Theorem 1.5 and it is more amenable to an inductive scheme.

1.5.2. The proof of Theorem 1.6 is somewhat different, and it is based exclusively on  $L_2$  methods. The main goal is to construct an appropriate collection of  $\sigma$ -algebras for which the corresponding projections behave like the lattice of projections described in Paragraph 1.4.3.

This goal boils down to classify all two-point correlations of finite, spreadable random arrays with square-integrable entries. Sections 5 and 6 are devoted to the proof of this classification; we note that this material is of independent interest, and it can also be read independently. The proof of Theorem 1.6 is completed in Section 7.

1.5.3. Finally, as we have already mentioned, in the last section of this paper, Section 8, we present an application of Theorem 1.4 which is related to the concentration results obtained in [DTV20].

## 2. APPROXIMATION BY A RANDOM ARRAY OF LOWER COMPLEXITY

The main result in this section—Proposition 2.1 below—asserts that any large subarray of a finite-valued, approximately spreadable random array  $\mathbf{X}$  can be approximated, in distribution, by a random array which is obtained by projecting the entries of  $\mathbf{X}$  on certain  $\sigma$ -algebras of “lower complexity”.

2.1. **The  $\sigma$ -algebras  $\Sigma(\mathcal{G}_\ell^s, \mathbf{X})$ .** Our first goal is to define the aforementioned  $\sigma$ -algebras. To this end we need to introduce some notation which will be used throughout this section and Section 4. Let  $n \geq d$  be positive integers, and let  $\mathcal{G}$  be a nonempty subset of  $\binom{[n]}{d}$ . For every finite-valued,  $d$ -dimensional random array  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  on  $[n]$  we set

$$(2.1) \quad \Sigma(\mathcal{G}, \mathbf{X}) := \sigma(\{X_s : s \in \mathcal{G}\});$$

that is,  $\Sigma(\mathcal{G}, \mathbf{X})$  denotes the  $\sigma$ -algebra generated by the random variables  $\langle X_s : s \in \mathcal{G} \rangle$ .

Moreover, for every pair  $F = \{i_1 < \dots < i_k\}$  and  $G = \{j_1 < \dots < j_k\}$  of nonempty subsets of  $\mathbb{N}$  with  $|F| = |G| = k$ , we define  $\mathbf{I}_{F,G} : F \rightarrow G$  by setting

$$(2.2) \quad \mathbf{I}_{F,G}(i_r) = j_r$$

for every  $r \in [k]$ . Notice that  $\mathbf{I}_{F,G} = \mathbf{I}_G \circ \mathbf{I}_F^{-1}$  where  $\mathbf{I}_F$  and  $\mathbf{I}_G$  denote the canonical isomorphisms associated with the sets  $F$  and  $G$  respectively. (See Paragraph 1.4.2.)



2.1.1. Next let  $n, d, \ell$  be positive integers with  $n \geq d$ . Also let  $F$  be a nonempty subset of  $[n]$ , and let  $\{j_1 < \dots < j_{|F|}\}$  denote the increasing enumeration of  $F$ . We say that  $F$  is  $\ell$ -sparse provided that

- $\ell \leq \min(F)$ ,
- $\max(F) \leq n - \ell$ , and
- if  $|F| \geq 2$ , then  $j_{i+1} - j_i \geq \ell$  for all  $i \in \{1, \dots, |F| - 1\}$ .

2.1.1.1. Now assume that  $d \geq 2$ . For every  $\ell$ -sparse  $x = \{j_1 < \dots < j_{d-1}\} \in \binom{[n]}{d-1}$  we set

$$(2.3) \quad \mathcal{R}_\ell^x := \binom{\left( \left( \bigcup_{r=1}^{d-1} \{j_r - \ell + 1, \dots, j_r\} \right) \cup \{n - \ell + 1, \dots, n\} \right)}{d}.$$

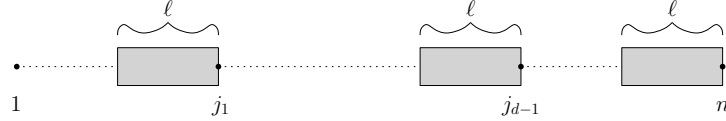


FIGURE 2. The set  $\mathcal{R}_\ell^x$ .

Moreover, for every  $\ell$ -sparse  $s \in \binom{[n]}{d}$  we define

$$(2.4) \quad \mathcal{G}_\ell^s := \bigcup_{x \in \binom{[n]}{d-1}} \mathcal{R}_\ell^x.$$

Finally, if  $\mathbf{X}$  is a finite-valued,  $d$ -dimensional random array on  $[n]$ , then  $\Sigma(\mathcal{G}_\ell^s, \mathbf{X})$  denotes the corresponding  $\sigma$ -algebra defined via formula (2.1); notice that

$$(2.5) \quad \Sigma(\mathcal{G}_\ell^s, \mathbf{X}) = \bigvee_{x \in \binom{[n]}{d-1}} \Sigma(\mathcal{R}_\ell^x, \mathbf{X}).$$

2.1.1.2. If  $d = 1$ , then for every  $\ell$ -sparse  $s \in \binom{[n]}{1}$  we set

$$(2.6) \quad \mathcal{G}_\ell^s := \binom{\{n - \ell + 1, \dots, n\}}{1}.$$

Of course, for every finite-valued,  $d$ -dimensional random array  $\mathbf{X}$  on  $[n]$ , the corresponding  $\sigma$ -algebra  $\Sigma(\mathcal{G}_\ell^s, \mathbf{X})$  is also defined via formula (2.1).

**2.2. The approximation.** We have the following proposition.

**Proposition 2.1.** *Let  $n, d, m, k$  be positive integers with  $k \geq d$  and  $m \geq 2$ , and let  $\theta > 0$ . Assume that*

$$(2.7) \quad n \geq (k + 1)k^{m \lfloor 1/\theta \rfloor + 1}$$

and set  $\ell_0 := k^{m \lfloor 1/\theta \rfloor}$ . Then every  $(k\ell_0)$ -sparse subset  $L$  of  $[n]$  of cardinality  $k$  has the following property. For every set  $\mathcal{X}$  with  $|\mathcal{X}| = m$ , every  $\eta \geq 0$  and every  $\mathcal{X}$ -valued,  $\eta$ -spreadable,  $d$ -dimensional random array  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  on  $[n]$  there exists  $\ell \in [\ell_0]$

such that for every nonempty subset  $\mathcal{F}$  of  $\binom{[L]}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(2.8) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E} [\mathbf{1}_{[X_s = a_s]} \mid \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right] \right| \leq k^d \sqrt{\theta + 15\eta m^{(k\ell_0(d+1))^d}}.$$

The rest of this section is devoted to the proof of Proposition 2.1.

2.2.1. *Step 1.* We start with the following lemma which is a consequence of spreadability.

**Lemma 2.2** (Shift invariance of projections). *Let  $n, d$  be positive integers with  $n \geq d$ , let  $s \in \binom{[n]}{d}$ , and let  $\mathcal{F}$  be a nonempty subset of  $\binom{[n]}{d}$ . Set  $F := s \cup (\cup \mathcal{F})$ . Also let  $G$  be a subset of  $[n]$  with  $|F| = |G|$ , and set*

$$(2.9) \quad t := \mathbf{I}_{F,G}(s) \quad \text{and} \quad \mathcal{G} := \{\mathbf{I}_{F,G}(s) : s \in \mathcal{F}\}.$$

Finally, let  $\mathcal{X}$  be a finite set, let  $\eta \geq 0$ , and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $\eta$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Then for every  $a \in \mathcal{X}$  we have

$$(2.10) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s = a]} \mid \Sigma(\mathcal{F}, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_t = a]} \mid \Sigma(\mathcal{G}, \mathbf{X})] \right\|_{L_2}^2 \right| \leq 5\eta |\mathcal{X}|^{|\mathcal{F}|}.$$

*Proof.* Fix  $a \in \mathcal{X}$ . For every collection  $\mathbf{a} = (a_u)_{u \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we set

$$(2.11) \quad B_{\mathbf{a}} := \bigcap_{u \in \mathcal{F}} [X_u = a_u] \quad \text{and} \quad C_{\mathbf{a}} := \bigcap_{u \in \mathcal{F}} [X_{\mathbf{I}_{F,G}(u)} = a_u].$$

Since the random array  $\mathbf{X}$  is  $\eta$ -spreadable, for every  $\mathbf{a} \in \mathcal{X}^{\mathcal{F}}$  we have

$$(2.12) \quad |\mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}(C_{\mathbf{a}})| \leq \eta \quad \text{and} \quad |\mathbb{P}([X_s = a] \cap B_{\mathbf{a}}) - \mathbb{P}([X_t = a] \cap C_{\mathbf{a}})| \leq \eta.$$

Set  $\mathcal{B} := \{\mathbf{a} \in \mathcal{X}^{\mathcal{F}} : \mathbb{P}(B_{\mathbf{a}}) > 0 \text{ and } \mathbb{P}(C_{\mathbf{a}}) > 0\}$ . By (2.12), for every  $\mathbf{a} \in \mathcal{X}^{\mathcal{F}} \setminus \mathcal{B}$  we have  $\mathbb{P}(B_{\mathbf{a}}) \leq \eta$  and  $\mathbb{P}(C_{\mathbf{a}}) \leq \eta$  and, consequently,

$$(2.13) \quad |\mathbb{P}([X_s = a] \mid B_{\mathbf{a}})^2 \mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}([X_t = a] \mid C_{\mathbf{a}})^2 \mathbb{P}(C_{\mathbf{a}})| \leq 2\eta.$$

Next, let  $\mathbf{a} \in \mathcal{B}$  be arbitrary, and observe that

$$(2.14) \quad \begin{aligned} & |\mathbb{P}([X_s = a] \mid B_{\mathbf{a}}) - \mathbb{P}([X_t = a] \mid C_{\mathbf{a}})| = \\ & = \left| \frac{\mathbb{P}([X_s = a] \cap B_{\mathbf{a}})}{\mathbb{P}(B_{\mathbf{a}})} - \frac{\mathbb{P}([X_t = a] \cap C_{\mathbf{a}})}{\mathbb{P}(C_{\mathbf{a}})} \right| \\ & \leq \frac{1}{\mathbb{P}(B_{\mathbf{a}})} |\mathbb{P}([X_s = a] \cap B_{\mathbf{a}}) - \mathbb{P}([X_t = a] \cap C_{\mathbf{a}})| + \mathbb{P}(C_{\mathbf{a}}) \left| \frac{1}{\mathbb{P}(B_{\mathbf{a}})} - \frac{1}{\mathbb{P}(C_{\mathbf{a}})} \right| \\ & \stackrel{(2.12)}{\leq} \frac{\eta}{\mathbb{P}(B_{\mathbf{a}})} + \frac{1}{\mathbb{P}(B_{\mathbf{a}})} |\mathbb{P}(C_{\mathbf{a}}) - \mathbb{P}(B_{\mathbf{a}})| \stackrel{(2.12)}{\leq} \frac{2\eta}{\mathbb{P}(B_{\mathbf{a}})}. \end{aligned}$$

On the other hand, we have  $\mathbb{P}([X_s = a] \mid B_{\mathbf{a}}) + \mathbb{P}([X_t = a] \mid C_{\mathbf{a}}) \leq 2$  and so, by (2.14),

$$(2.15) \quad \left| \mathbb{P}([X_s = a] \mid B_{\mathbf{a}})^2 - \mathbb{P}([X_t = a] \mid C_{\mathbf{a}})^2 \right| \leq \frac{4\eta}{\mathbb{P}(B_{\mathbf{a}})}.$$

Therefore, for every  $\mathbf{a} \in \mathcal{B}$ ,

$$\begin{aligned}
 (2.16) \quad & \left| \mathbb{P}([X_s = a] | B_{\mathbf{a}})^2 \mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}([X_t = a] | C_{\mathbf{a}})^2 \mathbb{P}(C_{\mathbf{a}}) \right| \leq \\
 & \leq \mathbb{P}(B_{\mathbf{a}}) \left| \mathbb{P}([X_s = a] | B_{\mathbf{a}})^2 - \mathbb{P}([X_t = a] | C_{\mathbf{a}})^2 \right| + \\
 & \quad + \mathbb{P}([X_t = a] | C_{\mathbf{a}})^2 \left| \mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}(C_{\mathbf{a}}) \right| \stackrel{(2.15), (2.12)}{\leq} 5\eta.
 \end{aligned}$$

By (2.13) and (2.16), we conclude that

$$\begin{aligned}
 (2.17) \quad & \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{F}, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}, \mathbf{X})] \right\|_{L_2}^2 \right| = \\
 & = \left| \sum_{\mathbf{a} \in \mathcal{X}^{\mathcal{F}}} \mathbb{P}([X_s = a] | B_{\mathbf{a}})^2 \mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}([X_t = a] | C_{\mathbf{a}})^2 \mathbb{P}(C_{\mathbf{a}}) \right| \leq 5\eta |\mathcal{X}|^{|\mathcal{F}|}
 \end{aligned}$$

as desired.  $\square$

2.2.2. *Step 2.* The next lemma follows from elementary properties of martingale difference sequences.

**Lemma 2.3** (Basic approximation). *Let  $n, d, m, k$  be positive integers with  $k \geq d$  and  $m \geq 2$ , and let  $\theta > 0$ . Assume that*

$$(2.18) \quad n \geq (d+1)k^{m\lceil 1/\theta \rceil + 1}$$

and set  $\ell_0 := k^{m\lceil 1/\theta \rceil}$ . Moreover, let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , let  $\eta \geq 0$ , and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $\eta$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Then for every  $(k\ell_0)$ -sparse  $t \in \binom{[n]}{d}$  there exists  $\ell \in [\ell_0]$  such that for every  $a \in \mathcal{X}$ ,

$$(2.19) \quad \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k\ell}^t, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{\ell}^t, \mathbf{X})] \right\|_{L_2} \leq \sqrt{\theta}.$$

*Proof.* Fix  $t \in \binom{[n]}{d}$  which is  $(k\ell_0)$ -sparse. For every  $a \in \mathcal{X}$  and  $r \in [m\lceil 1/\theta \rceil + 1]$ , we set

$$(2.20) \quad D_r^a := \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k^r}^t, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k^{r-1}}^t, \mathbf{X})].$$

Clearly, it enough to show that there exists  $r \in [m\lceil 1/\theta \rceil + 1]$  such that  $\|D_r^a\|_{L_2} \leq \sqrt{\theta}$  for every  $a \in \mathcal{X}$ . Assume, towards a contradiction, that for every  $r \in [m\lceil 1/\theta \rceil + 1]$  there exists  $a_r \in \mathcal{X}$  such that  $\|D_r^{a_r}\|_{L_2} > \sqrt{\theta}$ . Since  $|\mathcal{X}| = m$ , by the pigeonhole principle, there exist  $b \in \mathcal{X}$  and a subset  $R$  of  $[m\lceil 1/\theta \rceil + 1]$  with  $|R| = \lceil 1/\theta \rceil + 1$  such that  $a_r = b$ , which is equivalent to saying that  $\|D_r^b\|_{L_2} > \sqrt{\theta}$  for every  $r \in R$ . Now, observe that the sequence  $(\mathcal{G}_1^t, \dots, \mathcal{G}_{k^{m\lceil 1/\theta \rceil + 1}}^t)$  is increasing with respect to inclusion, which in turn implies, by (2.1), that the sequence  $(D_1^b, \dots, D_{m\lceil 1/\theta \rceil + 1}^b)$  is a martingale difference sequence. By the contractive property of conditional expectation, we obtain that

$$(2.21) \quad 1 \geq \|\mathbf{1}_{[X_t=b]}\|_{L_2}^2 \geq \sum_{r=1}^{m\lceil 1/\theta \rceil + 1} \|D_r^b\|_{L_2}^2 \geq \sum_{r \in R} \|D_r^b\|_{L_2}^2 > |R|\theta > 1$$

which is clearly a contradiction. The proof is completed.  $\square$

We will need the following consequence of Lemma 2.3.

**Corollary 2.4.** *Let  $n, d, m, k, \ell_0, \mathcal{X}, \eta, \mathbf{X}$  be as in Lemma 2.3. Then there exists  $\ell \in [\ell_0]$  such that for every  $(k\ell_0)$ -sparse  $s \in \binom{[n]}{d}$  and every  $a \in \mathcal{X}$  we have*

$$(2.22) \quad \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2} \leq \sqrt{\theta + 10\eta m^{(k\ell_0(d+1))^d}}.$$

*Proof.* Fix a  $(k\ell_0)$ -sparse  $t \in \binom{[n]}{d}$ . By Lemma 2.3, there exists  $\ell \in [\ell_0]$  such that for every  $a \in \mathcal{X}$  we have

$$(2.23) \quad \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2} \leq \sqrt{\theta}.$$

Since the set  $\mathcal{G}_\ell^t$  is contained in  $\mathcal{G}_{k\ell}^t$ , we see that  $\Sigma(\mathcal{G}_\ell^t, \mathbf{X})$  is a sub- $\sigma$ -algebra of  $\Sigma(\mathcal{G}_{k\ell}^t, \mathbf{X})$ . Hence, by (2.23), for every  $a \in \mathcal{X}$  we have

$$(2.24) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k\ell}^t, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_\ell^t, \mathbf{X})] \right\|_{L_2}^2 \right| \leq \theta.$$

Now let  $s \in \binom{[n]}{d}$  be an arbitrary  $(k\ell_0)$ -sparse subset of  $[n]$ . Set  $F := t \cup (\cup \mathcal{G}_\ell^t)$  and  $G := s \cup (\cup \mathcal{G}_\ell^s)$ , and notice that

$$(2.25) \quad s = \mathbf{I}_{F,G}(t) \quad \text{and} \quad \mathcal{G}_\ell^s = \{ \mathbf{I}_{F,G}(u) : u \in \mathcal{G}_\ell^t \}$$

where  $\mathbf{I}_{F,G}$  is as in (2.2). By Lemma 2.2 and the fact that  $|\mathcal{G}_\ell^t| \leq ((d+1)\ell)^d \leq (k\ell_0(d+1))^d$ , for every  $a \in \mathcal{X}$  we have

$$(2.26) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_\ell^t, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2}^2 \right| \leq 5\eta m^{(k\ell_0(d+1))^d}.$$

With identical arguments we obtain that

$$(2.27) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_t=a]} | \Sigma(\mathcal{G}_{k\ell}^t, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})] \right\|_{L_2}^2 \right| \leq 5\eta m^{(k\ell_0(d+1))^d}.$$

Finally, the fact that  $\mathcal{G}_\ell^s$  is contained in  $\mathcal{G}_{k\ell}^s$  yields that  $\Sigma(\mathcal{G}_\ell^s, \mathbf{X})$  is a sub- $\sigma$ -algebra of  $\Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})$ , and so, for every  $a \in \mathcal{X}$  we have

$$(2.28) \quad \begin{aligned} & \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2}^2 = \\ & \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2}^2 \right|. \end{aligned}$$

The desired estimate (2.22) follows from (2.24), (2.26), (2.27), (2.28) and the triangle inequality.  $\square$

**2.2.3. Step 3.** For the next step of the proof of Proposition 2.1 we need to introduce some auxiliary  $\sigma$ -algebras. Let  $n, d, \ell$  be positive integers with  $n \geq \ell(d+1)$ . Also let  $L$  be an  $\ell$ -sparse subset of  $[n]$  of cardinality at least  $d$ , set  $k := |L|$  and let  $\{i_1 < \dots < i_k\}$  denote the increasing enumeration of  $L$ . Moreover, let  $s = \{i_{l_1} < \dots < i_{l_d}\} \in \binom{[n]}{d}$ . First, we define the following subsets of  $[n]$ .

- (D1) We set  $R_\ell^{s,L,1} := \bigcup_{u=1}^{l_1} \{i_u - \ell + 1, \dots, i_u\}$ .
- (D2) If we have that  $d \geq 2$ , then we set  $R_\ell^{s,L,r} := \bigcup_{u=l_{r-1}+1}^{l_r} \{i_u - \ell + 1, \dots, i_u\}$  for every  $r \in \{2, \dots, d\}$ .
- (D3) If  $l_d < k$ , then we set  $\Delta_\ell^{s,L,n} := \{n - \ell + 1, \dots, n\} \cup \bigcup_{u=l_d+1}^k \{i_u - \ell + 1, \dots, i_u\}$ ; otherwise, we set  $\Delta_\ell^{s,L,n} = \{n - \ell + 1, \dots, n\}$ .

Next, we set

$$(2.29) \quad \mathcal{G}_\ell^{s,L} := \bigcup_{x \in \binom{[d]}{d-1}} \left( \Delta_\ell^{s,L,n} \cup \bigcup_{r \in x} R_\ell^{s,L,r} \right).$$

Finally, for every  $d$ -dimensional random array  $\mathbf{X}$  on  $[n]$  we define the corresponding  $\sigma$ -algebra  $\Sigma(\mathcal{G}_\ell^{s,L}, \mathbf{X})$  via formula (2.1).

We have the following lemma.

**Lemma 2.5** (Absorbtion). *Let  $n, d, m, k$  be positive integers with  $k \geq d$ , and let  $\theta > 0$ . Assume that*

$$(2.30) \quad n \geq (k+1)k^{m\lceil 1/\theta \rceil + 1}$$

and set  $\ell_0 := k^{m\lceil 1/\theta \rceil}$ . Then every  $(k\ell_0)$ -sparse subset  $L$  of  $[n]$  with  $|L| = k$  has the following property. For every set  $\mathcal{X}$  with  $|\mathcal{X}| = m$ , every  $\eta \geq 0$  and every  $\mathcal{X}$ -valued,  $\eta$ -spreadable,  $d$ -dimensional random array  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  on  $[n]$ , there exists  $\ell \in [\ell_0]$  such that for every  $a \in \mathcal{X}$  and every  $s \in \binom{[L]}{d}$  we have

$$(2.31) \quad \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^{s,L}, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2} \leq \sqrt{\theta + 15\eta m^{(k\ell_0(d+1))^d}}.$$

*Proof.* For notational convenience, we will assume that  $d \geq 2$ . The case “ $d = 1$ ” is similar. At any rate, in order to facilitate the reader, we shall indicate the necessary changes.

Let  $L, \mathcal{X}, \eta, \mathbf{X}$  be as in the statement of the lemma. We apply Corollary 2.4 and we obtain  $\ell \in [\ell_0]$  such that for every  $a \in \mathcal{X}$  and every  $(k\ell_0)$ -sparse  $s \in \binom{[n]}{d}$  we have the estimate (2.22). In what follows, this  $\ell$  will be fixed.

Let  $s = \{j_1 < \dots < j_d\} \in \binom{[L]}{d}$  be arbitrary; notice that  $s$  is  $(k\ell_0)$ -sparse. Let  $\Delta, R_1, \dots, R_d$  denote the unique subintervals of  $[n]$  with the following properties.

- We have  $|\Delta| = |\Delta_\ell^{s,L,n}|$  and  $\max(\Delta) = n$ , where  $\Delta_\ell^{s,L,n}$  is as in (D3).
- For every  $r \in [d]$  we have  $|R_r| = |R_\ell^{s,L,r}|$  and  $\max(R_r) = j_r$ , where  $R_\ell^{s,L,1}$  is as in (D1) and  $R_\ell^{s,L,r}$  is as in (D2) if  $r \geq 2$ .

We set

$$(2.32) \quad \mathcal{G} := \bigcup_{x \in \binom{[d]}{d-1}} \left( \Delta \cup \bigcup_{r \in x} R_r \right).$$

(If “ $d = 1$ ”, then we set  $\mathcal{G} := \binom{[n]}{1}$ .) Next, we define  $\Delta_\ell := \{n - \ell + 1, \dots, n\}$  and  $\Delta_{k\ell} := \{n - k\ell + 1, \dots, n\}$ ; moreover, for every  $r \in [d]$  we set  $R_\ell^r := \{j_r - \ell + 1, \dots, j_r\}$  and  $R_{k\ell}^r := \{j_r - k\ell + 1, \dots, j_r\}$ . With these choices, by (2.4) and (2.29), we have

$$(2.33) \quad \mathcal{G}_\ell^s = \bigcup_{x \in \binom{[d]}{d-1}} \left( \Delta_\ell \cup \bigcup_{r \in x} R_\ell^r \right) \quad \text{and} \quad \mathcal{G}_{k\ell}^s = \bigcup_{x \in \binom{[d]}{d-1}} \left( \Delta_{k\ell} \cup \bigcup_{r \in x} R_{k\ell}^r \right).$$

(If “ $d = 1$ ”, then we have  $\mathcal{G}_\ell^s = \binom{[n]}{1}$  and  $\mathcal{G}_{k\ell}^s = \binom{[n]}{1}$ .) Observing that

$$(2.34) \quad \ell \leq |\Delta_\ell^{s,L,n}|, |R_\ell^{s,L,1}|, \dots, |R_\ell^{s,L,d}| \leq |L|\ell = k\ell \leq k\ell_0,$$

we see that  $\Delta \subseteq \Delta_{k\ell}$  and  $R_r \subseteq R_{k\ell}^r$  for every  $r \in [d]$ , and moreover,  $\Delta_\ell \subseteq \Delta$  and  $R_\ell^r \subseteq R_r$  for every  $r \in [d]$ . By (2.32) and (2.33), we obtain that  $\mathcal{G}_\ell^s \subseteq \mathcal{G} \subseteq \mathcal{G}_{k\ell}^s$  which, in turn, implies that

$$(2.35) \quad \Sigma(\mathcal{G}_\ell^s, \mathbf{X}) \subseteq \Sigma(\mathcal{G}, \mathbf{X}) \subseteq \Sigma(\mathcal{G}_{k\ell}^s, \mathbf{X}).$$

By (2.22) and (2.35), for every  $a \in \mathcal{X}$  we have

$$(2.36) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2}^2 \right| \leq \theta + 10\eta m^{(k\ell_0(d+1))^d}.$$

On the other hand, setting  $F := \Delta \cup \bigcup_{j=1}^d R_j$  and  $G := \Delta_\ell^{s,L,n} \cup \bigcup_{r=1}^d R_\ell^{s,L,r}$ , we have

$$(2.37) \quad s = \mathbf{I}_{F,G}(s) \quad \text{and} \quad \mathcal{G}_\ell^{s,L} = \{ \mathbf{I}_{F,G}(t) : t \in \mathcal{G} \}$$

where  $\mathbf{I}_{F,G}$  is as in (2.2). By Lemma 2.2 and the fact that  $|\mathcal{G}| \leq ((k+1)\ell)^d \leq (k\ell_0(d+1))^d$ , for every  $a \in \mathcal{X}$  we have

$$(2.38) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}, \mathbf{X})] \right\|_{L_2}^2 - \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^{s,L}, \mathbf{X})] \right\|_{L_2}^2 \right| \leq 5\eta m^{(k\ell_0(d+1))^d}.$$

Finally recall that, by (2.35), we have  $\Sigma(\mathcal{G}_\ell^s, \mathbf{X}) \subseteq \Sigma(\mathcal{G}_\ell^{s,L}, \mathbf{X})$ . Therefore, the estimate (2.31) follows from (2.36), (2.38) and the triangle inequality.  $\square$

**2.2.4. Completion of the proof.** For every positive integer  $d$  let  $<_{\text{lex}}$  denote the lexicographical order on  $\binom{[d]}{n}$ . Specifically, for every distinct  $s = \{i_1 < \dots < i_d\} \in \binom{[d]}{n}$  and  $t = \{j_1 < \dots < j_d\} \in \binom{[d]}{n}$ , setting  $r_0 := \min \{r \in [d] : i_r \neq j_r\}$ , we have

$$(2.39) \quad s <_{\text{lex}} t \Leftrightarrow i_{r_0} < j_{r_0}.$$

We also isolate, for future use, the following fact. Although it is an elementary observation which follows readily from the relevant definitions, it is quite crucial for the proof of Proposition 2.1 and, to a large extent, it justifies the definition of the families of sets in (2.4) and (2.29).

**Fact 2.6.** *Let  $n, d, \ell$  be positive integers with  $n \geq \ell(d+1)$ . Also let  $L$  be an  $\ell$ -sparse subset of  $[n]$  with  $|L| \geq d$ . Then the following hold.*

- (i) *For every  $s, t \in \binom{L}{d}$  we have  $\mathcal{G}_\ell^s \subseteq \mathcal{G}_\ell^{t,L}$ .*
- (ii) *For every  $s \in \binom{L}{d}$  we have  $\{t \in \binom{L}{d} : s <_{\text{lex}} t\} \subseteq \mathcal{G}_\ell^{s,L}$ .*

We are now ready to give the proof of Proposition 2.1.

*Proof of Proposition 2.1.* Fix a  $(k\ell_0)$ -sparse subset  $L$  of  $[n]$  of cardinality  $k$ , and let  $\mathcal{X}, \eta, \mathbf{X}$  be as in the statement of the proposition. By Lemma 2.5, there exists  $\ell \in [\ell_0]$  such that for every  $a \in \mathcal{X}$  and every  $s \in \binom{L}{d}$  we have

$$(2.40) \quad \left| \left\| \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^{s,L}, \mathbf{X})] - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right\|_{L_2} \right| \leq \sqrt{\theta + 15\eta m^{(k\ell_0(d+1))^d}}.$$

We claim that  $\ell$  is as desired.

Indeed, let  $\mathcal{F}$  be subset of  $\binom{L}{d}$  and let  $(a_s)_{s \in \mathcal{F}}$  be a collection of elements of  $\mathcal{X}$ . Set  $\kappa := |\mathcal{F}|$  and let  $\{s_1 <_{\text{lex}} \dots <_{\text{lex}} s_\kappa\}$  denote the lexicographical increasing enumeration

of  $\mathcal{F}$ . Notice that  $\kappa = |\mathcal{F}| \leq \binom{L}{d} \leq k^d$ . Thus, in order to verify (2.8), by a telescopic argument, it is enough to show that for every  $r \in [\kappa]$  we have

$$(2.41) \quad \left| \mathbb{E} \left[ \left( \prod_{i=1}^{r-1} \mathbb{E} [\mathbf{1}_{[X_{s_i}=a_{s_i}]} | \Sigma(\mathcal{G}_\ell^{s_i}, \mathbf{X})] \right) \cdot \left( \prod_{i=r}^{\kappa} \mathbf{1}_{[X_{s_i}=a_{s_i}]} \right) \right] - \mathbb{E} \left[ \left( \prod_{i=1}^r \mathbb{E} [\mathbf{1}_{[X_{s_i}=a_{s_i}]} | \Sigma(\mathcal{G}_\ell^{s_i}, \mathbf{X})] \right) \cdot \left( \prod_{i=r+1}^{\kappa} \mathbf{1}_{[X_{s_i}=a_{s_i}]} \right) \right] \right| \leq \sqrt{\theta + 15\eta m^{(k\ell_0(d+1))^d}}.$$

(Here, we use the convention that the product of an empty family of functions is equal to the constant function 1.) So, fix  $r \in [\kappa]$ . By Fact 2.6, we see that

$$(2.42) \quad \begin{aligned} & \mathbb{E} \left[ \left( \prod_{i=1}^{r-1} \mathbb{E} [\mathbf{1}_{[X_{s_i}=a_{s_i}]} | \Sigma(\mathcal{G}_\ell^{s_i}, \mathbf{X})] \right) \cdot \left( \prod_{i=r}^{\kappa} \mathbf{1}_{[X_{s_i}=a_{s_i}]} \right) \right] = \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \left( \prod_{i=1}^{r-1} \mathbb{E} [\mathbf{1}_{[X_{s_i}=a_{s_i}]} | \Sigma(\mathcal{G}_\ell^{s_i}, \mathbf{X})] \right) \cdot \left( \prod_{i=r}^{\kappa} \mathbf{1}_{[X_{s_i}=a_{s_i}]} \right) \middle| \Sigma(\mathcal{G}_\ell^{s_r, L}, \mathbf{X}) \right] \right] = \\ &= \mathbb{E} \left[ \left( \prod_{i=1}^{r-1} \mathbb{E} [\mathbf{1}_{[X_{s_i}=a_{s_i}]} | \Sigma(\mathcal{G}_\ell^{s_i}, \mathbf{X})] \right) \cdot \mathbb{E} [\mathbf{1}_{[X_{s_r}=a_{s_r}]} | \Sigma(\mathcal{G}_\ell^{s_r, L}, \mathbf{X})] \cdot \left( \prod_{i=r+1}^{\kappa} \mathbf{1}_{[X_{s_i}=a_{s_i}]} \right) \right]. \end{aligned}$$

Inequality (2.41) follows from (2.40), (2.42) and the Cauchy–Schwarz inequality. The proof of Proposition 2.1 is completed.  $\square$

### 3. A CODING FOR DISTRIBUTIONS

The following proposition is the main result in this section.

**Proposition 3.1.** *Let  $d, m, \kappa_0$  be positive integers with  $d, m \geq 2$ , let  $\varepsilon > 0$ , and set*

$$(3.1) \quad u_0 = u_0(d, m, \kappa_0, \varepsilon) := 5d^2 d! m \kappa_0^{2^{d+1}} \varepsilon^{-2^{d+1}}.$$

*Let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , and let  $\mathcal{H} = \langle h^a : a \in \mathcal{X} \rangle$  be an  $\mathcal{X}$ -partition of unity defined on  $\mathcal{Y}^d$  where  $(\mathcal{Y}, \nu)$  is a finite probability space. (See Paragraph 1.3.1.) Then there exists a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $(\mathcal{Y} \times [u_0])^d$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{\mathbb{N}}{d}$  with  $|\mathcal{F}| \leq \kappa_0$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have*

$$(3.2) \quad \left| \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_s) d\nu(\mathbf{y}) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_s) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \varepsilon$$

*where: (i)  $\nu$  denotes the product measure on  $\mathcal{Y}^{\mathbb{N}}$  obtained by equipping each factor with the measure  $\nu$ , (ii)  $\boldsymbol{\mu}$  denotes the product measure on  $(\mathcal{Y} \times [u_0])^{\mathbb{N}}$  obtained by equipping each factor with the product of  $\nu$  and the uniform probability measure on  $[u_0]$ , and (iii) for every  $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ , every  $\boldsymbol{\omega} \in (\mathcal{Y} \times [u_0])^{\mathbb{N}}$  and every  $s \in \binom{\mathbb{N}}{d}$  by  $\mathbf{y}_s$  and  $\boldsymbol{\omega}_s$  we denote the restrictions on the coordinates determined by  $s$  of  $\mathbf{y}$  and  $\boldsymbol{\omega}$  respectively. (See also Paragraph 1.3.1.)*

Proposition 3.1 immediately yields the following corollary.

**Corollary 3.2.** *Let  $d, m, k \geq 2$  be integers with  $k \geq d$ , let  $\varepsilon > 0$ , and set*

$$(3.3) \quad u'_0 = u'_0(d, m, k, \varepsilon) := m^{4k^d+1} k^{2^{d+1}} \varepsilon^{-2^{d+1}}.$$

*Let  $\mathcal{X}, \mathcal{H}, (\mathcal{Y}, \nu)$  be as in Proposition 3.1. Then there exists a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $(\mathcal{Y} \times [u'_0])^d$  with the following property. Set  $\mathcal{E} := \langle \mathbf{1}_{E^a} : a \in \mathcal{X} \rangle$ , and let  $\mathbf{X}^{\mathcal{H}}$  and  $\mathbf{X}^{\mathcal{E}}$  denote the spreadable,  $d$ -dimensional random arrays on  $\mathbb{N}$  defined via (1.2) for  $\mathcal{H}$  and  $\mathcal{E}$  respectively. Then for every subset  $L$  of  $\mathbb{N}$  of cardinality at most  $k$  we have*

$$(3.4) \quad \rho_{\text{TV}}(P_L, Q_L) \leq \varepsilon$$

*where  $P_L$  and  $Q_L$  denote the laws of the subarrays  $\mathbf{X}^{\mathcal{H}}$  and  $\mathbf{X}^{\mathcal{E}}$  determined by  $L$  respectively.*

Corollary 3.2 asserts that the finite pieces of all<sup>8</sup> distributions of the form (1.2) are essentially generated by genuine partitions instead of partitions of unity. Besides its intrinsic interest, this information is important for the proof of Theorem 1.4.

The rest of this section is devoted to the proof of Proposition 3.1. We start by presenting some preparatory material.

**3.1. Box norms.** We will use below—as well as in Section 8—the box norms introduced by Gowers [Go07]. We shall recall the definition of these norms and a couple of their basic properties; for proofs, and a more complete presentation, we refer to [GT10, Appendix B] and [DKK20, Section 2].

Let  $d \geq 2$  be an integer, let  $(\Omega, \Sigma, \mu)$  be a probability space, and let  $\Omega^d$  be equipped with the product measure. For every integrable random variable  $h : \Omega^d \rightarrow \mathbb{R}$  we define its *box norm*  $\|h\|_{\square}$  by setting

$$(3.5) \quad \|h\|_{\square} := \left( \int \prod_{\epsilon \in \{0,1\}^d} h(\omega_{\epsilon}) d\mu(\omega) \right)^{1/2^d}$$

where  $\mu$  denotes the product measure on  $\Omega^{2^d}$  and, for every  $\omega = (\omega_1^0, \omega_1^1, \dots, \omega_d^0, \omega_d^1) \in \Omega^{2^d}$  and every  $\epsilon = (\epsilon_1, \dots, \epsilon_d) \in \{0,1\}^d$  we have  $\omega_{\epsilon} := (\omega_1^{\epsilon_1}, \dots, \omega_d^{\epsilon_d}) \in \Omega^d$ ; by convention, we set  $\|h\|_{\square} := +\infty$  if the integral in (3.5) does not exist.

The quantity  $\|\cdot\|_{\square}$  is a norm on the vector space  $\{h \in L_1 : \|h\|_{\square} < +\infty\}$ , and it satisfies the following Hölder-type inequality, known as the *Gowers–Cauchy–Schwarz inequality*: for every collection  $\langle h_{\epsilon} : \epsilon \in \{0,1\}^d \rangle$  of integrable random variables on  $\Omega^d$  we have

$$(3.6) \quad \left| \int \prod_{\epsilon \in \{0,1\}^d} h_{\epsilon}(\omega_{\epsilon}) d\mu(\omega) \right| \leq \prod_{\epsilon \in \{0,1\}^d} \|h_{\epsilon}\|_{\square}.$$

We will need the following simple fact which follows from Fubini's theorem and the Gowers–Cauchy–Schwarz inequality.

---

<sup>8</sup>We have stated Proposition 3.1 and Corollary 3.2 for finite probability spaces mainly because this is the context of Theorem 1.4. But of course, by an approximation argument, one easily sees that these results hold true in full generality.



**Fact 3.3.** *Let  $(\Omega, \Sigma, \mu)$  be a probability space, and let  $\boldsymbol{\mu}$  denote the product measure on  $\Omega^{\mathbb{N}}$ . Let  $d, k$  be positive integers with  $d \geq 2$ , and let  $f, g, h_1, \dots, h_k: \Omega^d \rightarrow [-1, 1]$  be random variables. Also let  $s_0, s_1, \dots, s_k \in \binom{\mathbb{N}}{d}$  with  $s_0 \neq s_i$  for every  $i \in [k]$ . Then,*

$$(3.7) \quad \left| \int (f(\boldsymbol{\omega}_{s_0}) - g(\boldsymbol{\omega}_{s_0})) \prod_{i=1}^k h_i(\boldsymbol{\omega}_{s_i}) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \|f - g\|_{\square}.$$

(Here, we follow the notational conventions in Paragraph 1.3.1.)

**3.2. Random selection.** We will also need the following lemma.

**Lemma 3.4.** *Let  $d, m \geq 2$  be integers, let  $\varepsilon > 0$ , and set*

$$(3.8) \quad n_0 = n_0(d, m, \varepsilon) := 5d^2 d! m \varepsilon^{-2^{d+1}}.$$

*Also let  $\lambda_1, \dots, \lambda_m \geq 0$  such that  $\lambda_1 + \dots + \lambda_m = 1$ . Then for every finite set  $V$  with  $|V| \geq n_0$  there exists a partition  $\langle E_1, \dots, E_m \rangle$  of  $V^d$  into nonempty symmetric<sup>9</sup> sets such that  $\|\mathbf{1}_{E_j} - \lambda_j\|_{\square} \leq \varepsilon$  for every  $j \in [m]$ . (Here, we view  $V$  as a probability space equipped with the uniform probability measure.)*

Lemma 3.4 is based on a (standard) random selection and the bounded differences inequality. We present the details in Appendix A.

**3.3. Proof of Proposition 3.1.** Let  $\mathcal{X}, \mathcal{H} = \langle h^a : a \in \mathcal{X} \rangle, (\mathcal{Y}, \nu)$  be as in the statement of the proposition. Without loss of generality, we may assume that  $\nu(y) > 0$  for every  $y \in \mathcal{Y}$ , and consequently, we have  $\sum_{a \in \mathcal{X}} h^a(\mathbf{y}) = 1$  for every  $\mathbf{y} \in \mathcal{Y}^d$ . By Lemma 3.4 and the choice of  $u_0$  in (3.1), for every  $\mathbf{y} \in \mathcal{Y}^d$  there exists a partition  $\langle E_{\mathbf{y}}^a : a \in \mathcal{X} \rangle$  of  $[u_0]^d$  such that for every  $a \in \mathcal{X}$  we have

$$(3.9) \quad \|\mathbf{1}_{E_{\mathbf{y}}^a} - h^a(\mathbf{y})\|_{\square} \leq \frac{\varepsilon}{\kappa_0}.$$

For every  $a \in \mathcal{X}$  we set

$$(3.10) \quad E^a := \bigcup_{\mathbf{y} \in \mathcal{Y}^d} \{\mathbf{y}\} \times E_{\mathbf{y}}^a,$$

and we observe that the family  $\langle E^a : a \in \mathcal{X} \rangle$  is a partition of  $(\mathcal{Y} \times [u_0])^d$  into nonempty sets. We claim that this partition  $\langle E^a : a \in \mathcal{X} \rangle$  is as desired.

Indeed, let  $\mathcal{F}$  be a nonempty subset of  $\binom{\mathbb{N}}{d}$  with  $|\mathcal{F}| \leq \kappa_0$  and let  $(a_s)_{s \in \mathcal{F}}$  be a collection of elements of  $\mathcal{X}$ . Set  $\kappa := |\mathcal{F}|$ , and let  $\{s_1, \dots, s_{\kappa}\}$  be an enumeration of  $\mathcal{F}$ . Also let  $\boldsymbol{\lambda}$  denote the product measure on  $[u_0]^{\mathbb{N}}$  obtained by equipping each factor with the uniform

<sup>9</sup>Recall that a subset  $E$  of a Cartesian product  $V^d$  is called *symmetric* if for every  $(v_1, \dots, v_d) \in V^d$  and every permutation  $\pi$  of  $[d]$  we have that  $(v_1, \dots, v_d) \in E$  if and only if  $(v_{\pi(1)}, \dots, v_{\pi(d)}) \in E$ ; in particular, for any symmetric set  $E$ , the set  $\{(v_1, \dots, v_d) \in E : v_1, \dots, v_d \text{ are mutually distinct}\}$  can be identified with a  $d$ -uniform hypergraph on  $V$ .

probability measure. First observe that, by Fact 3.3 and (3.9), for every  $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$  and every  $j \in [\kappa]$  we have

$$(3.11) \quad \left| \int \left( \prod_{i < j} h^{a_{s_i}}(\mathbf{y}_{s_i}) \right) \cdot (h^{a_{s_j}}(\mathbf{y}_{s_j}) - \mathbf{1}_{E_{\mathbf{y}_{s_j}}^{a_{s_j}}}(\mathbf{z}_{s_j})) \cdot \left( \prod_{i > j} \mathbf{1}_{E_{\mathbf{y}_{s_i}}^{a_{s_i}}}(\mathbf{z}_{s_i}) \right) d\lambda(\mathbf{z}) \right| \leq \frac{\varepsilon}{\kappa_0}$$

where, as in Section 2, we use the convention that the product of an empty family of functions is equal to the constant function 1. Hence, by a telescopic argument and the fact that  $\kappa \leq \kappa_0$ , we obtain that for every  $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ ,

$$(3.12) \quad \left| \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_s) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E_{\mathbf{y}_s}^{a_s}}(\mathbf{z}_s) d\lambda(\mathbf{z}) \right| \leq \varepsilon.$$

On the other hand, by Fubini's theorem, we have

$$(3.13) \quad \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\omega_s) d\mu(\omega) = \int \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E_{\mathbf{y}_s}^{a_s}}(\mathbf{z}_s) d\lambda(\mathbf{z}) d\nu(\mathbf{y}).$$

Therefore, (3.2) follows from (3.12) and (3.13). The proof of Proposition 3.1 is completed.

#### 4. PROOFS OF THEOREMS 1.4 AND 1.5

In this section we present the proofs of Theorems 1.4 and 1.5. As already noted, we will actually prove a slightly stronger theorem—Theorem 4.1 below—whose proof occupies Subsections 4.1 up to 4.6. The deduction of Theorems 1.4 and 1.5 from Theorem 4.1 is given in Subsection 4.7.

**4.1. Initializing various numerical invariants.** We start by introducing some numerical invariants. The reader is advised to skip this section at first reading.

4.1.1. First, we define  $\theta: \mathbb{N}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ ,  $\ell_0: \mathbb{N}^3 \times \mathbb{R}^+ \rightarrow \mathbb{N}$ ,  $\bar{m}: \mathbb{N}^3 \times \mathbb{R}^+ \rightarrow \mathbb{N}$  and  $\bar{\varepsilon}: \mathbb{N}^3 \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  by setting

$$(4.1) \quad \theta(d, k, \varepsilon) := \frac{\varepsilon^2}{27 \cdot k^{2d}}$$

$$(4.2) \quad \ell_0(d, m, k, \varepsilon) := k^{m \lfloor 1/\theta(d, k, \varepsilon) \rfloor}$$

$$(4.3) \quad \bar{m}(d, m, k, \varepsilon) := m^{\ell_0(d, m, k, \varepsilon)}$$

$$(4.4) \quad \bar{\varepsilon}(d, m, k, \varepsilon) := \frac{\varepsilon}{8} \bar{m}(d, m, k, \varepsilon)^{-k^{d-1}}.$$

4.1.2. By recursion on  $d$ , for every pair  $m, k$  of positive integers with  $k \geq d$  and every  $\varepsilon > 0$ , we define the quantities  $\eta(d, m, k, \varepsilon)$ ,  $n_0(d, m, k, \varepsilon)$  and  $v(d, m, k, \varepsilon)$ . For “ $d = 1$ ” we set

$$(4.5) \quad \eta(1, m, k, \varepsilon) := \frac{\varepsilon^3}{2^{12} \cdot k^3} m^{-3\ell_0(1, m, k, \varepsilon)}$$

$$(4.6) \quad n_0(1, m, k, \varepsilon) := (k+1)k \ell_0(1, m, k, \varepsilon)$$

$$(4.7) \quad v(1, m, k, \varepsilon) := 2^{22} m^{\ell_0(1, m, k, \varepsilon) + 1} k^8 \varepsilon^{-8}.$$

Next, let  $d \geq 2$  be an integer and assume that  $\eta(d-1, m, k, \varepsilon)$ ,  $n_0(d-1, m, k, \varepsilon)$  and  $v(d-1, m, k, \varepsilon)$  have been defined for every choice of admissible parameters. For notational simplicity set  $\bar{m} := \bar{m}(d, m, k, \varepsilon)$  and  $\bar{\varepsilon} := \bar{\varepsilon}(d, m, k, \varepsilon)$ , and define

$$(4.8) \quad \eta(d, m, k, \varepsilon) := \min \left\{ \frac{\varepsilon^3}{2^{12} \cdot k^{3d}} m^{-\left(k(d+1)\ell_0(d, m, k, \varepsilon)\right)^d}, \eta(d-1, \bar{m}, k, \bar{\varepsilon}) \right\}$$

$$(4.9) \quad n_0(d, m, k, \varepsilon) := k \ell_0(d, m, k, \varepsilon) \cdot (n_0(d-1, \bar{m}, k, \bar{\varepsilon}) + 1)$$

$$(4.10) \quad v(d, m, k, \varepsilon) := 4^{24} m k^{d2^{d+2}} \varepsilon^{-2^{d+2}} v(d-1, \bar{m}, k, \bar{\varepsilon}).$$

**4.2. The main result.** We are ready to state the main result in this section.

**Theorem 4.1.** *Let  $d, m, k$  be positive integers with  $m \geq 2$  and  $k \geq d$ , let  $\varepsilon > 0$ , and let  $\eta(d, m, k, \varepsilon)$ ,  $n_0(d, m, k, \varepsilon)$  and  $v(d, m, k, \varepsilon)$  be the quantities defined in Subsection 4.1. Also let  $n \geq n_0(d, m, k, \varepsilon)$  be a positive integer, let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $\eta(d, m, k, \varepsilon)$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Then there exist a finite probability space  $(\Omega, \mu)$  with  $|\Omega| \leq v(d, m, k, \varepsilon)$  and a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $\Omega^{\{0\} \cup [d]}$  such that for every  $M \in \binom{[n]}{k}$ , every nonempty subset  $\mathcal{F}$  of  $\binom{M}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have*

$$(4.11) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\omega_{\{0\} \cup s}) d\mu(\omega) \right| \leq \varepsilon$$

where  $\mu$  denotes the product measure on  $\Omega^{\{0\} \cup \mathbb{N}}$  and, for every  $s = \{j_1 < \dots < j_d\} \in \binom{[n]}{d}$  and every  $\omega = (\omega_i)_{i \in \{0\} \cup \mathbb{N}} \in \Omega^{\{0\} \cup \mathbb{N}}$ , by  $\omega_{\{0\} \cup s} := (\omega_0, \omega_{j_1}, \dots, \omega_{j_d})$  we denote the restriction of  $\omega$  on the coordinates determined by  $\{0\} \cup s$ .

**4.3. Toolbox.** Our next goal is to collect some preliminary results which are part of the proof of Theorem 4.1, but they are not related with the main argument. Specifically, we have the following lemma.

**Lemma 4.2.** *Let  $n, d, m, \ell$  be positive integers with  $m \geq 2$  and  $n \geq (d+1)\ell$ , and let  $s, t \in \binom{[n]}{d}$  be  $\ell$ -sparse. (See Paragraph 2.1.1.) Also let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , let  $\eta \geq 0$ , and let  $\mathbf{X} = \langle X_u : u \in \binom{[n]}{d} \rangle$  be an  $\mathcal{X}$ -valued,  $\eta$ -spreadable,  $d$ -dimensional random array on  $[n]$ . Set  $F := s \cup (\cup \mathcal{G}_\ell^s)$  and  $G := t \cup (\cup \mathcal{G}_\ell^t)$  where  $\mathcal{G}_\ell^s$  and  $\mathcal{G}_\ell^t$  are as Subsection 2.1. Moreover, for every collection  $\mathbf{a} = (a_u)_{u \in \mathcal{G}_\ell^s}$  of elements of  $\mathcal{X}$  set*

$$(4.12) \quad B_{\mathbf{a}} := \bigcap_{u \in \mathcal{G}_\ell^s} [X_u = a_u] \quad \text{and} \quad C_{\mathbf{a}} := \bigcap_{u \in \mathcal{G}_\ell^t} [X_{I_{F,G}(u)} = a_u]$$

where  $I_{F,G}$  is as in (2.2). (Note that  $I_{F,G}(s) = t$ .) Finally, for every  $a \in \mathcal{X}$  define

$$(4.13) \quad f_a := \sum_{\mathbf{a} \in \mathcal{X}^{\mathcal{G}_\ell^s}} \mathbb{P}([X_t = a] | C_{\mathbf{a}}) \mathbf{1}_{B_{\mathbf{a}}}.$$

Then, for every  $a \in \mathcal{X}$  we have

$$(4.14) \quad \|f_a - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})]\|_{L_2} \leq 2\sqrt{\eta^{2/3} m^{\ell(d+1)^d}}.$$

*Proof.* Observe that for every  $a \in \mathcal{X}$  we have

$$(4.15) \quad f_a - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] = \sum_{\mathbf{a} \in \mathcal{X}^{\mathcal{G}_\ell^s}} \left( \mathbb{P}([X_t = a] | C_{\mathbf{a}}) - \mathbb{P}([X_s = a] | B_{\mathbf{a}}) \right) \mathbf{1}_{B_{\mathbf{a}}}.$$

Hence, if  $\eta = 0$ , then (4.14) follows immediately by (4.15) and the spreadability of  $\mathbf{X}$ . So in what follows we may assume that  $\eta > 0$ . Set  $\mathcal{A} := \{\mathbf{a} \in \mathcal{X}^{\mathcal{G}_\ell^s} : \mathbb{P}(B_{\mathbf{a}}) \leq \eta^{2/3}\}$  and  $\mathcal{B} := \mathcal{X}^{\mathcal{G}_\ell^s} \setminus \mathcal{A}$ . Notice that for every  $a \in \mathcal{X}$  and every  $\mathbf{a} \in \mathcal{A}$  we have the trivial estimate

$$(4.16) \quad |\mathbb{P}([X_t = a] | C_{\mathbf{a}}) - \mathbb{P}([X_s = a] | B_{\mathbf{a}})| \leq 1.$$

Moreover, by the  $\eta$ -spreadability of  $\mathbf{X}$ , for every  $a \in \mathcal{X}$  and every  $\mathbf{a} \in \mathcal{X}^{\mathcal{G}_\ell^s}$  we have

$$(4.17) \quad |\mathbb{P}(C_{\mathbf{a}}) - \mathbb{P}(B_{\mathbf{a}})| \leq \eta \quad \text{and} \quad |\mathbb{P}([X_t = a] \cap C_{\mathbf{a}}) - \mathbb{P}([X_s = a] \cap B_{\mathbf{a}})| \leq \eta.$$

Hence, for every  $a \in \mathcal{X}$  and every  $\mathbf{a} \in \mathcal{B}$  we have

$$(4.18) \quad \begin{aligned} |\mathbb{P}([X_t = a] | C_{\mathbf{a}}) - \mathbb{P}([X_s = a] | B_{\mathbf{a}})| &= \left| \frac{\mathbb{P}([X_t = a] \cap C_{\mathbf{a}})}{\mathbb{P}(C_{\mathbf{a}})} - \frac{\mathbb{P}([X_s = a] \cap B_{\mathbf{a}})}{\mathbb{P}(B_{\mathbf{a}})} \right| \\ &\leq \frac{1}{\mathbb{P}(B_{\mathbf{a}})} |\mathbb{P}([X_t = a] \cap C_{\mathbf{a}}) - \mathbb{P}([X_s = a] \cap B_{\mathbf{a}})| + \mathbb{P}(C_{\mathbf{a}}) \left| \frac{1}{\mathbb{P}(C_{\mathbf{a}})} - \frac{1}{\mathbb{P}(B_{\mathbf{a}})} \right| \\ &\leq \eta^{1/3} + \frac{1}{\mathbb{P}(B_{\mathbf{a}})} |\mathbb{P}(B_{\mathbf{a}}) - \mathbb{P}(C_{\mathbf{a}})| \leq 2\eta^{1/3}. \end{aligned}$$

By (4.15), (4.16) and (4.18), we conclude that for every  $a \in \mathcal{X}$ ,

$$(4.19) \quad \|f_a - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})]\|_{L_2}^2 \leq \sum_{\mathbf{a} \in \mathcal{B}} 4\eta^{2/3} \mathbb{P}(B_{\mathbf{a}}) + |\mathcal{A}| \eta^{2/3} \leq 4\eta^{2/3} m^{|\mathcal{G}_\ell^s|}.$$

Inequality (4.14) follows from (4.19) after observing that  $|\mathcal{G}_\ell^s| \leq (\ell(d+1))^d$ .  $\square$

We will also need the following consequence of Proposition 3.1.

**Corollary 4.3.** *Let  $d, m, \kappa_0$  be positive integers with  $m \geq 2$ , let  $\varepsilon > 0$ , and set*

$$(4.20) \quad u_0 = u_0(d, m, \kappa_0) := 5(d+1)^2(d+1)! m \kappa_0^{2^{d+2}} \varepsilon^{-2^{d+2}}.$$

*Let  $\mathcal{X}$  be a set with  $|\mathcal{X}| = m$ , and let  $\mathcal{H} = \langle h^a : a \in \mathcal{X} \rangle$  be an  $\mathcal{X}$ -partition of unity defined on  $\mathcal{Y}^{\{0\} \cup [d]}$  where  $(\mathcal{Y}, \nu)$  is a finite probability space. Then there exist a finite probability space  $(\Omega, \mu)$  with  $|\Omega| \leq u_0 |\mathcal{Y}|$  and a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $\Omega^{\{0\} \cup [d]}$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{[d]}{d}$  with  $|\mathcal{F}| \leq \kappa_0$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have*

$$(4.21) \quad \left| \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_{\{0\} \cup s}) d\nu(\mathbf{y}) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\mu(\boldsymbol{\omega}) \right| \leq \varepsilon.$$

(Here, we follow the conventions in Proposition 3.1 and Theorem 4.1.)

**4.4. The inductive hypothesis.** We have already mentioned in the introduction that the proof of Theorem 4.1 proceeds by induction on  $d$ . Specifically, for every positive integer  $d$  by  $\mathbf{P}(d)$  we shall denote the following statement.

Let the parameters  $m, k, \varepsilon, n_0(d, m, k, \varepsilon), \eta(d, m, k, \varepsilon), v(d, m, k, \varepsilon)$  and the notation be as in Theorem 4.1. Then for every integer  $n \geq n_0(d, m, k, \varepsilon)$ , every set  $\mathcal{X}$  with  $|\mathcal{X}| = m$  and every  $\mathcal{X}$ -valued,  $\eta(d, m, k, \varepsilon)$ -spreadable,  $d$ -dimensional random array  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  on  $[n]$  there exist a finite probability space  $(\Omega, \mu)$  with  $|\Omega| \leq v(d, m, k, \varepsilon)$  and a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $\Omega^{\{0\} \cup [d]}$  such that for every  $M \in \binom{[n]}{k}$ , every nonempty subset  $\mathcal{F}$  of  $\binom{M}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.22) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\omega_{\{0\} \cup s}) d\mu(\omega) \right| \leq \varepsilon.$$

Notice that Theorem 4.1 is equivalent to the validity of  $\mathbf{P}(d)$  for every integer  $d \geq 1$ .

**4.5. The base case “ $d = 1$ ”.** In this subsection we establish  $\mathbf{P}(1)$ . We note that this case is, essentially, the analogue of the results of Diaconis and Freedman [DF80] for approximately spreadable random vectors. The proofs, however, are rather different, and the bounds we obtain are quite weaker than those in [DF80]; this is mainly due to the fact that we are dealing with random vectors whose distribution is much less symmetric.

We proceed to the details. Let  $m, k$  be positive integers with  $m \geq 2$ , and let  $\varepsilon > 0$ . For notational convenience, we set  $\theta := \theta(1, k, \varepsilon)$ ,  $\ell_0 := \ell_0(1, m, k, \varepsilon)$  and  $\eta := \eta(1, k, m, \varepsilon)$  where  $\theta(1, k, \varepsilon)$ ,  $\ell_0(1, m, k, \varepsilon)$  and  $\eta(1, k, m, \varepsilon)$  are as in (4.1), (4.2) and (4.5) respectively. Let  $n, \mathcal{X}$  and  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{1} \rangle$  be as in  $\mathbf{P}(1)$ , and define

$$(4.23) \quad L := \{jk\ell_0 : j \in [k]\}.$$

Notice that  $L$  is a  $(k\ell_0)$ -sparse subset of  $[n]$  with  $|L| = k$ . Since  $n \geq n_0(1, m, k, \varepsilon)$ , by the choice of  $n_0(1, m, k, \varepsilon)$  in (4.6), Proposition 2.1 and the choice of  $\ell_0$ , there exists  $\ell \in [\ell_0]$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.24) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E} [ \mathbf{1}_{[X_s = a_s]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X}) ] \right] \right| \leq k \sqrt{\theta + 15\eta m^{2k^{m \lceil 1/\theta \rceil + 1}}}.$$

Fix  $t_0 \in \binom{L}{1}$  and set  $\mathcal{G} := \mathcal{G}_{\ell}^{t_0}$ . By (2.6), we see that  $\mathcal{G}_\ell^s = \mathcal{G}$  for every  $s \in \binom{L}{1}$ . By Lemma 4.2, the previous observation and the fact that  $\ell \leq \ell_0$ , for every  $s \in \binom{L}{1}$  and every  $a \in \mathcal{X}$  we have

$$(4.25) \quad \left\| \mathbb{E} [ \mathbf{1}_{[X_s = a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X}) ] - \mathbb{E} [ \mathbf{1}_{[X_{t_0} = a]} | \Sigma(\mathcal{G}, \mathbf{X}) ] \right\|_{L_2} \leq 2\eta^{1/3} m^{\ell_0}.$$

Moreover, notice that

$$(4.26) \quad k \sqrt{\theta + 15\eta m^{2k^{m \lceil 1/\theta \rceil + 1}}} + 2k\eta^{1/3} m^{\ell_0} \leq \frac{\varepsilon}{4}.$$

By (4.24)–(4.26), the Cauchy–Schwarz inequality, the fact that  $\binom{L}{1} = k$  and a telescopic argument, we conclude that for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$ ,

$$(4.27) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E} [\mathbf{1}_{[X_{t_0} = a_s]} \mid \Sigma(\mathcal{G}, \mathbf{X})] \right] \right| \leq \frac{\varepsilon}{4}.$$

Next, set  $\mathcal{Y} := \mathcal{X}^{\mathcal{G}}$  and define a probability measure  $\nu$  on  $\mathcal{Y}$  by the rule

$$(4.28) \quad \nu(\mathbf{a}) := \mathbb{P} \left( \bigcap_{s \in \mathcal{G}} [X_s = a_s] \right)$$

for every  $\mathbf{a} = (a_s)_{s \in \mathcal{G}} \in \mathcal{Y}$ . Moreover, for every  $a \in \mathcal{X}$  define  $h'_a : \mathcal{Y} \rightarrow [0, 1]$  by setting for every  $\mathbf{a} = (a_s)_{s \in \mathcal{G}} \in \mathcal{Y}$ ,

$$(4.29) \quad h'_a(\mathbf{a}) := \mathbb{P} \left( [X_{t_0} = a] \mid \bigcap_{s \in \mathcal{G}} [X_s = a_s] \right).$$

Observe that  $\langle h'_a : a \in \mathcal{X} \rangle$  is an  $\mathcal{X}$ -partition of unity, and for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.30) \quad \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E} [\mathbf{1}_{[X_{t_0} = a_s]} \mid \Sigma(\mathcal{G}, \mathbf{X})] \right] = \int \prod_{s \in \mathcal{F}} h'_{a_s}(\mathbf{a}) d\nu(\mathbf{a}).$$

This information is already strong enough, but we need to write it in a form which is suitable for the induction.

Specifically, for every  $a \in \mathcal{X}$  we define the function  $h^a : \mathcal{Y}^{\{0\} \cup [1]} \rightarrow [0, 1]$  by setting  $h^a(\mathbf{a}_0, \mathbf{a}_1) := h'_a(\mathbf{a}_0)$  for every  $(\mathbf{a}_0, \mathbf{a}_1) \in \mathcal{Y}^{\{0\} \cup [1]}$ . Again observe that  $\langle h^a : a \in \mathcal{X} \rangle$  is an  $\mathcal{X}$ -partition of unity, and for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.31) \quad \int \prod_{s \in \mathcal{F}} h'_{a_s}(\mathbf{a}) d\nu(\mathbf{a}) = \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_{\{0\} \cup s}) d\nu(\mathbf{y})$$

where  $\nu$  denotes the product measure on  $\mathcal{Y}^{\{0\} \cup \mathbb{N}}$  obtained by equipping each factor with the measure  $\nu$ . By the choice of  $v(1, k, m, \varepsilon)$  in (4.7), the fact that  $|\mathcal{Y}| \leq m^{k^{m^{1/\theta}}}$  and Corollary 4.3 applied for “ $\kappa_0 = k$ ”, “ $d = 1$ ” and “ $\varepsilon = \varepsilon/4$ ”, there exist a finite probability space  $(\Omega, \mu)$  with  $|\Omega| \leq v(1, k, m, \varepsilon)$  and a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $\Omega^{\{0\} \cup [1]}$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.32) \quad \left| \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_{\{0\} \cup s}) d\nu(\mathbf{y}) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \frac{\varepsilon}{4}$$

and so, by (4.27), (4.30), (4.31) and (4.32),

$$(4.33) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \frac{\varepsilon}{2}.$$

Finally, by (4.33) and the  $\eta$ -spreadability of  $\mathbf{X}$ , we see that if  $M \in \binom{[n]}{k}$  is arbitrary, then for every nonempty subset  $\mathcal{F}$  of  $\binom{[M]}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$ ,

$$(4.34) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\omega_{\{0\} \cup s}) d\mu(\omega) \right| \leq \frac{\varepsilon}{2} + \eta \stackrel{(4.5)}{\leq} \varepsilon.$$

The proof of the case “ $d = 1$ ” is completed.

**4.6. The general inductive step.** Let  $d \geq 2$  be an integer, and assume that  $\mathbf{P}(d-1)$  holds true. We will show that  $\mathbf{P}(d)$  also holds true. Clearly, this is enough to complete the proof of Theorem 4.1.

We fix a pair  $m, k$  of positive integers with  $k \geq d$  and  $m \geq 2$ , and  $\varepsilon > 0$ . As in the previous subsection, for notation convenience, we set

$$(4.35) \quad \ell_0 := \ell_0(d, m, k, \ell), \quad \bar{m} := \bar{m}(d, m, k, \ell) \quad \text{and} \quad \bar{\varepsilon} := \bar{\varepsilon}(d, m, k, \ell).$$

where  $\ell_0(d, m, k, \ell)$ ,  $\bar{m}(d, m, k, \ell)$  and  $\bar{\varepsilon}(d, m, k, \ell)$  are as in (4.2), (4.3) and (4.4) respectively. Also let the parameters  $n_0(d, m, k, \varepsilon)$  and

$$(4.36) \quad \eta := \eta(d, m, k, \varepsilon)$$

be as in Subsection 4.1, and let  $n, \mathcal{X}$  and  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be as in  $\mathbf{P}(d)$ . We set

$$(4.37) \quad Q := \{j k \ell_0 : j \in [n_0(d-1, \bar{m}, k, \bar{\varepsilon})]\}$$

and

$$(4.38) \quad L := \{j k \ell_0 : j \in [k]\}.$$

Notice that both  $L$  and  $Q$  are  $(k\ell_0)$ -sparse subsets of  $[n]$ . *All these data will fixed in the rest of this subsection.*

**4.6.1. Step 1: application of the approximation.** First observe that, by the selection in Subsection 4.1, we have

$$(4.39) \quad k^d \sqrt{\theta(d, k, \varepsilon) + 15\eta(d, m, k, \varepsilon) m^{\kappa(d+1)\ell_0}} \leq \frac{\varepsilon}{8}.$$

Since  $L$  is a  $(k\ell_0)$ -sparse subset of  $[n]$  with  $|L| = k$  and  $n \geq n_0(d, m, k, \varepsilon)$ , by Proposition 2.1 and (4.39), there exists  $\ell \in [\ell_0]$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{[L]}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.40) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E} [\mathbf{1}_{[X_s = a_s]} \mid \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right] \right| \leq \frac{\varepsilon}{8}.$$

4.6.2. *Step 2: application of the shift invariance property.* Fix  $t_0 \in \binom{Q}{d}$ , and set  $\mathcal{G} := \mathcal{G}_\ell^{t_0}$  and  $M_{t_0} := t_0 \cup (\cup \mathcal{G}_\ell^{t_0})$ . Moreover, for every  $s \in \binom{Q}{d}$ , every  $\mathbf{a} = (a_u)_{u \in \mathcal{G}} \in \mathcal{X}^{\mathcal{G}}$  and every  $a \in \mathcal{X}$  set

$$(4.41) \quad M_s := s \cup (\cup \mathcal{G}_\ell^s), \quad B_{\mathbf{a}}^s := \bigcap_{u \in \mathcal{G}} [X_{I_{M_{t_0}, M_s}(u)} = a_u] \quad \text{and} \quad \lambda_{\mathbf{a}}^a := \mathbb{P}([X_{t_0} = a] | B_{\mathbf{a}}^{t_0})$$

where  $I_{M_{t_0}, M_s}$  is as in (2.2). Finally, for every  $s \in \binom{Q}{d}$  and every  $a \in \mathcal{X}$  set

$$(4.42) \quad f_s^a := \sum_{\mathbf{a} \in \mathcal{X}^{\mathcal{G}}} \lambda_{\mathbf{a}}^a \mathbf{1}_{B_{\mathbf{a}}^s}$$

and notice that, by Lemma 4.2 and the fact that  $\ell \leq \ell_0$ ,

$$(4.43) \quad \|f_s^a - \mathbb{E}[\mathbf{1}_{[X_s=a]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})]\|_{L_2} \leq 2\sqrt{\eta^{2/3} m^{(\ell_0(d+1))^d}}.$$

On the other hand, it is easy to see that

$$(4.44) \quad 2k^d \sqrt{\eta^{2/3} m^{(\ell_0(d+1))^d}} \leq \frac{\varepsilon}{8}.$$

By (4.43) and (4.44), the Cauchy–Schwartz inequality, the observation that  $\|f_s^a\|_{L_\infty} \leq 1$ , the fact that  $|\binom{L}{d}| \leq k^d$  and a telescopic argument, we obtain that for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$ ,

$$(4.45) \quad \left| \mathbb{E} \left[ \prod_{s \in \mathcal{F}} \mathbb{E}[\mathbf{1}_{[X_s=a_s]} | \Sigma(\mathcal{G}_\ell^s, \mathbf{X})] \right] - \mathbb{E} \left[ \prod_{s \in \mathcal{F}} f_s^{a_s} \right] \right| \leq \frac{\varepsilon}{8}.$$

4.6.3. *Step 3: application of the inductive hypothesis.* Fix  $y_0 \in \binom{Q}{d-1}$ , and set  $\mathcal{R} := \mathcal{R}_\ell^{y_0}$  and  $L_{y_0} := \cup \mathcal{R}_\ell^{y_0}$  where  $\mathcal{R}_\ell^{y_0}$  is as in (2.3). Furthermore, for every  $x \in \binom{Q}{d-1}$  and every  $\mathbf{b} = (b_u)_{u \in \mathcal{R}} \in \mathcal{X}^{\mathcal{R}}$  set

$$(4.46) \quad L_x := \cup \mathcal{R}_\ell^x \quad \text{and} \quad C_{\mathbf{b}}^x := \bigcap_{u \in \mathcal{R}} [X_{I_{L_{y_0}, L_x}(u)} = b_u].$$

(Here,  $I_{L_{y_0}, L_x}$  is as in (2.2).) Next, set

$$(4.47) \quad \mathcal{Z} := \mathcal{X}^{\mathcal{R}}$$

and let  $\mathbf{Y} := \langle Y_x : x \in \binom{Q}{d-1} \rangle$  denote the  $\mathcal{Z}$ -valued,  $(d-1)$ -dimensional random array on  $Q$  defined by setting  $[Y_x = \mathbf{b}] = C_{\mathbf{b}}^x$  for every  $x \in \binom{Q}{d-1}$  and every  $\mathbf{b} \in \mathcal{Z}$ . Since the random array  $\mathbf{X}$  is  $\eta$ -spreadable and  $\eta = \eta(d, m, k, \varepsilon) \leq \eta(d-1, \bar{m}, k, \bar{\varepsilon})$ , we see that  $\mathbf{Y}$  is  $\eta(d-1, \bar{m}, k, \bar{\varepsilon})$ -spreadable. Moreover, by (4.37), we have that  $|Q| = n_0(d-1, \bar{m}, k, \bar{\varepsilon})$ . Therefore, by the fact that  $|\mathcal{Z}| \leq \bar{m}$  and our inductive hypothesis that property  $\mathbf{P}(d-1)$  holds true, there exist a finite probability space  $(\mathcal{Y}, \nu)$  with  $|\mathcal{Y}| \leq v(d-1, \bar{m}, k, \bar{\varepsilon})$  and a partition  $\langle E'_{\mathbf{b}} : \mathbf{b} \in \mathcal{Z} \rangle$  of  $\mathcal{Y}^{\{0\} \cup [d-1]}$  such that for every nonempty subset  $\Gamma$  of  $\binom{L}{d-1}$  and every collection  $(\mathbf{b}_x)_{x \in \Gamma}$  of elements of  $\mathcal{Z}$  we have

$$(4.48) \quad \left| \mathbb{P} \left( \bigcap_{x \in \Gamma} [Y_x = \mathbf{b}_x] \right) - \int \prod_{x \in \Gamma} \mathbf{1}_{E'_{\mathbf{b}_x}}(\mathbf{y}_{\{0\} \cup x}) d\nu(\mathbf{y}) \right| \leq \bar{\varepsilon}.$$



4.6.4. *Step 4: compatibility.* It is convenient to introduce the following notation. For every  $s \in \binom{[N]}{d}$  set

$$(4.49) \quad \partial s := \binom{s}{d-1}.$$

Next, given  $\mathbf{a} = (a_u)_{u \in \mathcal{G}} \in \mathcal{X}^{\mathcal{G}}$  and  $\beta = (\mathbf{b}^z)_{z \in \partial t_0} = \langle b_u^z : u \in \mathcal{R}, z \in \partial t_0 \rangle \in \mathcal{Z}^{\partial t_0}$ , we say that the pair  $(\mathbf{a}, \beta)$  is *compatible* provided that for every  $u \in \mathcal{G}$ , every  $u' \in \mathcal{R}$  and every  $z \in \partial t_0$ , if we have  $I_{L_{y_0}, L_z}(u') = u$ , then  $a_u = b_{u'}^z$ . Set

$$(4.50) \quad \mathbf{B} := \{\beta \in \mathcal{Z}^{\partial t_0} : \text{there exists } \mathbf{a} \in \mathcal{X}^{\mathcal{G}} \text{ such that the pair } (\mathbf{a}, \beta) \text{ is compatible}\}.$$

Notice that for every  $\beta \in \mathbf{B}$  there exists a unique  $\mathbf{a} \in \mathcal{X}^{\mathcal{G}}$  such that the pair  $(\mathbf{a}, \beta)$  is compatible, and conversely, for every  $\mathbf{a} \in \mathcal{X}^{\mathcal{G}}$  there exists a unique  $\beta \in \mathbf{B}$  such that the pair  $(\mathbf{a}, \beta)$  is compatible. This observation enables us to define the map  $T: \mathbf{B} \rightarrow \mathcal{X}^{\mathcal{G}}$  by setting  $T(\beta)$  to be the unique element of  $\mathcal{X}^{\mathcal{G}}$  such that the pair  $(T(\beta), \beta)$  is compatible. Observe that for every  $\beta = (\mathbf{b}^z)_{z \in \partial t_0} \in \mathbf{B}$  and every  $s \in \binom{[Q]}{d}$  we have the identity

$$(4.51) \quad B_{T(\beta)}^s = \bigcap_{x \in \partial s} C_{\mathbf{b}^{\mathbf{I}_{s, t_0}(x)}}^x$$

where  $B_{T(\beta)}^s$  is as in (4.41), and for every  $x \in \partial s$  the event  $C_{\mathbf{b}^{\mathbf{I}_{s, t_0}(x)}}^x$  is as in (4.46); on the other hand, notice that for every  $(\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0} \setminus \mathbf{B}$  and every  $s \in \binom{[Q]}{d}$  we have

$$(4.52) \quad \bigcap_{x \in \partial s} C_{\mathbf{b}^{\mathbf{I}_{s, t_0}(x)}}^x = \emptyset.$$

Having these observations in mind, for every  $a \in \mathcal{X}$  and every  $\beta \in \mathcal{Z}^{\partial t_0}$  we define

$$(4.53) \quad \lambda_{\beta}^a := \begin{cases} \lambda_{T(\beta)}^a & \text{if } \beta \in \mathbf{B}, \\ 0 & \text{otherwise} \end{cases}$$

where  $\lambda_{T(\beta)}^a$  is as in (4.41). By (4.51), (4.52), (4.53) and (4.42), it follows in particular that for every  $s \in \binom{[Q]}{d}$  and every  $a \in \mathcal{X}$  we have

$$(4.54) \quad f_s^a = \sum_{\beta = (\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}} \lambda_{\beta}^a \prod_{x \in \partial s} \mathbf{1}_{C_{\mathbf{b}^{\mathbf{I}_{s, t_0}(x)}}^x}.$$

4.6.5. *Step 5: definition of the partition of unity.* Now, for every  $a \in \mathcal{X}$  we define a function  $h^a: \mathcal{Y}^{\{0\} \cup [d]} \rightarrow [0, 1]$  by setting for every  $\mathbf{y} \in \mathcal{Y}^{\{0\} \cup [d]}$ ,

$$(4.55) \quad h^a(\mathbf{y}) := \sum_{\beta = (\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}} \lambda_{\beta}^a \prod_{x \in \partial [d]} \mathbf{1}_{E'_{\mathbf{b}^{\mathbf{I}_{t_0}(x)}}}(\mathbf{y}_{\{0\} \cup x}).$$

For every  $\beta = (\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}$  the map

$$(4.56) \quad \mathcal{Y}^{\{0\} \cup [d]} \ni \mathbf{y} \mapsto \prod_{x \in \partial [d]} \mathbf{1}_{E'_{\mathbf{b}^{\mathbf{I}_{t_0}(x)}}}(\mathbf{y}_{\{0\} \cup x})$$

is clearly boolean, and so, it is equal to the indicator function of some subset of  $\mathcal{Y}^{\{0\} \cup [d]}$  which we shall denote by  $D_{\beta}$ . Using the fact that the family  $\langle E'_{\mathbf{b}} : \mathbf{b} \in \mathcal{Z} \rangle$  is a partition

of  $\mathcal{Y}^{\{0\} \cup [d-1]}$ , we see that the family  $\langle D_\beta : \beta \in \mathcal{Z}^{\partial t_0} \rangle$  is also a partition of  $\mathcal{Y}^{\{0\} \cup [d]}$  with possibly empty parts; therefore, by (4.53) and (4.41), we conclude that the collection  $\mathcal{H} = \langle h^a : a \in \mathcal{X} \rangle$  is an  $\mathcal{X}$ -partition of unity.

4.6.6. *Step 6: application of the coding.* Recall that  $|\mathcal{Y}| \leq v(d-1, \bar{m}, k, \bar{\varepsilon})$ . Therefore, by (4.10) and Corollary 4.3 applied for “ $\kappa_0 = \binom{k}{d}$ ” and “ $\varepsilon = \varepsilon/8$ ”, there exist a finite probability space  $(\Omega, \mu)$  with  $|\Omega| \leq v(d, m, k, \varepsilon)$  and a partition  $\langle E^a : a \in \mathcal{X} \rangle$  of  $\Omega^{\{0\} \cup [d]}$  such that for every nonempty subset  $\mathcal{F}$  of  $\binom{L}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.57) \quad \left| \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_{\{0\} \cup s}) d\nu(\mathbf{y}) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\mu(\boldsymbol{\omega}) \right| \leq \frac{\varepsilon}{8}.$$

4.6.7. *Step 7: verification of the inductive hypothesis.* Let  $\mathcal{F}$  be an arbitrary nonempty subset of  $\binom{L}{d}$  and let  $(a_s)_{s \in \mathcal{F}}$  be an arbitrary collection of elements of  $\mathcal{X}$ . We set

$$(4.58) \quad \Gamma := \left\{ x \in \binom{L}{d-1} : x \in \partial s \text{ for some } s \in \mathcal{F} \right\}.$$

By (4.54), we have

$$(4.59) \quad \begin{aligned} \prod_{s \in \mathcal{F}} f_s^{a_s} &= \prod_{s \in \mathcal{F}} \sum_{\beta = (\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}} \lambda_\beta^{a_s} \prod_{x \in \partial s} \mathbf{1}_{C_{\mathbf{b}^{\mathbf{I}_s, t_0}(x)}}^x \\ &= \sum_{\langle \mathbf{b}^{z,s} : z \in \partial t_0, s \in \mathcal{F} \rangle \in (\mathcal{Z}^{\partial t_0})^\mathcal{F}} \prod_{s \in \mathcal{F}} \left( \lambda_{(\mathbf{b}^{z,s})_{z \in \partial t_0}}^{a_s} \prod_{x \in \partial s} \mathbf{1}_{C_{\mathbf{b}^{\mathbf{I}_s, t_0}(x), s}}^x \right) \\ &= \sum_{\langle \mathbf{b}^{z,s} : z \in \partial t_0, s \in \mathcal{F} \rangle \in (\mathcal{Z}^{\partial t_0})^\mathcal{F}} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{z,s})_{z \in \partial t_0}}^{a_s} \right) \prod_{s \in \mathcal{F}} \prod_{x \in \partial s} \mathbf{1}_{C_{\mathbf{b}^{\mathbf{I}_s, t_0}(x), s}}^x. \\ &= \sum_{\langle \mathbf{b}^{z,s} : z \in \partial t_0, s \in \mathcal{F} \rangle \in (\mathcal{Z}^{\partial t_0})^\mathcal{F}} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{z,s})_{z \in \partial t_0}}^{a_s} \right) \prod_{x \in \Gamma} \prod_{\{s \in \mathcal{F} : x \in \partial s\}} \mathbf{1}_{C_{\mathbf{b}^{\mathbf{I}_s, t_0}(x), s}}^x. \end{aligned}$$

Fix  $x \in \Gamma$  and let  $s_1, s_2 \in \mathcal{F}$  such that  $x \in \partial s_1$  and  $x \in \partial s_2$ . Note that if we have  $\mathbf{b}^{\mathbf{I}_{s_1, t_0}(x), s_1} \neq \mathbf{b}^{\mathbf{I}_{s_2, t_0}(x), s_2}$  then the events  $C_{\mathbf{b}^{\mathbf{I}_{s_1, t_0}(x), s_1}}^x$  and  $C_{\mathbf{b}^{\mathbf{I}_{s_2, t_0}(x), s_2}}^x$  are disjoint; in particular, all these terms in the above sum vanish. On the other hand, in the remaining terms, the last product collapses since it consists of indicators of the same event. Thus, we have

$$(4.60) \quad \begin{aligned} \prod_{s \in \mathcal{F}} f_s^{a_s} &\stackrel{(4.59)}{=} \sum_{(\mathbf{b}^x)_{x \in \Gamma} \in \mathcal{Z}^\Gamma} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{\mathbf{I}_{t_0, s}(z)})_{z \in \partial t_0}}^{a_s} \right) \prod_{x \in \Gamma} \mathbf{1}_{C_{\mathbf{b}^x}}^x \\ &= \sum_{(\mathbf{b}^x)_{x \in \Gamma} \in \mathcal{Z}^\Gamma} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{\mathbf{I}_{t_0, s}(z)})_{z \in \partial t_0}}^{a_s} \right) \prod_{x \in \Gamma} \mathbf{1}_{[Y_x = \mathbf{b}^x]}. \end{aligned}$$

Since  $|\mathcal{Z}^\Gamma| \leq (\bar{m})^{k^{d-1}}$  and  $\bar{\varepsilon} = (\varepsilon/8)(\bar{m})^{-k^{d-1}}$ , by (4.60) and (4.48), we have

$$(4.61) \quad \left| \mathbb{E} \left[ \prod_{s \in \mathcal{F}} f_s^{a_s} \right] - \sum_{(\mathbf{b}^x)_{x \in \Gamma} \in \mathcal{Z}^\Gamma} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{\mathbf{I}_{t_0, s}(z)})_{z \in \partial t_0}}^{a_s} \right) \int \prod_{x \in \Gamma} \mathbf{1}_{E_{\mathbf{b}^x}^x}(\mathbf{y}_{\{0\} \cup x}) d\nu(\mathbf{y}) \right| \leq \frac{\varepsilon}{8}.$$

Moreover, arguing precisely as above and using the fact that the family  $\langle E'_\mathbf{b} : \mathbf{b} \in \mathcal{Z} \rangle$  is a partition, we obtain that

$$\begin{aligned}
 (4.62) \quad & \sum_{(\mathbf{b}^x)_{x \in \Gamma} \in \mathcal{Z}^\Gamma} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{I_{t_0, s}(z)}, s)}^{a_s} \right)_{z \in \partial t_0} \int \prod_{x \in \Gamma} \mathbf{1}_{E'_{\mathbf{b}^x}}(\mathbf{y}_{\{0\} \cup x}) d\nu(\mathbf{y}) = \\
 &= \int \sum_{\langle \mathbf{b}^{z, s} : z \in \partial t_0, s \in \mathcal{F} \rangle \in (\mathcal{Z}^{\partial t_0})^\mathcal{F}} \left( \prod_{s \in \mathcal{F}} \lambda_{(\mathbf{b}^{z, s})_{z \in \partial t_0}}^{a_s} \right) \prod_{s \in \mathcal{F}} \prod_{x \in \partial s} \mathbf{1}_{E'_{\mathbf{b}^{I_{s, t_0}(x), s}}}(\mathbf{y}_{\{0\} \cup x}) d\nu(\mathbf{y}) \\
 &= \int \sum_{\langle \mathbf{b}^{z, s} : z \in \partial t_0, s \in \mathcal{F} \rangle \in (\mathcal{Z}^{\partial t_0})^\mathcal{F}} \prod_{s \in \mathcal{F}} \left( \lambda_{(\mathbf{b}^{z, s})_{z \in \partial t_0}}^{a_s} \prod_{x \in \partial s} \mathbf{1}_{E'_{\mathbf{b}^{I_{s, t_0}(x), s}}}(\mathbf{y}_{\{0\} \cup x}) \right) d\nu(\mathbf{y}) \\
 &= \int \prod_{s \in \mathcal{F}} \sum_{(\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}} \lambda_{(\mathbf{b}^z)_{z \in \partial t_0}}^{a_s} \prod_{x \in \partial s} \mathbf{1}_{E'_{\mathbf{b}^{I_{s, t_0}(x)}}}(\mathbf{y}_{\{0\} \cup x}) d\nu(\mathbf{y}) \\
 &= \int \prod_{s \in \mathcal{F}} \sum_{(\mathbf{b}^z)_{z \in \partial t_0} \in \mathcal{Z}^{\partial t_0}} \lambda_{(\mathbf{b}^z)_{z \in \partial t_0}}^{a_s} \prod_{x \in \partial[d]} \mathbf{1}_{E'_{\mathbf{b}^{I_{t_0}(x)}}}(\mathbf{y}_{\{0\} \cup I_s(x)}) d\nu(\mathbf{y}) \\
 &\stackrel{(4.55)}{=} \int \prod_{s \in \mathcal{F}} h^{a_s}(\mathbf{y}_{\{0\} \cup s}) d\nu(\mathbf{y}).
 \end{aligned}$$

By (4.40), (4.45), (4.61), (4.62) and (4.57), for every nonempty subset  $\mathcal{F}$  of  $\binom{[L]}{d}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.63) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \frac{\varepsilon}{2}.$$

Finally, using the  $\eta$ -spreadability of  $\mathbf{X}$  and (4.63), we conclude that for every  $M \in \binom{[n]}{k}$ , every nonempty subset  $\mathcal{F}$  of  $\binom{[M]}{1}$  and every collection  $(a_s)_{s \in \mathcal{F}}$  of elements of  $\mathcal{X}$  we have

$$(4.64) \quad \left| \mathbb{P} \left( \bigcap_{s \in \mathcal{F}} [X_s = a_s] \right) - \int \prod_{s \in \mathcal{F}} \mathbf{1}_{E^{a_s}}(\boldsymbol{\omega}_{\{0\} \cup s}) d\boldsymbol{\mu}(\boldsymbol{\omega}) \right| \leq \frac{\varepsilon}{2} + \eta \stackrel{(4.8)}{\leq} \varepsilon.$$

This shows that property **P**( $d$ ) is satisfied, and so the entire proof of Theorem 4.1 is completed.

**4.7. Proofs of Theorems 1.4 and 1.5.** Invoking the definition of all relevant parameters in Subsection 4.1 and proceeding by induction on  $d$ , it is not hard to see that

$$(4.65) \quad \eta(d, m, k, \varepsilon)^{-1} \leq \exp^{(2d)} \left( \frac{2^8 m^2 k^{3d}}{\varepsilon^2} \right)$$

$$(4.66) \quad n_0(d, m, k, \varepsilon) \leq \exp^{(2d)} \left( \frac{2^8 m^2 k^{3d}}{\varepsilon^2} \right)$$

$$(4.67) \quad v(d, m, k, \varepsilon) \leq \exp^{(2d)} \left( \frac{2^8 m^2 k^{3d}}{\varepsilon^2} \right)$$

for every triple  $d, m, k$  of positive integers with  $m \geq 2$  and  $k \geq d$ , and every  $0 < \varepsilon \leq 1$ .

Thus, both Theorem 1.4 and Theorem 1.5 follow from Theorem 4.1 after taking into account the estimates in (4.65)–(4.67) and the choice of the constant  $C(d, m, k, \varepsilon)$  in (1.4).

*Remark 4.4.* The tower type dependence of the parameters  $\eta(d, m, k, \varepsilon)$ ,  $n_0(d, m, k, \varepsilon)$  and  $v(d, m, k, \varepsilon)$  with respect to the dimension  $d$  is, of course, a byproduct of the inductive nature of the proof of Theorem 4.1. It would be very interesting—and also important for certain applications—if these bounds could be improved to a single, or even double, exponential behavior.

## 5. ORBITS

The following definition plays an important role in the proof of Theorem 1.6.

**Definition 5.1** (Orbits). *Let  $\mathbf{X} = \langle X_i : i \in I \rangle$  be a family of real-valued random variables defined on a common probability space, indexed by a set  $I$  with  $|I| \geq 2$ , and such that  $\|X_i\|_{L_2} = 1$  for every  $i \in I$ . Also let  $\eta \geq 0$ . We say that  $\mathbf{X}$  is an  $\eta$ -orbit (and simply an orbit if  $\eta = 0$ ) if for every pair  $\{i_1, j_1\}$  and  $\{i_2, j_2\}$  of doubletons of  $I$  we have*

$$(5.1) \quad |\mathbb{E}[X_{i_1}X_{j_1}] - \mathbb{E}[X_{i_2}X_{j_2}]| \leq \eta.$$

Arguably, the simplest example of an orbit is a (finite or infinite) sequence of independent random variables with zero mean and unit variance. Note, however, that the notion of an orbit is significantly less restrictive than independence—we will see several more refined examples of orbits in Sections 6 and 7.

It is intuitively clear that an orbit is a stochastic process which “everywhere looks the same”. We formalize this basic intuition in the following proposition.

**Proposition 5.2** (Universality). *Let  $\eta \geq 0$ , and let  $\mathbf{X} = \langle X_i : i \in I \rangle$  be an  $\eta$ -orbit. Also let  $\mathcal{F}, \mathcal{G}$  be finite subsets of  $I$  with  $|\mathcal{F}|, |\mathcal{G}| \geq 2$ , and set*

$$(5.2) \quad Z_{\mathcal{F}} := \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} X_i \quad \text{and} \quad Z_{\mathcal{G}} := \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i.$$

Then we have

$$(5.3) \quad \|Z_{\mathcal{F}} - Z_{\mathcal{G}}\|_{L_2} \leq 2 \left( \frac{1}{\min\{|\mathcal{F}|, |\mathcal{G}|\}} + \eta \right)^{1/2}.$$

*Proof.* Fix  $\kappa, \ell \in I$  with  $\kappa \neq \ell$ , and set  $\delta := \mathbb{E}[X_{\kappa}X_{\ell}]$ . Since  $\mathbf{X}$  is an  $\eta$ -orbit, we see that  $\delta - \eta \leq \mathbb{E}[X_iX_j] \leq \delta + \eta$  for every  $i, j \in I$  with  $i \neq j$ ; also recall that  $\mathbb{E}[X_i^2] = 1$  for every  $i \in I$ . Therefore,

$$(5.4) \quad \begin{aligned} \mathbb{E}[Z_{\mathcal{F}}^2] &= \frac{1}{|\mathcal{F}|^2} \sum_{i, j \in \mathcal{F}} \mathbb{E}[X_iX_j] = \frac{1}{|\mathcal{F}|^2} \sum_{i \in \mathcal{F}} \mathbb{E}[X_i^2] + \frac{1}{|\mathcal{F}|^2} \sum_{\substack{i, j \in \mathcal{F} \\ i \neq j}} \mathbb{E}[X_iX_j] \\ &\leq \frac{1}{|\mathcal{F}|} + \left(1 - \frac{1}{|\mathcal{F}|}\right)(\delta + \eta). \end{aligned}$$

Similarly, we obtain that

$$(5.5) \quad \mathbb{E}[Z_{\mathcal{G}}^2] \leq \frac{1}{|\mathcal{G}|} + \left(1 - \frac{1}{|\mathcal{G}|}\right)(\delta + \eta).$$

On the other hand, we have

$$\begin{aligned}
 (5.6) \quad \mathbb{E}[Z_{\mathcal{F}}Z_{\mathcal{G}}] &= \frac{1}{|\mathcal{F}| \cdot |\mathcal{G}|} \sum_{i \in \mathcal{F}, j \in \mathcal{G}} \mathbb{E}[X_i X_j] \\
 &= \frac{1}{|\mathcal{F}| \cdot |\mathcal{G}|} \sum_{i \in \mathcal{F} \cap \mathcal{G}} \mathbb{E}[X_i^2] + \frac{1}{|\mathcal{F}| \cdot |\mathcal{G}|} \sum_{\substack{i \in \mathcal{F}, j \in \mathcal{G} \\ i \neq j}} \mathbb{E}[X_i X_j] \\
 &\geq \frac{|\mathcal{F} \cap \mathcal{G}|}{|\mathcal{F}| \cdot |\mathcal{G}|} + \left(1 - \frac{|\mathcal{F} \cap \mathcal{G}|}{|\mathcal{F}| \cdot |\mathcal{G}|}\right) (\delta - \eta).
 \end{aligned}$$

Finally, by the Cauchy–Schwarz inequality, we have  $|\delta| \leq 1$ . Using this observation, the estimate (5.3) follows from the identity  $\|Z_{\mathcal{F}} - Z_{\mathcal{G}}\|_{L_2}^2 = \mathbb{E}[Z_{\mathcal{F}}^2] + \mathbb{E}[Z_{\mathcal{G}}^2] - 2\mathbb{E}[Z_{\mathcal{F}}Z_{\mathcal{G}}]$  and inequalities (5.4)–(5.6).  $\square$

Proposition 5.2 will mostly be used in the following form. (The proof follows immediately from Proposition 5.2 and the Cauchy–Schwarz inequality.)

**Corollary 5.3.** *Let  $\eta \geq 0$ , let  $\mathbf{X} = \langle X_i : i \in I \rangle$  be an  $\eta$ -orbit, and let  $\mathcal{F}, \mathcal{G}$  be finite subsets of  $I$  with  $|\mathcal{F}|, |\mathcal{G}| \geq 2$ . Then for every random variable  $Y$  with  $\|Y\|_{L_2} \leq 1$  we have*

$$(5.7) \quad \left| \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \mathbb{E}[X_i Y] - \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mathbb{E}[X_i Y] \right| \leq 2 \left( \frac{1}{\min\{|\mathcal{F}|, |\mathcal{G}|\}} + \eta \right)^{1/2}.$$

In particular, if  $\vartheta \geq 0$  is such that

- (a)  $|\mathbb{E}[X_i Y] - \mathbb{E}[X_j Y]| \leq \vartheta$  for every  $i, j \in \mathcal{F}$ , and
- (b)  $|\mathbb{E}[X_i Y] - \mathbb{E}[X_j Y]| \leq \vartheta$  for every  $i, j \in \mathcal{G}$ ,

then for every  $i \in \mathcal{F}$  and every  $j \in \mathcal{G}$  we have

$$(5.8) \quad |\mathbb{E}[X_i Y] - \mathbb{E}[X_j Y]| \leq 2 \left( \frac{1}{\min\{|\mathcal{F}|, |\mathcal{G}|\}} + \eta \right)^{1/2} + 2\vartheta.$$

## 6. COMPARING TWO-POINT CORRELATIONS OF SPREADABLE RANDOM ARRAYS

**6.1. Motivation.** Let  $n \geq 4$  be an integer, and assume that  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{2} \rangle$  is a real-valued, two-dimensional random array on  $[n]$  such that  $\|X_s\|_{L_2} = 1$  for all  $s \in \binom{[n]}{2}$ . We wish to compare the correlations

$$\alpha := \mathbb{E}[X_{\{1,2\}} X_{\{3,4\}}] \quad \text{and} \quad \beta := \mathbb{E}[X_{\{1,3\}} X_{\{2,4\}}].$$

Of course, if  $\mathbf{X}$  is exchangeable, then  $\alpha = \beta$ . On the other hand, if  $\mathbf{X}$  is spreadable, then Kallenberg’s representation theorem [Kal92] and an ultraproduct argument yield that  $\alpha = \beta + o_{n \rightarrow \infty}(1)$  but with an ineffective error term. We will see, however, that

$$(6.1) \quad |\alpha - \beta| \leq \frac{6}{\sqrt{n}}.$$

In other words, the symmetries of a finite, spreadable, high-dimensional random array with square-integrable entries, impose explicit restrictions on its two-point correlations.

In order to see that (6.1) is satisfied, let  $n \geq 10$  be an integer (if  $n \leq 9$ , then (6.1) is straightforward), let  $\ell$  be the largest positive integer such that  $2\ell + 3 < n$ , and notice that  $\ell \geq n/4$ . Also observe that, by the spreadability of  $\mathbf{X}$ , for every  $i \in \{2, \dots, \ell + 1\}$  we have  $\beta = \mathbb{E}[X_{\{1, \ell+2\}} X_{\{i, n\}}]$ , and on the other hand for every  $j \in \{\ell + 3, \dots, 2\ell + 3\}$  we have  $\alpha = \mathbb{E}[X_{\{1, \ell+2\}} X_{\{j, n\}}]$ . Therefore, setting

$$Y := (1/\ell) \sum_{i=2}^{\ell+1} X_{\{i, n\}} \quad \text{and} \quad Z := (1/\ell) \sum_{j=\ell+3}^{2\ell+3} X_{\{j, n\}},$$

we obtain that

$$(6.2) \quad \alpha = \mathbb{E}[X_{\{1, \ell+2\}} Z] \quad \text{and} \quad \beta = \mathbb{E}[X_{\{1, \ell+2\}} Y].$$

The main observation, which follows readily from the spreadability of  $\mathbf{X}$ , is that the process  $\langle X_{\{k, n\}} : k \in \{2, \dots, \ell + 1\} \cup \{\ell + 3, \dots, 2\ell + 3\} \rangle$  is an orbit in the sense of Definition 5.1. This information together with (6.2) and Corollary 5.3 yield (6.1).

6.2. Our goal in this section is to study of the phenomenon outlined above and to characterize, combinatorially, when two two-point correlations of a spreadable random array essentially coincide. To this end, we need the following analogue of the notion of an aligned pair of partial maps which was introduced in Paragraph 1.4.1.

**Definition 6.1** (Aligned pair of sets). *Let  $d$  be a positive integer, and let  $s_1, s_2 \in \binom{\mathbb{N}}{d}$  be distinct. We say that the pair  $\{s_1, s_2\}$  is aligned if there exists a proper (possibly empty) subset  $G$  of  $[d]$  such that: (i)  $I_{s_1} \upharpoonright G = I_{s_2} \upharpoonright G$ , and (ii)  $I_{s_1}([d] \setminus G) \cap I_{s_2}([d] \setminus G) = \emptyset$ . (Here,  $I_{s_1}$  and  $I_{s_2}$  denote the canonical isomorphisms associated with the sets  $s_1$  and  $s_2$ ; see Paragraph 1.4.2.) We call the set  $G$  the root of  $\{s_1, s_2\}$  and we denote it by  $r(s_1, s_2)$ .*

We have the following proposition.

**Proposition 6.2.** *Let  $n, d$  be positive integers with  $n \geq 4d + 2$ , and let  $s_1, s_2, t_1, t_2 \in \binom{[n]}{d}$  with  $s_1 \neq s_2$  and  $t_1 \neq t_2$ . Assume that the pairs  $\{s_1, s_2\}$  and  $\{t_1, t_2\}$  are aligned and have the same root. If  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  is a real-valued, spreadable,  $d$ -dimensional random array on  $[n]$  such that  $\|X_s\|_{L_2} = 1$  for all  $s \in \binom{[n]}{d}$ , then*

$$(6.3) \quad |\mathbb{E}[X_{s_1} X_{s_2}] - \mathbb{E}[X_{t_1} X_{t_2}]| \leq \frac{8d^2}{\sqrt{n}}.$$

*Remark 6.3.* By considering spreadable, high-dimensional random arrays whose distribution is of the form (1.2), it is not hard to see that the assumption in Proposition 6.2 (namely, the fact that the pairs  $\{s_1, s_2\}$  and  $\{t_1, t_2\}$  are aligned and have the same root) is essentially optimal.

The proof of Proposition 6.2 is based on the following lemma.

**Lemma 6.4.** *Let  $n, d, s_1, s_2, t_1, t_2$  be as in Proposition 6.2. Assume that the pairs  $\{s_1, s_2\}$  and  $\{t_1, t_2\}$  are aligned, and that there exists  $i_0 \in [d]$  with the following properties.*

(i) We have  $I_{s_1}(i_0) < I_{s_2}(i_0)$  and  $I_{t_2}(i_0) < I_{t_1}(i_0)$ .

(ii) We have  $I_{s_1} \upharpoonright ([d] \setminus \{i_0\}) = I_{t_1} \upharpoonright ([d] \setminus \{i_0\})$  and  $I_{s_2} \upharpoonright ([d] \setminus \{i_0\}) = I_{t_2} \upharpoonright ([d] \setminus \{i_0\})$ .

If  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  is a real-valued, spreadable,  $d$ -dimensional random array on  $[n]$  such that  $\|X_s\|_{L_2} = 1$  for all  $s \in \binom{[n]}{d}$ , then

$$(6.4) \quad |\mathbb{E}[X_{s_1} X_{s_2}] - \mathbb{E}[X_{t_1} X_{t_2}]| \leq \frac{4}{\sqrt{n}}.$$

*Proof.* Since  $\mathbf{X}$  is spreadable, we may assume that there exist subintervals  $L_1$  and  $L_2$  of  $[n]$  with  $|L_1| = |L_2| = \lfloor (n - 2d + 1)/2 \rfloor$  and satisfying the following properties.

(P1) We have  $\max(L_1) < I_{s_1}(i_0) < \min(L_2)$ .

(P2) If  $i_0 > 1$ , then  $I_{s_1}(i_0 - 1) < \min(L_1)$  and  $I_{s_2}(i_0 - 1) < \min(L_1)$ .

(P3) If  $i_0 < d$ , then  $\max(L_2) < I_{s_1}(i_0 + 1)$  and  $\max(L_2) < I_{s_2}(i_0 + 1)$ .

Set  $L := L_1 \cup L_2$  and  $\ell := \lfloor (n - 2d + 1)/2 \rfloor$ ; also set

$$g_j := (s_2 \setminus \{I_{s_2}(i_0)\}) \cup \{I_L(j)\} \in \binom{[n]}{d}$$

for every  $j \in [2\ell]$ . Using the spreadability of  $\mathbf{X}$  again, we obtain that

(P4)  $\mathbb{E}[X_{s_1} X_{g_j}] = \mathbb{E}[X_{t_1} X_{t_2}]$  for every  $j \in [\ell]$ , and

(P5)  $\mathbb{E}[X_{s_1} X_{g_j}] = \mathbb{E}[X_{s_1} X_{s_2}]$  for every  $j \in [2\ell] \setminus [\ell]$ .

By the spreadability of  $\mathbf{X}$  once again, we see that the collection  $\langle X_{g_j} : j \in [2\ell] \rangle$  is an orbit and, moreover,  $\mathbb{E}[X_{s_1} X_{g_i}] = \mathbb{E}[X_{s_1} X_{g_j}]$  if either  $i, j \in [\ell]$ , or  $i, j \in [2\ell] \setminus [\ell]$ . The result follows using the previous remarks, Corollary 5.3 and the fact that  $n \geq 4d + 2$ .  $\square$

We are ready to give the proof of Proposition 6.2.

*Proof of Proposition 6.2.* Note that, by applying successively Lemma 6.4 at most  $d^2$  times, we obtain the following.

Let  $\mathfrak{s}_1, \mathfrak{s}_2, \mathfrak{s}_3, \mathfrak{s}_4 \in \binom{[n]}{d}$  with  $\mathfrak{s}_1 \neq \mathfrak{s}_2$  and  $\mathfrak{s}_3 \neq \mathfrak{s}_4$ . Assume that the pairs  $\{\mathfrak{s}_1, \mathfrak{s}_2\}$  and  $\{\mathfrak{s}_3, \mathfrak{s}_4\}$  are aligned and have the same root  $G$ . Assume, moreover, that

(i)  $I_{\mathfrak{s}_1} \upharpoonright G = I_{\mathfrak{s}_3} \upharpoonright G$ ,

(ii)  $I_{\mathfrak{s}_2} \upharpoonright G = I_{\mathfrak{s}_4} \upharpoonright G$ , and

(iii) for every interval  $I$  of  $[d] \setminus G$  we have  $\max(I_{\mathfrak{s}_3}(I)) < \min(I_{\mathfrak{s}_4}(I))$ .

Then we have  $|\mathbb{E}[X_{\mathfrak{s}_1} X_{\mathfrak{s}_2}] - \mathbb{E}[X_{\mathfrak{s}_3} X_{\mathfrak{s}_4}]| \leq 4d^2/\sqrt{n}$ .

The estimate (6.3) follows using this observation, the triangle inequality and the spreadability of the random array  $\mathbf{X}$ .  $\square$

## 7. PROOF OF THEOREM 1.6

In this section we give the proof of Theorem 1.6. As already noted, the proof is based on the results obtained in Sections 5 and 6; in particular, the reader is advised to review this material, as well as the terminology and notation introduced in Paragraphs 1.4.1 and 1.4.2, before reading this section.

**7.1. Existence of decomposition.** The main step is the following proposition which establishes the existence of the desired decomposition.

**Proposition 7.1.** *Let  $n, d, \kappa, k$  be positive integers with  $\kappa \geq 2$  and  $n \geq 2\kappa^2 d(k+1)^{d+1}$ , and set*

$$(7.1) \quad \gamma = \gamma(n, d, \kappa) := \left( \frac{1}{\kappa} + \frac{8d^2}{\sqrt{n}} \right)^{1/2}.$$

*Then there exists a subset  $N$  of  $[n]$  with  $|N| = k$  and satisfying the following property. If  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  is a real-valued, spreadable,  $d$ -dimensional random array on  $[n]$  such that  $\|X_s\|_{L_2} = 1$  for all  $s \in \binom{[n]}{d}$ , then there exists a real-valued stochastic process  $\mathbf{\Delta} = \langle \Delta_p : p \in \text{PartIncr}([d], N) \rangle$  such that the following hold true.*

- (i) *For every  $s \in \binom{N}{d}$  we have  $X_s = \sum_{F \subseteq [d]} \Delta_{\mathbf{I}_s \upharpoonright F}$ .*
- (ii) *For every  $p \in \text{PartIncr}([d], N)$  with  $p \neq \emptyset$  we have  $|\mathbb{E}[\Delta_p]| \leq 2^d \gamma$ .*
- (iii) *If  $p_1, p_2 \in \text{PartIncr}([d], N)$  are distinct and the pair  $\{p_1, p_2\}$  is aligned, then we have  $|\mathbb{E}[\Delta_{p_1} \Delta_{p_2}]| \leq 2^{2d+2} \gamma$ .*

The bulk of this section is devoted to the proof of Proposition 7.1—it spans Paragraphs 7.1.1 up to 7.1.4. The proof of Theorem 1.6 is completed in Subsection 7.2.

**7.1.1. Definitions/Notation.** This is the heart of the proof of Proposition 7.1. Our goal is to define the set  $N$  and the process  $\mathbf{\Delta}$ . This task is combinatorially delicate, and it requires a number of preparatory steps. In what follows, let  $d, n, \kappa, k, \mathbf{X}$  be as in Proposition 7.1.

**7.1.1.1.** We start by selecting two sequences  $(L_1, \dots, L_k)$  and  $(D_1, \dots, D_k, D_{k+1})$  of subintervals of  $[n-1]$  with the following properties.

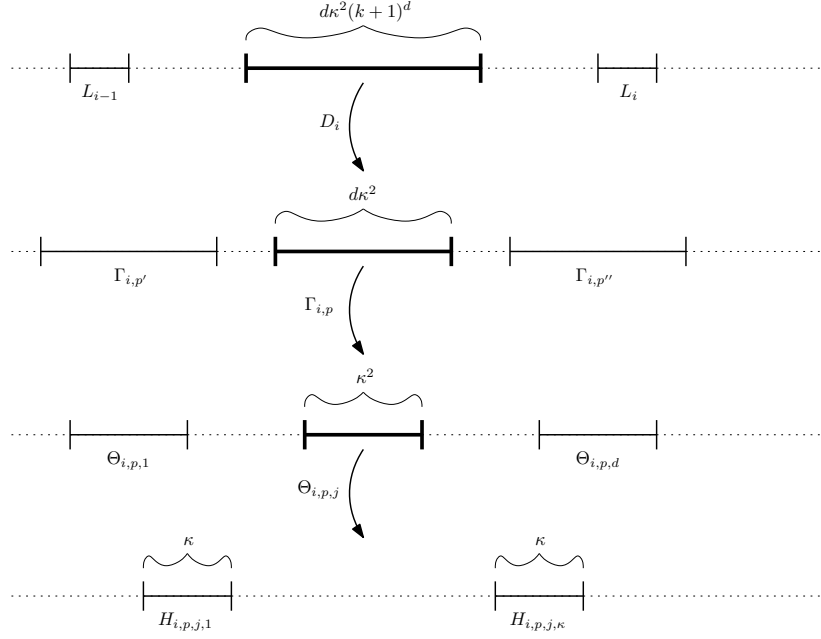
- For every  $i \in [k]$  we have  $|L_i| = \kappa$ .
- For every  $i \in [k+1]$  we have  $|D_i| = d\kappa^2(k+1)^d$ .
- For every  $i \in [k]$  we have  $\max(D_i) < \min(L_i) \leq \max(L_i) < \min(D_{i+1})$ .

We set

$$(7.2) \quad N := \{ \min(L_i) : i \in [k] \} \quad \text{and} \quad \tilde{N} := N \cup \{n\}.$$

Moreover, for every  $i \in [k+1]$  we select a collection  $(\Gamma_{i,p} : p \in \text{PartIncr}([d], N))$  of pairwise disjoint subintervals of  $D_i$  of length  $d\kappa^2$ . Next, for every  $i \in [k+1]$  and every  $p \in \text{PartIncr}([d], N)$  let  $(\Theta_{i,p,1}, \dots, \Theta_{i,p,d})$  denote the unique finite sequence of successive subintervals of  $\Gamma_{i,p}$  of length  $\kappa^2$ . Finally, for every  $i \in [k+1]$ , every  $p \in \text{PartIncr}([d], N)$  and every  $j \in [d]$  by  $(H_{i,p,j,1}, \dots, H_{i,p,j,\kappa})$  we denote the unique finite sequence of successive subintervals of  $\Theta_{i,p,j}$  of length  $\kappa$ .




 FIGURE 3. The sets  $D_i$ ,  $\Gamma_{i,p}$ ,  $\Theta_{i,p,j}$  and  $H_{i,p,j,r}$ .

7.1.1.2. Next, for every  $p \in \text{PartIncr}([d], N)$  we define a subset  $\mathcal{O}_p$  of  $\binom{[n]}{[d]}$  as follows. If  $\text{dom}(p) = [d]$ , then we set

$$(7.3) \quad \mathcal{O}_p := \{\text{Im}(p)\}.$$

Otherwise, if  $\text{dom}(p) \subsetneq [d]$ , then let  $(K_1, \dots, K_b)$  denote the unique finite sequence of subintervals of  $[d] \setminus \text{dom}(p)$  of maximal length which cover the set  $[d] \setminus \text{dom}(p)$ —that is,  $[d] \setminus \text{dom}(p) = K_1 \cup \dots \cup K_b$ —and such that  $\max(K_a) < \min(K_{a+1})$  if  $b \geq 2$  and  $a \in [b-1]$ . (Note that, in  $b \geq 2$ , then for every  $a \in [b-1]$  we have that  $\max(K_a) + 1 \in \text{dom}(p)$ .) Also let  $\tilde{p}: \text{dom}(p) \cup \{d+1\} \rightarrow \tilde{N}$  denote the extension of  $p$  which satisfies  $\tilde{p}(d+1) = n$ . For every  $r \in [\kappa]$  we set

$$s_{p,r} := \text{Im}(p) \cup \left\{ \min(H_{i,p,j,r}) : a \in [b], i = \Gamma_N^{-1}(\tilde{p}(\max(K_a) + 1)) \text{ and } j \in \{1, \dots, |K_a|\} \right\}$$

and we define

$$(7.4) \quad \mathcal{O}_p := \{s_{p,r} : r \in [\kappa]\}.$$

We also set

$$(7.5) \quad \mathcal{G}_p := \bigcup_{G \subseteq \text{dom}(p)} \mathcal{O}_p \upharpoonright G$$

and

$$(7.6) \quad \mathcal{O} := \bigcup_{p \in \text{PartIncr}([d], N)} \mathcal{O}_p.$$

Note that if  $p, p' \in \text{PartIncr}([d], N)$  are distinct, then  $\mathcal{O}_p \cap \mathcal{O}_{p'} = \emptyset$ .

Finally, for every  $s \in \mathcal{O}$  and every  $G \subseteq \text{dom}(p)$ —where  $p \in \text{PartIncr}([d], N)$  denotes the unique partial map such that  $s \in \mathcal{O}_p$ —we define a subset  $\mathcal{O}'_{s,G}$  of  $\binom{[n]}{d}$  as follows. For every  $r \in [\kappa]$  we set  $t_{s,G,r} := p(G) \cup \{v + r - 1 : v \in s \setminus p(G)\} \in \binom{[n]}{d}$  and we define

$$(7.7) \quad \mathcal{O}'_{s,G} := \{t_{s,G,r} : r \in [\kappa]\}.$$

7.1.1.3. We are now in a position to introduce the stochastic process  $\Delta$ . First, for every  $p \in \text{PartIncr}([d], N)$  we set

$$(7.8) \quad Y_p := \frac{1}{|\mathcal{O}_p|} \sum_{s \in \mathcal{O}_p} X_s$$

(notice that, by (7.3), we have  $Y_{I_s} = X_s$  for every  $s \in \binom{[n]}{d}$ ), and we define

$$(7.9) \quad \Delta_p := \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} Y_{p \upharpoonright G}.$$

We also set

$$(7.10) \quad \mathcal{A}_p := \sigma(\{X_s : s \in \mathcal{G}_p\});$$

that is,  $\mathcal{A}_p$  is the  $\sigma$ -algebra generated by the random variables  $\langle X_s : s \in \mathcal{G}_p \rangle$ .

7.1.2. *Basic properties.* We isolate, for future use, the following basic properties of the construction presented in Paragraph 7.1.1.

- (P1) For every  $p \in \text{PartIncr}([d], N)$  and every subset  $G$  of  $\text{dom}(p)$  the random variable  $Y_{p \upharpoonright G}$  is  $\mathcal{A}_p$ -measurable.
- (P2) Let  $p_1, p_2 \in \text{PartIncr}([d], N)$  be distinct and such that the pair  $\{p_1, p_2\}$  is aligned, and assume that  $r(p_1, p_2) \neq \text{dom}(p_1)$ . Then for every  $s \in \mathcal{O}_{p_1}$  the family of random variables  $\langle X_t : t \in \mathcal{O}_{p_1 \wedge p_2} \cup \mathcal{O}'_{s, r(p_1, p_2)} \rangle$  is an  $(8d^2/\sqrt{n})$ -orbit in the sense of Definition 5.1.
- (P3) Let  $p_1, p_2 \in \text{PartIncr}([d], N)$  be distinct and such that the pair  $\{p_1, p_2\}$  is aligned, and assume that  $r(p_1, p_2) \neq \text{dom}(p_1)$ . Also let  $s \in \mathcal{O}_{p_1}$  be arbitrary. Then for every  $s' \in \mathcal{O}'_{s, r(p_1, p_2)}$  we have that  $\mathbb{E}[X_{s'} | \mathcal{A}_{p_2}] = \mathbb{E}[X_s | \mathcal{A}_{p_2}]$ .

Property (P1) follows immediately by (7.8) and the fact that  $\mathcal{O}_{p \upharpoonright G} \subseteq \mathcal{G}_p$ . In order to see that property (P2) is satisfied notice that, since  $p_1 \neq p_1 \wedge p_2$ , if  $t_1, t_2 \in \mathcal{O}_{p_1 \wedge p_2} \cup \mathcal{O}'_{s, r(p_1, p_2)}$  are distinct, then  $I_{t_1} \upharpoonright r(p_1, p_2) = I_{t_2} \upharpoonright r(p_1, p_2) = p_1 \wedge p_2$  and the pair  $\{t_1, t_2\}$  is aligned with  $r(t_1, t_2) = r(p_1, p_2)$ . Using this observation, property (P2) follows from Proposition 6.2. Finally, for property (P3) we first observe that for every  $F \subseteq \text{dom}(p_2)$  we have that  $p_1 \neq (p_2 \upharpoonright F)$  and  $\text{dom}(p_1 \wedge p_2 \upharpoonright F) \subseteq r(p_1, p_2)$ . Therefore, for every  $i \in [d] \setminus r(p_1, p_2)$  and every  $F \subseteq \text{dom}(p_2)$  we have that  $I_s(i) \notin \cup \mathcal{O}_{p_2 \upharpoonright F}$  and, consequently,

$I_s(i) \notin \cup \mathcal{G}_{p_2}$ . On the other hand, by the definition of the set  $\mathcal{O}'_{s,r(p_1,p_2)}$  in (7.7), there exists a collection  $\langle J_i : i \in [d] \setminus r(p_1, p_2) \rangle$  of disjoint intervals of length  $\kappa$  such that for every  $i \in [d] \setminus r(p_1, p_2)$  we have that  $I_s(i) = \min(J_i)$ ,  $I_{s'}(i) \in J_i$  and  $J_i \cap (\cup \mathcal{G}_{p_2}) = \emptyset$ . Taking into account these remarks, property (P3) follows from the spreadability of  $\mathbf{X}$ .

7.1.3. *Compatibility of projections.* The following lemma shows that the projections associated with the  $\sigma$ -algebras defined in (7.10) behave like a lattice of projections when applied to the random variables defined in (7.8).

**Lemma 7.2.** *Let  $d, n, \kappa, k, \mathbf{X}$  be as in Proposition 7.1, and let  $\gamma$  be as in (7.1). Also let  $N \subseteq [n]$ ,  $\mathbf{Y} = \langle Y_p : p \in \text{PartIncr}([d], N) \rangle$  and  $\langle \mathcal{A}_p : p \in \text{PartIncr}([d], N) \rangle$  be as in Paragraph 7.1.1. Then for every distinct  $p_1, p_2 \in \text{PartIncr}([d], N)$  such that the pair  $\{p_1, p_2\}$  is aligned we have*

$$(7.11) \quad \left\| \mathbb{E}[Y_{p_1} | \mathcal{A}_{p_2}] - Y_{p_1 \wedge p_2} \right\|_{L_2} \leq 2\gamma.$$

*Proof.* If  $\text{dom}(p_1) = r(p_1, p_2)$ , then  $p_1 = p_1 \wedge p_2$  and, by property (P1), the random variable  $Y_{p_1}$  is  $\mathcal{A}_{p_2}$ -measurable; hence, in this case, the result is straightforward. Therefore, we may assume that  $\text{dom}(p_1) \setminus r(p_1, p_2) \neq \emptyset$ . By (7.8), it is enough to show that for every  $s \in \mathcal{O}_{p_1}$  we have

$$(7.12) \quad \left\| \mathbb{E}[X_s | \mathcal{A}_{p_2}] - Y_{p_1 \wedge p_2} \right\|_{L_2} \leq 2\gamma.$$

To this end, let  $s \in \mathcal{O}_{p_1}$  be arbitrary. Since  $p_1 \wedge p_2 = p_2 \upharpoonright r(p_1, p_2)$ , using property (P1) again, we see that  $Y_{p_1 \wedge p_2}$  is  $\mathcal{A}_{p_2}$ -measurable and, consequently,

$$(7.13) \quad Y_{p_1 \wedge p_2} = \mathbb{E}[Y_{p_1 \wedge p_2} | \mathcal{A}_{p_2}].$$

By property (P2), the process  $\langle X_t : t \in \mathcal{O}_{p_1 \wedge p_2} \cup \mathcal{O}'_{s,r(p_1,p_2)} \rangle$  is an  $(8d^2/\sqrt{n})$ -orbit in the sense of Definition 5.1. Moreover,  $|\mathcal{O}_{p_1 \wedge p_2}| = |\mathcal{O}'_{s,r(p_1,p_2)}| = \kappa$ . By Proposition 5.2, the definition of  $Y_{p_1 \wedge p_2}$  in (7.8) and the choice of  $\gamma$ , we have

$$(7.14) \quad \left\| Y_{p_1 \wedge p_2} - \frac{1}{\kappa} \sum_{t \in \mathcal{O}'_{s,r(p_1,p_2)}} X_t \right\|_{L_2} \leq 2\gamma$$

and so, by the contractive property of conditional expectation and (7.13),

$$(7.15) \quad \left\| Y_{p_1 \wedge p_2} - \frac{1}{\kappa} \sum_{t \in \mathcal{O}'_{s,r(p_1,p_2)}} \mathbb{E}[X_t | \mathcal{A}_{p_2}] \right\|_{L_2} \leq 2\gamma.$$

The estimate (7.12) follows from (7.15) and property (P3).  $\square$

The following corollary of Lemma 7.2 is the last ingredient needed for the proof of Proposition 7.1.

**Corollary 7.3.** *Let  $d, n, \kappa, k, \mathbf{X}$  be as in Proposition 7.1, and let  $\gamma$  be as in (7.1). Also let  $N \subseteq [n]$  and  $\mathbf{Y} = \langle Y_p : p \in \text{PartIncr}([d], N) \rangle$  be as in Paragraph 7.1.1. Then for every distinct  $p_1, p_2 \in \text{PartIncr}([d], N)$  such that the pair  $\{p_1, p_2\}$  is aligned we have*

$$(7.16) \quad \left| \mathbb{E}[Y_{p_1} Y_{p_2}] - \mathbb{E}[Y_{p_1 \wedge p_2}^2] \right| \leq 4\gamma.$$

*Proof.* We first observe that

$$(7.17) \quad |\mathbb{E}[Y_{p_1} Y_{p_2}] - \mathbb{E}[Y_{p_1 \wedge p_2}^2]| \leq |\mathbb{E}[Y_{p_2}(Y_{p_1} - Y_{p_1 \wedge p_2})]| + |\mathbb{E}[Y_{p_1 \wedge p_2}(Y_{p_2} - Y_{p_1 \wedge p_2})]|.$$

By (7.8), we see that  $\|Y_{p_2}\|_{L_2} \leq 1$ . Since  $Y_{p_2}$  and  $Y_{p_1 \wedge p_2}$  are  $\mathcal{A}_{p_2}$ -measurable—which follows from property (P1)—by the Cauchy–Schwartz inequality and Lemma 7.2, the first term in the right hand-side of (7.17) can be estimated by

$$(7.18) \quad \begin{aligned} |\mathbb{E}[Y_{p_2}(Y_{p_1} - Y_{p_1 \wedge p_2})]| &= |\mathbb{E}[\mathbb{E}[Y_{p_2}(Y_{p_1} - Y_{p_1 \wedge p_2}) \mid \mathcal{A}_{p_2}]]| \\ &= |\mathbb{E}[Y_{p_2}(\mathbb{E}[Y_{p_1} \mid \mathcal{A}_{p_2}] - Y_{p_1 \wedge p_2})]| \\ &\leq \|\mathbb{E}[Y_{p_1} \mid \mathcal{A}_{p_2}] - Y_{p_1 \wedge p_2}\|_{L_2} \stackrel{(7.11)}{\leq} 2\gamma. \end{aligned}$$

Similarly, we obtain that

$$(7.19) \quad |\mathbb{E}[Y_{p_1 \wedge p_2}(Y_{p_2} - Y_{p_1 \wedge p_2})]| \leq \|\mathbb{E}[Y_{p_2} \mid \mathcal{A}_{p_1 \wedge p_2}] - Y_{p_1 \wedge p_2}\|_{L_2} \stackrel{(7.11)}{\leq} 2\gamma.$$

Inequality (7.16) follows by combining (7.17), (7.18) and (7.19).  $\square$

7.1.4. *Proof of Proposition 7.1.* Let  $N$  be as in (7.2). Moreover, given the random array  $\mathbf{X}$ , let  $\mathbf{Y} = \langle Y_p : p \in \text{PartIncr}([d], N) \rangle$  and  $\mathbf{\Delta} = \langle \Delta_p : p \in \text{PartIncr}([d], N) \rangle$  be the real-valued stochastic processes defined in (7.8) and (7.9) respectively.

We claim that  $N$  and  $\mathbf{\Delta}$  are as desired. To this end we first observe that  $|N| = k$ . For part (i), let  $s \in \binom{N}{d}$  be arbitrary. Notice that for every  $G \subseteq [d]$  the quantity

$$(7.20) \quad \sum_{G \subseteq F \subseteq [d]} (-1)^{|F \setminus G|}$$

is equal to 1 if  $G = [d]$ , and 0 otherwise. Therefore,

$$(7.21) \quad \begin{aligned} \sum_{F \subseteq [d]} \Delta_{I_s \upharpoonright F} &\stackrel{(7.9)}{=} \sum_{F \subseteq [d]} \left( \sum_{G \subseteq F} (-1)^{|F \setminus G|} Y_{I_s \upharpoonright G} \right) \\ &= \sum_{G \subseteq [d]} Y_{I_s \upharpoonright G} \left( \sum_{G \subseteq F \subseteq [d]} (-1)^{|F \setminus G|} \right) \stackrel{(7.20)}{=} Y_{I_s} \stackrel{(7.8)}{=} X_s. \end{aligned}$$

For part (ii), fix  $p \in \text{PartIncr}([d], N)$  with  $p \neq \emptyset$  and observe that

$$(7.22) \quad \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} = 0.$$

Since  $\|Y_\emptyset\|_{L_2} \leq 1$ , by the Cauchy–Schwarz inequality and Lemma 7.2, we obtain that

$$\begin{aligned}
 (7.23) \quad |\mathbb{E}[\Delta_p]| &\stackrel{(7.9)}{=} \left| \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} \mathbb{E}[Y_{p \uparrow G}] \right| \\
 &= \left| \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} \mathbb{E}[\mathbb{E}[Y_{p \uparrow G} \mid \mathcal{A}_\emptyset]] \right| \\
 &\leq \left| \sum_{G \subseteq \text{dom}(p)} (-1)^{|\text{dom}(p) \setminus G|} \mathbb{E}[Y_\emptyset] \right| + \sum_{\emptyset \neq G \subseteq \text{dom}(p)} |\mathbb{E}[\mathbb{E}[Y_{p \uparrow G} \mid \mathcal{A}_\emptyset] - Y_\emptyset]| \\
 &\stackrel{(7.22)}{\leq} \sum_{\emptyset \neq G \subseteq \text{dom}(p)} \|\mathbb{E}[Y_{p \uparrow G} \mid \mathcal{A}_\emptyset] - Y_\emptyset\|_{L_2} \stackrel{(7.11)}{\leq} 2^d \gamma.
 \end{aligned}$$

Finally, for part (iii), let  $p_1, p_2 \in \text{PartIncr}([d], N)$  be distinct such that the pair  $\{p_1, p_2\}$  is aligned. Without loss of generality we may assume that  $\text{dom}(p_1) \setminus r(p_1, p_2) \neq \emptyset$ . (If not, then we will work with  $p_2$ .) By Corollary 7.3, we have

$$\begin{aligned}
 |\mathbb{E}[\Delta_{p_1} \Delta_{p_2}]| &\stackrel{(7.9)}{=} \left| \sum_{\substack{G \subseteq \text{dom}(p_1) \\ H \subseteq \text{dom}(p_2)}} (-1)^{|\text{dom}(p_1) \setminus G|} (-1)^{|\text{dom}(p_2) \setminus H|} \mathbb{E}[Y_{p_1 \uparrow G} Y_{p_2 \uparrow H}] \right| \\
 &\stackrel{(7.16)}{\leq} \left| \sum_{\substack{G \subseteq \text{dom}(p_1) \\ H \subseteq \text{dom}(p_2)}} (-1)^{|\text{dom}(p_1)| + |\text{dom}(p_2)| + |G| + |H|} \mathbb{E}[Y_{p_1 \uparrow G \cap H \cap r(p_1, p_2)}^2] \right| + 2^{2d+2} \gamma;
 \end{aligned}$$

on the other hand, our assumption that  $\text{dom}(p_1) \setminus r(p_1, p_2) \neq \emptyset$  yields that

$$\sum_{\tilde{G} \subseteq \text{dom}(p_1) \setminus r(p_1, p_2)} (-1)^{|\tilde{G}|} = 0,$$

and so,

$$\begin{aligned}
 &\left| \sum_{\substack{G \subseteq \text{dom}(p_1) \\ H \subseteq \text{dom}(p_2)}} (-1)^{|\text{dom}(p_1)| + |\text{dom}(p_2)| + |G| + |H|} \mathbb{E}[Y_{p_1 \uparrow G \cap H \cap r(p_1, p_2)}^2] \right| = \\
 &= \left| \sum_{\substack{K \subseteq r(p_1, p_2) \\ G, H \subseteq r(p_1, p_2) \setminus K \\ G \cap H = \emptyset}} \mathbb{E}[Y_{p_1 \uparrow K}^2] (-1)^{|G| + |H|} \sum_{\tilde{H} \subseteq \text{dom}(p_2) \setminus r(p_1, p_2)} (-1)^{|\tilde{H}|} \sum_{\tilde{G} \subseteq \text{dom}(p_1) \setminus r(p_1, p_2)} (-1)^{|\tilde{G}|} \right| = 0.
 \end{aligned}$$

Therefore,  $|\mathbb{E}[\Delta_{p_1} \Delta_{p_2}]| \leq 2^{2d+2} \gamma$ . The proof of Proposition 7.1 is completed.

**7.2. Proof of Theorem 1.6.** Let  $d$  be a positive integer, let  $\varepsilon > 0$ , and let  $c$  and  $n_0$  be as in (1.7) and (1.8) respectively. Fix an integer  $n \geq n_0$ . We set

$$(7.24) \quad \kappa := \left\lceil \frac{2^{4d+5}}{\varepsilon^2} \right\rceil \quad \text{and} \quad k := \left\lfloor \frac{1}{2^9} \left( \frac{\varepsilon^4}{2^5 d} \right)^{\frac{1}{d+1}} \sqrt[n]{n} \right\rfloor$$

and we observe that  $\kappa \geq 2$  and  $n \geq 2\kappa^2 d(k+1)^{d+1}$ . Let  $N$  be the subset of  $[n]$  obtained by Proposition 7.1 applied for  $n, d$  and the positive integers  $\kappa, k$  defined above. By the choices of  $c, n_0, k$  and the fact that  $|N| = k$ , it is easy to see that  $|N| \geq c \sqrt[n]{n}$ . Next, let  $\mathbf{X}$  be a  $d$ -dimensional, spreadable, random array on  $[n]$  as in Theorem 1.6, and let  $\mathbf{\Delta}$

be the real-valued stochastic process obtained by Proposition 7.1 when applied to  $\mathbf{X}$ . By the definition of the constant  $\gamma$  in (7.1) and using again the choices of  $\kappa$  and  $k$ , it is not hard to check that parts (i), (ii) and (iii) of Proposition 7.1 yield the corresponding parts of Theorem 1.6. Thus, we only need to verify part (iv), that is, the fact that the process  $\Delta$  is (essentially) unique.

Indeed, set

$$(7.25) \quad \ell := \lceil \varepsilon^{-1} + 2^{2d} \rceil, \quad k_0 := \left\lfloor \frac{k - \ell(d-1)}{\ell(d-1) + 1} \right\rfloor \quad \text{and} \quad L := \{I_N((\ell(d-1)+1)j) : j \in [k_0]\}$$

and observe that  $L$  is a subset of  $N$  with  $|L| \geq (\varepsilon^{-1} + 2^{2d})d^{-1}k = (\varepsilon^{-1} + 2^{2d})d^{-1}|N|$ . We will show that the set  $L$  is as desired. To this end, we first observe the following property which follows from the definition of the set  $L$ .

- (A) For every  $p \in \text{PartIncr}([d], L)$  there exists a sequence  $(s_j^p)_{j=1}^\ell$  in  $\binom{N}{d}$  such that for every distinct  $i, j \in [\ell]$  the pair  $\{s_i^p, s_j^p\}$  is aligned in the sense of Definition 6.1, and satisfies  $I_{s_i^p} \wedge I_{s_j^p} = p$ .

Now, let  $\mathbf{Z} = \langle Z_p : p \in \text{PartIncr}([d], N) \rangle$  be a real-valued stochastic process which satisfies parts (i) and (iii) of Theorem 1.6. By part (i) applied for  $\Delta$  and  $\mathbf{Z}$ , for every  $s \in \binom{N}{d}$  we have

$$1 = \|X_s\|_{L_2}^2 = \sum_{F \subseteq [d]} \|\Delta_{I_s \upharpoonright F}\|_{L_2}^2 + \sum_{\substack{F, G \subseteq [d] \\ F \neq G}} \mathbb{E}[\Delta_{I_s \upharpoonright F} \Delta_{I_s \upharpoonright G}]$$

and

$$1 = \|X_s\|_{L_2}^2 = \sum_{F \subseteq [d]} \|Z_{I_s \upharpoonright F}\|_{L_2}^2 + \sum_{\substack{F, G \subseteq [d] \\ F \neq G}} \mathbb{E}[Z_{I_s \upharpoonright F} Z_{I_s \upharpoonright G}]$$

and therefore, by part (iii), for every  $F \subseteq [d]$  we have

$$(7.26) \quad \|\Delta_{I_s \upharpoonright F}\|_{L_2}^2 \leq 1 + 2^{2d}\varepsilon \quad \text{and} \quad \|Z_{I_s \upharpoonright F}\|_{L_2}^2 \leq 1 + 2^{2d}\varepsilon.$$

**Claim 7.4.** *Let  $p \in \text{PartIncr}([d], L)$ , and let  $(s_j^p)_{j=1}^\ell$  be the corresponding sequence in  $\binom{N}{d}$  described in property (A). Then for every  $F \subseteq [d]$  the following hold.*

- (i) *If  $F \subseteq \text{dom}(p)$ , then we have*

$$(7.27) \quad \frac{1}{\ell} \sum_{j=1}^{\ell} \Delta_{I_{s_j^p} \upharpoonright F} = \Delta_{p \upharpoonright F} \quad \text{and} \quad \frac{1}{\ell} \sum_{j=1}^{\ell} Z_{I_{s_j^p} \upharpoonright F} = Z_{p \upharpoonright F}.$$

- (ii) *If  $F \setminus \text{dom}(p) \neq \emptyset$ , then we have*

$$(7.28) \quad \left\| \frac{1}{\ell} \sum_{j=1}^{\ell} \Delta_{I_{s_j^p} \upharpoonright F} \right\|_{L_2} \leq \sqrt{2\varepsilon} \quad \text{and} \quad \left\| \frac{1}{\ell} \sum_{j=1}^{\ell} Z_{I_{s_j^p} \upharpoonright F} \right\|_{L_2} \leq \sqrt{2\varepsilon}.$$

*Proof of Claim 7.4.* By property (A), for every  $j \in [n]$  we have that  $I_{s_j^p} \upharpoonright \text{dom}(p) = p$ ; (7.27) follows from this observation. On the other hand, invoking property (A) again, we see that if  $F \setminus \text{dom}(p) \neq \emptyset$ , then for every distinct  $j_1, j_2 \in [\ell]$  the partial maps

$I_{s_{j_1}^p}$  and  $I_{s_{j_2}^p}$  are distinct and the pair  $\{I_{s_{j_1}^p}, I_{s_{j_2}^p}\}$  is aligned. Taking into account this remark, (7.28) follows from (7.26), the fact that the processes  $\Delta$  and  $Z$  satisfy part (iii) of Theorem 1.6, and the choice of  $\ell$  in (7.25).  $\square$

After this preliminary discussion, for every  $p \in \text{PartIncr}([d], L)$  we will show that

$$(7.29) \quad \|\Delta_p - Z_p\|_{L_2} \leq 2^{(\text{dom}(p)+1)+d+1} \sqrt{2\varepsilon}$$

with the convention that  $\binom{1}{2} = 0$ ; clearly, this is enough to complete the proof. We will proceed by induction on the cardinality of  $\text{dom}(p)$ . If “ $\text{dom}(p) = 0$ ”, then this is equivalently to saying that  $p = \emptyset$ ; in this case, by (7.27) and using the fact that part (i) of Theorem 1.6 is satisfied for  $\Delta$  and  $Z$ , we see that

$$\frac{1}{\ell} \sum_{j=1}^{\ell} X_{s_j^\emptyset} = \Delta_\emptyset + \sum_{\emptyset \neq F \subseteq [d]} \frac{1}{\ell} \sum_{j=1}^{\ell} \Delta_{I_{s_j^\emptyset} \upharpoonright F} \quad \text{and} \quad \frac{1}{\ell} \sum_{j=1}^{\ell} X_{s_j^\emptyset} = Z_\emptyset + \sum_{\emptyset \neq F \subseteq [d]} \frac{1}{\ell} \sum_{j=1}^{\ell} Z_{I_{s_j^\emptyset} \upharpoonright F}$$

and so, by (7.28), we obtain that

$$\|\Delta_\emptyset - Z_\emptyset\|_{L_2} \leq 2^{d+1} \sqrt{2\varepsilon}.$$

Next, let  $u \in [\ell]$  and assume that (7.29) has been proved for every partial map whose domain has size strictly less than  $u$ . Fix  $p \in \text{PartIncr}([d], L)$  with  $|\text{dom}(p)| = u$ . Using again (7.27) and the validity of part (i) of Theorem 1.6 for  $\Delta$  and  $Z$ , we see that

$$\begin{aligned} \frac{1}{\ell} \sum_{j=1}^{\ell} X_{s_j^p} &= \sum_{F \subseteq \text{dom}(p)} \Delta_{p \upharpoonright F} + \sum_{\substack{F \subseteq [d] \\ F \setminus \text{dom}(p) \neq \emptyset}} \frac{1}{\ell} \sum_{j=1}^{\ell} \Delta_{I_{s_j^p} \upharpoonright F} \\ &= \sum_{F \subseteq \text{dom}(p)} Z_{p \upharpoonright F} + \sum_{\substack{F \subseteq [d] \\ F \setminus \text{dom}(p) \neq \emptyset}} \frac{1}{\ell} \sum_{j=1}^{\ell} Z_{I_{s_j^p} \upharpoonright F}. \end{aligned}$$

Invoking this identity, (7.28) and the inductive assumptions, we conclude that

$$\begin{aligned} \|\Delta_p - Z_p\|_{L_2} &= \left\| \sum_{F \not\subseteq \text{dom}(p)} (Z_{p \upharpoonright F} - \Delta_{p \upharpoonright F}) + \sum_{\substack{F \subseteq [d] \\ F \setminus \text{dom}(p) \neq \emptyset}} \frac{1}{\ell} \sum_{j=1}^{\ell} (Z_{I_{s_j^p} \upharpoonright F} - \Delta_{I_{s_j^p} \upharpoonright F}) \right\|_{L_2} \\ &\leq (2^u - 1) 2^{\binom{u}{2} + d + 1} \sqrt{2\varepsilon} + 2(2^d - 2^u) \sqrt{2\varepsilon} \leq 2^{\binom{u+1}{2} + d + 1} \sqrt{2\varepsilon}. \end{aligned}$$

This completes the proof of the general inductive step, and consequently, the entire proof of Theorem 1.6 is completed.

## 8. CONNECTION WITH CONCENTRATION

**8.1. Overview.** We are about to present an application of Theorem 1.4 which supplements the concentration results obtained in [DTV20]. To put things in a proper context, we first recall the main problem addressed in [DTV20].

**Problem 8.1.** Let  $n \geq d \geq 2$  be integers, and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be an approximately spreadable,  $d$ -dimensional random array on  $[n]$  whose entries take values in a finite set  $\mathcal{X}$ . Also let  $f: \mathcal{X}^{\binom{[n]}{d}} \rightarrow \mathbb{R}$  be a function, and assume that  $\mathbb{E}[f(\mathbf{X})] = 0$  and  $\|f(\mathbf{X})\|_{L_p} = 1$  for some  $p > 1$ . Under what condition on  $\mathbf{X}$  can we find a large subset  $I$  of  $[n]$  such that, setting  $\mathcal{F}_I := \sigma(\{X_s : s \in \binom{I}{d}\})$ , the random variable  $\mathbb{E}[f(\mathbf{X}) | \mathcal{F}_I]$  is concentrated around its mean?

Note that Problem 8.1 is somewhat distinct from the traditional setting of concentration of smooth functions (see, e.g., [Le01, BLM13]). It is particularly relevant in a combinatorial context since functions on discrete sets are, usually, highly nonsmooth. We refer the reader to the introduction of [DTV20] for further motivation, and to [DK16] for a broader discussion on this “conditional concentration” and its applications.

8.1.1. *The box independence condition.* In [DTV20] it was shown<sup>10</sup> that an affirmative answer to Problem 8.1 can be obtained if—and essentially only if—the random array  $\mathbf{X}$  satisfies a certain correlation condition to which we refer as the *box independence condition*. In order to state this condition we need to introduce some terminology. Let  $n, d$  be integers with  $n \geq 2d$  and  $d \geq 2$ ; we say that a subset of  $\binom{[n]}{d}$  is a  $d$ -dimensional box of  $[n]$  if it is of the form

$$(8.1) \quad \left\{ s \in \binom{[n]}{d} : |s \cap H_i| = 1 \text{ for all } i \in [d] \right\}.$$

where  $H_1, \dots, H_d$  are 2-element subsets of  $[n]$  which satisfy  $\max(H_i) < \min(H_{i+1})$  for every  $i \in [d-1]$ .

**Definition 8.2** (Box independence condition). Let  $n, d$  be integers with  $n \geq 2d$  and  $d \geq 2$ , let  $\mathcal{X}$  be a finite set with  $|\mathcal{X}| \geq 2$ , and let  $\mathbf{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$  be a  $d$ -dimensional random array on  $[n]$  with  $\mathcal{X}$ -valued entries. Also let  $\vartheta \geq 0$ . We say that  $\mathbf{X}$  satisfies the  $\vartheta$ -box independence condition if there exists a subset  $\mathcal{S}$  of  $\mathcal{X}$  with  $|\mathcal{S}| = |\mathcal{X}| - 1$  such that for every  $d$ -dimensional box  $B$  of  $[n]$  and every  $a \in \mathcal{S}$  we have

$$(8.2) \quad \left| \mathbb{P}\left(\bigcap_{s \in B} [X_s = a]\right) - \prod_{s \in B} \mathbb{P}([X_s = a]) \right| \leq \vartheta.$$

Thus, for instance, if “ $d = 2$ ” and “ $\mathcal{X} = \{0, 1\}$ ”, then the  $\vartheta$ -box independence condition is equivalent to saying that for every  $i, j, k, \ell \in [n]$  with  $i < j < k < \ell$  we have

$$(8.3) \quad \left| \mathbb{E}[X_{\{i,k\}} X_{\{i,\ell\}} X_{\{j,k\}} X_{\{j,\ell\}}] - \mathbb{E}[X_{\{i,k\}}] \mathbb{E}[X_{\{i,\ell\}}] \mathbb{E}[X_{\{j,k\}}] \mathbb{E}[X_{\{j,\ell\}}] \right| \leq \vartheta.$$

8.1.2. The main result in this section—Proposition 8.3 below—is a characterization of the box independence condition in terms of the distributional decomposition obtained in Theorem 1.4; as will become clear in the ensuing discussion, the main advantage of this characterization lies in the fact that that it enables us to employ further analytical and combinatorial tools in the broader context of Problem 8.1. In a nutshell, it asserts

<sup>10</sup>See, in particular, [DTV20, Theorems 1.5 and 2.3].



that a random array  $\mathbf{X}$  satisfies condition (8.2) if and only if its distribution is close to a distribution of the form (1.3) where for “almost every”  $j \in J$  and every  $a \in \mathcal{X}$  the random variable  $h_j^a$  is *box uniform* and its average  $\mathbb{E}[h_j^a]$  is roughly equal to the expected value  $\mathbb{P}([X_{[d]} = a])$ .

8.1.3. *Box uniformity.* The aforementioned box uniformity is a well-known pseudorandomness property—see, e.g., [Ró15]—which is defined using the box norms. Specifically, let  $d \geq 2$  be an integer, let  $(\Omega, \Sigma, \mu)$  be a probability space, and let  $\Omega^d$  be equipped with the product measure. Also let  $\varrho > 0$ . We say that an integrable random variable  $h: \Omega^d \rightarrow \mathbb{R}$  is  $\varrho$ -*box uniform* provided that

$$(8.4) \quad \|h - \mathbb{E}[h]\|_{\square} \leq \varrho$$

where  $\|\cdot\|_{\square}$  denotes the corresponding box norm. (See Subsection 3.1.)

8.2. **The characterization.** We have the following proposition.

**Proposition 8.3.** *Let  $d, m \geq 2$  be integers, and let  $0 < \varepsilon \leq 1$ . Let  $C = C(d, m, 2d, \varepsilon)$  be as in (1.4), let  $n, \mathcal{X}, \mathbf{X}$  be as in Theorem 1.4, and set  $\delta_a := \mathbb{P}([X_{[d]} = a])$  for every  $a \in \mathcal{X}$ . Finally, let  $J, \Omega, \boldsymbol{\lambda} = \langle \lambda_j : j \in J \rangle$  and  $\mathcal{H} = \langle h_j^a : j \in J, a \in \mathcal{X} \rangle$  be as in Theorem 1.4 when applied to the random array  $\mathbf{X}$  for the parameters  $d, m, \varepsilon$  and  $k = 2d$ . Then the following hold.*

(i) *Let  $\varrho > 0$ , and set*

$$(8.5) \quad \vartheta = \vartheta(d, \varepsilon, \varrho) := 2^d(2\varepsilon + 4\varrho).$$

*Assume that there is a subset  $G$  of  $J$  such that: (a)  $\sum_{j \in G} \lambda_j \geq 1 - \varrho$ , and (b) for every  $j \in G$  and every  $a \in \mathcal{X}$  we have  $|\mathbb{E}[h_j^a] - \delta_a| \leq \varrho$  and  $\|h_j^a - \mathbb{E}[h_j^a]\|_{\square} \leq \varrho$ . Then,  $\mathbf{X}$  is  $\vartheta$ -box independent.*

(ii) *Conversely, let  $\vartheta > 0$ , and set*

$$(8.6) \quad \varrho = \varrho(d, m, \varepsilon, \vartheta) := 2^{d+7} m^3 (\varepsilon^{1/12^d} + \vartheta^{1/12^d}).$$

*Assume that  $\mathbf{X}$  is  $\vartheta$ -box independent. Then there exists a subset  $G$  of  $J$  such that: (a)  $\sum_{j \in G} \lambda_j \geq 1 - \varrho$ , and (b) for every  $j \in G$  and every  $a \in \mathcal{X}$  we have  $|\mathbb{E}[h_j^a] - \delta_a| \leq \varrho$  and  $\|h_j^a - \mathbb{E}[h_j^a]\|_{\square} \leq \varrho$ .*

*Proof.* First we argue for part (i). Let  $B$  be a  $d$ -dimensional box of  $[n]$ , and fix  $a \in \mathcal{X}$ . By (1.6) and part (a) of our assumptions, we have

$$(8.7) \quad \left| \mathbb{P}\left(\bigcap_{s \in B} [X_s = a]\right) - \sum_{j \in G} \lambda_j \int \prod_{s \in B} h_j^a(\omega_s) d\boldsymbol{\mu}_j(\omega) \right| \leq \varepsilon + \varrho.$$

Let  $j \in G$  be arbitrary, and observe that  $\|h_j^a\|_{\square} \leq 1$ . By the  $\varrho$ -box uniformity of  $h_j^a$ , the Gowers–Cauchy–Schwarz inequality (3.6) and a telescopic argument, we see that

$$(8.8) \quad \left| \int \prod_{s \in B} h_j^a(\omega_s) d\boldsymbol{\mu}_j(\omega) - \prod_{s \in B} \mathbb{E}[h_j^a] \right| \leq 2^d \varrho$$

and so, using the fact  $|\mathbb{E}[h_j^a] - \delta_a| \leq \varrho$ , we obtain that

$$(8.9) \quad \left| \int \prod_{s \in B} h_j^a(\omega_s) d\mu_j(\omega) - \delta_a^{2^d} \right| \leq 2^{d+1} \varrho.$$

On the other hand, since  $\mathbf{X}$  is  $(1/C)$ -spreadable, we have

$$(8.10) \quad \left| \delta_a^{2^d} - \prod_{s \in B} \mathbb{P}([X_s = a]) \right| \leq \frac{2^d}{C}.$$

By (8.7)–(8.10), assumption (a), the fact that  $1/C \leq \varepsilon$  and the choice of  $\vartheta$  in (8.5), we conclude that

$$\left| \mathbb{P}\left(\bigcap_{s \in B} [X_s = a]\right) - \prod_{s \in B} \mathbb{P}([X_s = a]) \right| \leq \vartheta$$

which yields that the random array  $\mathbf{X}$  is  $\vartheta$ -box independent.

We proceed to the proof of part (ii). We will need the following fact which follows from [DTV20, Lemma 4.6 and Subsection 5.2] and the fact that  $n \geq C \geq \varepsilon^{-1}$ . It shows that the box independence condition is inherited to subsets of  $d$ -dimensional boxes.

**Fact 8.4.** *Let the notation and assumptions be as in part (ii) of Proposition 8.3, and set*

$$(8.11) \quad \Theta := 100 2^{2d} m^{2^d} (2\varepsilon^{1/4^d} + \vartheta^{1/4^d}).$$

*Then for every  $d$ -dimensional box  $B$  of  $[n]$ , every nonempty subset  $F$  of  $B$  and every  $a \in \mathcal{X}$  we have*

$$(8.12) \quad \left| \mathbb{P}\left(\bigcap_{s \in F} [X_s = a]\right) - \prod_{s \in F} \mathbb{P}([X_s = a]) \right| \leq \Theta.$$

Now, fix  $a \in \mathcal{X}$ , and set  $s_1 := \{2i - 1 : i \in [d]\} \in \binom{[n]}{d}$  and  $s_2 := \{2i : i \in [d]\} \in \binom{[n]}{d}$ . By (1.6), we have

$$(8.13) \quad \left| \delta_a - \sum_{j \in J} \lambda_j \mathbb{E}[h_j^a] \right| \leq \varepsilon,$$

$$(8.14) \quad \left| \mathbb{P}([X_{s_1} = a] \cap [X_{s_2} = a]) - \sum_{j \in J} \lambda_j \mathbb{E}[h_j^a]^2 \right| \leq \varepsilon.$$

By Fact 8.4, the  $(1/C)$ -spreadability of  $\mathbf{X}$  and (8.14), we see that

$$(8.15) \quad \left| \delta_a^2 - \sum_{j \in J} \lambda_j \mathbb{E}[h_j^a]^2 \right| \leq \varepsilon + \Theta + \frac{2}{C}.$$

Thus, setting

$$(8.16) \quad \varrho_1 := 2m(4\varepsilon + \Theta)^{1/4},$$

by (8.13), (8.15), the fact that  $1/C \leq \varepsilon$ , Chebyshev's inequality and a union bound, we obtain a subset  $G_1$  of  $J$  such that  $\sum_{j \in G_1} \lambda_j \geq 1 - \varrho_1$  and  $|\mathbb{E}[h_j^a] - \delta_a| \leq \varrho_1$  for every  $j \in G_1$  and every  $a \in \mathcal{X}$ .

Again, let  $a \in \mathcal{X}$  be arbitrary. We shall estimate the quantity

$$(8.17) \quad \sum_{j \in G_1} \lambda_j \|h_j^a - \mathbb{E}[h_j^a]\|_{\square}^{2^d} \stackrel{(3.5)}{=} \sum_{H \subseteq \{0,1\}^d} (-1)^{2^d - |H|} \times \\ \times \left( \sum_{j \in G_1} \lambda_j \mathbb{E}[h_j^a]^{2^d - |H|} \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) \right).$$

(Here, as in Section 2, we use the convention that the product of an empty family of functions is equal to the constant function 1.) To this end, let  $H$  be an arbitrary subset of  $\{0,1\}^d$ . Notice first that, by the choice of  $G_1$ , we have

$$(8.18) \quad \left| \sum_{j \in G_1} \lambda_j \mathbb{E}[h_j^a]^{2^d - |H|} \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) - \delta_a^{2^d - |H|} \sum_{j \in J} \lambda_j \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) \right| \leq (2^d + 1)\varrho_1.$$

Next observe that if  $H$  is nonempty, then, by (1.6), we may select a  $d$ -dimensional box  $B$  of  $[n]$  and a nonempty subset  $F$  of  $B$  with  $|F| = |H|$  and such that

$$(8.19) \quad \left| \mathbb{P}\left(\bigcap_{s \in F} [X_s = a]\right) - \sum_{j \in J} \lambda_j \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) \right| \leq \varepsilon.$$

Thus, by Fact 8.4, the  $(1/C)$ -spreadability of  $\mathbf{X}$  and the fact that  $|F| = |H|$ , we have

$$(8.20) \quad \left| \delta_a^{|H|} - \sum_{j \in J} \lambda_j \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) \right| \leq \varepsilon + \Theta + \frac{2^d}{C}.$$

By (8.18) and (8.20), we see that for every (possibly empty) subset  $H$  of  $\{0,1\}^d$ ,

$$(8.21) \quad \left| \sum_{j \in G_1} \lambda_j \mathbb{E}[h_j^a]^{2^d - |H|} \int \prod_{\epsilon \in H} h_j^a(\omega_\epsilon) d\mu(\omega) - \delta_a^{2^d - |H|} \right| \leq \varepsilon + \Theta + \frac{2^d}{C} + (2^d + 1)\varrho_1.$$

By (8.17) and (8.21), we conclude that for every  $a \in \mathcal{X}$  we have

$$(8.22) \quad \sum_{j \in G_1} \lambda_j \|h_j^a - \mathbb{E}[h_j^a]\|_{\square}^{2^d} \leq 2^d \left( \varepsilon + \Theta + \frac{2^d}{C} + (2^d + 1)\varrho_1 \right).$$

Using once again the fact that  $1/C \leq \varepsilon$ , (8.22), the choice of  $\varrho$ ,  $\Theta$  and  $\varrho_1$  in (8.6), (8.11) and (8.16) respectively, Markov's inequality and a union bound, we may select a subset  $G$  of  $G_1$  with  $\sum_{j \in G} \lambda_j \geq 1 - \varrho$  and such that  $\|h_j^a - \mathbb{E}[h_j^a]\|_{\square} \leq \varrho$  for every  $j \in G$  and every  $a \in \mathcal{X}$ . Since  $G \subseteq G_1$  and  $\varrho_1 \leq \varrho$ , we see that  $G$  is as desired. The proof of Proposition 8.3 is completed.  $\square$

## APPENDIX A. PROOF OF LEMMA 3.4

Let  $d, m, \varepsilon$  be as in the statement of Lemma 3.4, and let  $n_0$  be as in (3.8). Fix coefficients  $\lambda_1, \dots, \lambda_m \geq 0$  with  $\lambda_1 + \dots + \lambda_m = 1$ , and let  $V$  be a finite set with  $|V| \geq n_0$ . Observe that, by the choice of  $n_0$ , we have

$$(A.1) \quad \left( \frac{1}{|V|^{d/3}} + \frac{(1+d!m)2d^2}{|V|} \right)^{1/2^d} \leq \varepsilon \quad \text{and} \quad 1 - 2m \exp\left(-\frac{2}{d!4^d} |V|^{d/3}\right) > 0.$$

We define an equivalence relation  $\sim$  on  $V^d$  by setting

$$(A.2) \quad (v_1, \dots, v_d) \sim (v'_1, \dots, v'_d) \Leftrightarrow \text{there exists a permutation } \pi \text{ of } [d] \\ \text{such that } v'_i = v_{\pi(i)} \text{ for all } i \in [d],$$

and for every  $e \in V^d$  by  $[e] := \{e' \in V^d : e' \sim e\}$  we denote the  $\sim$ -equivalence class of  $e$ . We also set  $\text{Sym}(V^d) := V^d / \sim$ .

Next, we fix a collection  $\mathbf{X} = \langle X_{\mathbf{e}} : \mathbf{e} \in \text{Sym}(V^d) \rangle$  of  $[m]$ -valued, independent random variables defined on some probability space  $(\Omega, \Sigma, \mathbb{P})$  which satisfy  $\mathbb{P}([X_{\mathbf{e}} = j]) = \lambda_j$  for every  $\mathbf{e} \in \text{Sym}(V^d)$  and every  $j \in [m]$ . Moreover, for every  $j \in [m]$  we define  $f_j : [m]^{\text{Sym}(V^d)} \rightarrow \mathbb{R}^+$  by setting for every  $\mathbf{x} = (x_{\mathbf{e}})_{\mathbf{e} \in \text{Sym}(V^d)} \in [m]^{\text{Sym}(V^d)}$ ,

$$(A.3) \quad f_j(\mathbf{x}) = \|\mathbf{1}_{\{e \in V^d : x_{[e]} = j\}} - \lambda_j\|_{\square}^{2^d}.$$

Recall that for every  $\mathbf{v} = (v_1^0, v_1^1, \dots, v_d^0, v_d^1) \in V^{2d}$  and every  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_d) \in \{0, 1\}^d$  we set  $\mathbf{v}_{\boldsymbol{\epsilon}} = (v_1^{\epsilon_1}, \dots, v_d^{\epsilon_d}) \in V^d$ . Let  $\mathcal{I}$  denote the subset of  $V^{2d}$  consisting of all strings with distinct entries. Observe that for every  $\mathbf{x} = (x_{\mathbf{e}})_{\mathbf{e} \in \text{Sym}(V^d)} \in [m]^{\text{Sym}(V^d)}$  and every  $j \in [m]$  we have

$$(A.4) \quad f_j(\mathbf{x}) \stackrel{(3.5)}{=} \frac{1}{|V|^{2d}} \sum_{\mathbf{v} \in V^{2d}} \prod_{\boldsymbol{\epsilon} \in \{0,1\}^d} (\mathbf{1}_{\{j\}}(x_{[\mathbf{v}_{\boldsymbol{\epsilon}]})} - \lambda_j) \\ = \frac{1}{|V|^{2d}} \sum_{\mathbf{v} \in \mathcal{I}} \sum_{H \subseteq \{0,1\}^d} (-\lambda_j)^{2^d - |H|} \prod_{\boldsymbol{\epsilon} \in H} \mathbf{1}_{\{j\}}(x_{[\mathbf{v}_{\boldsymbol{\epsilon}]})} + \\ + \frac{1}{|V|^{2d}} \sum_{\mathbf{v} \in V^{2d} \setminus \mathcal{I}} \prod_{\boldsymbol{\epsilon} \in \{0,1\}^d} (\mathbf{1}_{\{j\}}(x_{[\mathbf{v}_{\boldsymbol{\epsilon}]})} - \lambda_j)$$

and

$$(A.5) \quad \left| \frac{1}{|V|^{2d}} \sum_{\mathbf{v} \in V^{2d} \setminus \mathcal{I}} \prod_{\boldsymbol{\epsilon} \in \{0,1\}^d} (\mathbf{1}_{\{j\}}(x_{[\mathbf{v}_{\boldsymbol{\epsilon}]})} - \lambda_j) \right| \leq \frac{|V^{2d} \setminus \mathcal{I}|}{|V|^{2d}} \leq \frac{2d^2}{|V|}.$$

On the other hand, by the definition of  $\mathcal{I}$ , for every  $\mathbf{v} = (v_1^0, v_1^1, \dots, v_d^0, v_d^1) \in \mathcal{I}$  and every distinct  $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \in \{0, 1\}^d$  we have  $[\mathbf{v}_{\boldsymbol{\epsilon}_1}] \neq [\mathbf{v}_{\boldsymbol{\epsilon}_2}]$ . Therefore, by the independence of the entries of  $\mathbf{X}$  and linearity of expectation, for every  $\mathbf{v} = (v_1^0, v_1^1, \dots, v_d^0, v_d^1) \in \mathcal{I}$  and every  $j \in [m]$  we have

$$(A.6) \quad \mathbb{E} \left[ \sum_{H \subseteq \{0,1\}^d} (-\lambda_j)^{2^d - |H|} \prod_{\boldsymbol{\epsilon} \in H} \mathbf{1}_{\{j\}}(X_{[\mathbf{v}_{\boldsymbol{\epsilon}]})} \right] = \sum_{H \subseteq \{0,1\}^d} (-\lambda_j)^{2^d - |H|} \lambda_j^{|H|} = 0.$$

By (A.4), (A.5) and (A.6), for every  $j \in [m]$  we obtain that

$$(A.7) \quad \mathbb{E}[f_j(\mathbf{X})] \leq \frac{2d^2}{|V|}.$$

Moreover, by (A.4) again, if  $\mathbf{x} = (x_{\mathbf{e}})_{\mathbf{e} \in \text{Sym}(V^d)}$  and  $\mathbf{y} = (y_{\mathbf{e}})_{\mathbf{e} \in \text{Sym}(V^d)}$  in  $[m]^{\text{Sym}(V^d)}$  are such that  $|\{\mathbf{e} \in \text{Sym}(V^d) : x_{\mathbf{e}} \neq y_{\mathbf{e}}\}| \leq 1$ , then for every  $j \in [m]$  we have

$$(A.8) \quad |f_j(\mathbf{x}) - f_j(\mathbf{y})| \leq d! \left(\frac{2}{|V|}\right)^d.$$

By (A.8) and the bounded differences inequality (see, e.g., [BLM13, Theorem 6.2]), for every  $j \in [m]$  and every  $t > 0$  we have the estimate

$$(A.9) \quad \mathbb{P}\left(|f_j(\mathbf{X}) - \mathbb{E}[f_j(\mathbf{X})]| > t\right) \leq 2 \exp\left(-\frac{2t^2|V|^d}{d!4^d}\right).$$

Applying (A.9) for “ $t = 1/|V|^{d/3}$ ” and using (A.7), for every  $j \in [m]$  we have

$$(A.10) \quad \mathbb{P}\left(f_j(\mathbf{X}) \leq \frac{1}{|V|^{d/3}} + \frac{2d^2}{|V|}\right) \geq 1 - 2 \exp\left(-\frac{2}{d!4^d} |V|^{d/3}\right).$$

By (A.10), a union bound and (A.1), we select  $\omega_0 \in \Omega$  which belongs to the event

$$(A.11) \quad \bigcap_{j \in [m]} \left[f_j(\mathbf{X}) \leq \frac{1}{|V|^{d/3}} + \frac{2d^2}{|V|}\right].$$

We select a partition  $\langle D_1, \dots, D_m \rangle$  of  $\text{Sym}(V^d)$  into nonempty sets such that

$$(A.12) \quad |D_j \triangle \{\mathbf{e} \in \text{Sym}(V^d) : X_{\mathbf{e}}(\omega_0) = j\}| \leq m$$

for every  $j \in [m]$ . Finally, we set  $E_j := \{e \in V^d : [e] \in D_j\}$  for every  $j \in [m]$ . By (A.8), the choice of  $\omega_0$  and (A.1), we see that the partition  $\langle E_1, \dots, E_m \rangle$  is as desired. The proof of Lemma 3.4 is completed.

## REFERENCES

- [Ald81] D. J. Aldous, *Representations for partially exchangeable arrays of random variables*, J. Multivariate Anal. 11 (1981), 581–598.
- [Ald85] D. J. Aldous, *Exchangeability and related topics*, In “École d’été de probabilités de Saint-Flour XIII, 1983”, Lecture Notes in Mathematics, vol. 1117, Springer, 1985, 1–198.
- [Ald10] D. J. Aldous, *More uses of exchangeability: representations of complex random structures*, in “Probability and mathematical genetics. Papers in honour of Sir John Kingman”, London Mathematical Society Lecture Note Series, Vol. 378, Cambridge University Press, 2010, 35–63.
- [Au08] T. Austin, *On exchangeable random variables and the statistics of large graphs and hypergraphs*, Probability Surveys 5 (2008), 80–145.
- [Au13] T. Austin, *Exchangeable random arrays*, preprint (2013), available at <https://www.math.ucla.edu/~tim/ExchnotesforIISc.pdf>.
- [BLM13] S. Boucheron, G. Lugosi and P. Massart, *Concentration Inequalities. A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.
- [DJ08] P. Diaconis and S. Janson, *Graph limits and exchangeable random graphs*, Rend. Mat. Appl. 28 (2008), 33–61.
- [DF80] P. Diaconis and D. Freedman, *Finite exchangeable sequences*, Ann. Probab. 8 (1980), 745–764.

- [DK16] P. Dodos and V. Kanellopoulos, *Ramsey Theory for Product Spaces*, Mathematical Surveys and Monographs, Vol. 212, American Mathematical Society, 2016.
- [DKK20] P. Dodos, V. Kanellopoulos and Th. Karageorgos,  *$L_p$  regular sparse hypergraphs: box norms*, *Fund. Math.* 248 (2020), 49–77.
- [DTV20] P. Dodos, K. Tyros and P. Valettas, *Concentration estimates for functions of finite high-dimensional random arrays*, preprint (2021).
- [ES81] B. Efron and C. Stein, *The jackknife estimate of variance*, *Ann. Statist.* 9 (1981), 586–596.
- [FT85] D. H. Fremlin and M. Talagrand, *Subgraphs of random graphs*, *Trans. Amer. Math. Soc.* 291 (1985), 551–582.
- [Go07] W. T. Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, *Ann. Math.* 166 (2007), 897–946.
- [GT10] B. Green and T. Tao, *Linear equations in primes*, *Ann. Math.* 171 (2010), 1753–1850.
- [Hoe48] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, *Ann. Math. Stat.* 19 (1948), 293–325.
- [Hoo79] D. N. Hoover, *Relations on probability spaces and arrays of random variables*, preprint (1979), available at <https://www.stat.berkeley.edu/~aldous/Research/hover.pdf>.
- [Kal92] O. Kallenberg, *Symmetries on random arrays and set-indexed processes*, *J. Theor. Probab.* 5 (1992), 727–765.
- [Kal05] O. Kallenberg, *Probabilistic symmetries and invariance principles*, *Probability and its Applications* (New York), Springer, 2005.
- [Le01] M. Ledoux, *The Concentration of Measure Phenomenon*, *Mathematical Surveys and Monographs*, Vol. 89, American Mathematical Society, 2001.
- [Ra30] F. P. Ramsey, *On a problem of formal logic*, *Proc. London Math. Soc.* 30 (1930), 264–286.
- [Ró15] V. Rödl, *Quasi-randomness and the regularity method in hypergraphs*, in “Proceedings of the International Congress of Mathematicians” Vol. I, 571–599, 2015.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ATHENS, PANEPISTIMIOPOLIS 157 84, ATHENS, GREECE  
*Email address:* `pdodos@math.uoa.gr`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ATHENS, PANEPISTIMIOPOLIS 157 84, ATHENS, GREECE  
*Email address:* `ktyros@math.uoa.gr`

MATHEMATICS DEPARTMENT, UNIVERSITY OF MISSOURI, COLUMBIA, MO, 65211  
*Email address:* `valettasp@missouri.edu`