The Sample Complexity of Replicable Realizable PAC Learning

Kasper Green Larsen* Markus Engelund Mathiasen* Chirag Pabbaraju†
Clement Svendsen*

Abstract

In this paper, we consider the problem of replicable realizable PAC learning. We construct a particularly hard learning problem and show a sample complexity lower bound with a close to $(\log|\mathcal{H}|)^{3/2}$ dependence on the size of the hypothesis class $\mathcal{H}$. Our proof uses several novel techniques and works by defining a particular Cayley graph associated with $\mathcal{H}$ and analyzing a suitable random walk on this graph by examining the spectral properties of its adjacency matrix. Furthermore, we show an almost matching upper bound for the lower bound instance, meaning *if* a stronger lower bound exists, one would have to consider a different instance of the problem.

1 Introduction

Replicability in science is a notion of being able to replicate the work of fellow scientists with your own experiments and data. In recent years, replicability has become a more discussed topic as it has been pointed out in multiple Nature articles that we might face what is called a *reproducibility crisis* [1, 2]. When it comes to algorithms in machine learning, one might argue that these are perfectly reproducible, as long as the researchers share the source code along with the training data and the internal randomness used by the algorithm. However, one might ask if it would be possible to design algorithms which give the same result even without sharing the training data. This would have the advantage of other researchers being able to verify that the training data was not cherry-picked, since they can run the algorithm on their own training data. The notion of *replicable* algorithms was introduced by Impagliazzo et al. [3] as a theoretical property of learning algorithms which captures this notion of being able to replicate the output of the algorithm even with new training data. More formally, they define a ρ-replicable learning algorithm as follows.

Definition 1.1 (ρ-replicability [3])**.** Let $\mathcal{A}$ be a randomized algorithm. Then $\mathcal{A}$ is ρ-replicable if there exists an $n \in \mathbb{N}$ such that for all distributions $\mathcal{D}$ over some domain $\mathcal{X}$, it holds that

$$\Pr_{S_1,S_2,r}[\mathcal{A}(S_1;r) = \mathcal{A}(S_2;r)] \geq 1-\rho$$

where $S_1, S_2 \sim \mathcal{D}^n$ are independent, and r denotes the internal randomness used by $\mathcal{A}$.

That is, as long as $\mathcal{A}$ sees a sample from the same underlying distribution $\mathcal{D}$, then it will with high probability output the same classifier. We should remark that even though $\mathcal{A}$ is run on different training data, we still require the same internal randomness to be used for both runs. This turns out to be a necessary condition for many scenarios if we want such a guarantee, since

*Computer Science Department, Aarhus University, Email: {larsen, markusm, clementks}@cs.au.dk
†Computer Science Department, Stanford University, Email: cpabbara@stanford.edu

otherwise, simple tasks such as mean estimation become impossible to do ρ-replicably if we don't share the internal randomness [4].

In this work, we consider the setup of probably approximately correct learning (PAC learning) [5], which is the classic theoretical model for supervised learning. More specifically, we work with *binary classification* in the *realizable* setting. In this setup, we are interested in designing an algorithm which, with high probability, produces a classifier with good accuracy on new data.

More formally, a learning problem consists of a domain $\mathcal{X}$, a label space $\{0,1\}$, a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, an unknown true hypothesis $h^\star \in \mathcal{H}$, and an unknown distribution $\mathcal{D}$ over $\mathcal{X}$. We denote the error of a classifier h as $\mathrm{er}_{\mathcal{D}}(h) = \Pr_{x\sim\mathcal{D}}[h(x) \neq h^\star(x)]$. Then, we say that a (randomized) algorithm $\mathcal{A}$ is a PAC learner for $\mathcal{H}$ if there exists a function $n : (0,1)^2 \to \mathbb{N}$ such that for any $\varepsilon, \delta \in (0,1)$, $h^\star \in \mathcal{H}$ and distribution $\mathcal{D}$, if $\mathcal{A}$ is given at least $n(\varepsilon,\delta)$ i.i.d. samples from $\mathcal{D}$ labeled according to $h^\star$, then with probability at least $1-\delta$ over the randomness of the samples and the internal randomness of $\mathcal{A}$, the classifier g produced by $\mathcal{A}$ will have $\mathrm{er}_{\mathcal{D}}(g) \leq \varepsilon$. The function n is referred to as the *sample complexity* of $\mathcal{A}$.

PAC learning for binary classification has been studied extensively [6–13], and it is known that learnability is characterized by the VC dimension of $\mathcal{H}$ and that the optimal sample complexity in the realizable setting is $\Theta\big(\frac{1}{\varepsilon}(\mathrm{VC}(\mathcal{H}) + \log(1/\delta))\big)$ [6, 10, 11]. Replicable PAC learning differs from classical PAC learning in the sense that it is the Littlestone dimension [14] rather than the VC dimension that characterizes learnability. In particular, in [15, Cor. 2], it is shown that a class that is *privately* learnable has finite Littlestone dimension, whereas in [16, Thm 3.1], it is shown that replicability implies privacy. Hence, a class that is replicably PAC learnable necessarily has finite Littlestone dimension. However, the sample complexity of replicable PAC learning is not fully understood. In the agnostic setting [1], there is an upper bound on the sample complexity for infinite $\mathcal{H}$ which is *exponential* in the Littlestone dimension [17, Thm 1.4] while for finite $\mathcal{H}$, there is an upper bound which is *quadratic* in $\log|\mathcal{H}|$ with an almost matching lower bound [16, Thm 5.13].

1.1 Main Results

The main result of this paper is a lower bound for replicable realizable PAC learning with finite hypothesis classes.

Theorem 1.2 (Replicable Learning Lower Bound)**.** *For any integer $d \geq 10^{11}$, and positive reals $\varepsilon, \delta, \rho \leq 10^{-4}$, there exists a domain $\mathcal{X}$, a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ with VC-dimension d, such that for any algorithm $\mathcal{A}$ there is a distribution $\mathcal{D}$ over $\mathcal{X}$ for which $\mathcal{A}$ needs at least*

$$n = \widetilde{\Omega}\left(\frac{(\log|\mathcal{H}|)^{3/2}}{\varepsilon}\right)$$

labeled samples from $\mathcal{D}$ in order to be a ρ-replicable PAC learner for $\mathcal{H}$ with error ε and failure probability δ. Here $\widetilde{\Omega}$ hides logarithmic factors in $\log|\mathcal{H}|$ and $1/\varepsilon$.

This is the first lower bound for replicable realizable PAC learning beyond the lower bounds for the non-replicable setup where one only needs $\Omega(\frac{1}{\varepsilon}(\log|\mathcal{H}| + \log(1/\delta))$. While our lower bound doesn't scale with ρ and δ, it does show a stronger dependence on $\log|\mathcal{H}|$. As mentioned in the introduction, it is already known that one needs $\Omega(\log^2|\mathcal{H}|)$ samples if we go to the agnostic setting.

[1] In agnostic PAC learning, the data is not assumed to be labeled by some $h^\star \in \mathcal{H}$, but is instead drawn from a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, and the error is measured relative to the hypothesis in $\mathcal{H}$ that has the best error with respect to $\mathcal{D}$.

A natural question is therefore if the true dependence is in fact $(\log|\mathcal{H}|)^2$. It turns out that for the instances we consider in the lower bound, this is not the case, since we can construct an algorithm with an almost matching upper bound for these instances.

Theorem 1.3 (Replicable Learning Upper Bound)**.** *There exists an algorithm $\mathcal{A}$ such that for every instance $(\mathcal{X}, \mathcal{H}, \mathcal{D})$ shown to be hard in the proof of Theorem 1.2, and for every $\varepsilon, \delta, \rho \in (0,1)$, $\mathcal{A}$ is a ρ-replicable PAC learner on this instance with sample complexity*

$$n = \widetilde{O}\left(\frac{(\log|\mathcal{H}|)^{3/2}}{\rho\varepsilon}\right).$$

This means that $(\log|\mathcal{H}|)^{3/2}$ is not just an artifact of our proof, and if a stronger lower bound exists, one has to change our instances in some way to prove it. However, it remains an open question whether such an instance exists, or if one can construct an upper bound which applies to all instances.

Beyond the above two theorems, we believe that the main contribution of this paper lies on the novel technical ideas used in both the upper and lower bound. We describe these in more detail in Section 2 along with an overview of the proofs. We believe the techniques presented here could prove useful in other contexts.

Open Problems. Our results naturally raise the intriguing question of pin pointing the exact sample complexity of replicable realizable PAC learning with finite $\mathcal{H}$. We make the careful conjecture that a ρ^{-1} dependency on the replicability parameter is in fact possible for arbitrary $\mathcal{H}$ and distribution $\mathcal{D}$. Regarding the dependency on $\log|\mathcal{H}|$, we are more divided. Our difficulties in extending our upper bound to other hypothesis sets $\mathcal{H}$ might suggest a $(\log|\mathcal{H}|)^2$ lower bound. On the other hand, our hard instance $(\mathcal{X}, \mathcal{H}, \mathcal{D})$ used in the lower bound is very similar to canonically hard instances in standard realizable PAC learning with no replicability requirements and thus could on the other hand suggest that $(\log|\mathcal{H}|)^{3/2}$ is the right behavior. We hope our work may inspire further progress in understanding replicable learning.

1.2 Further Related Work

Since the paper by Impagliazzo et al. [3], replicable algorithms have been designed for a wide variety of problems such as clustering [18], learning half spaces [3, 19], online learning [20], mean estimation [21], reinforcement learning [22] and distribution testing [23]. Furthermore, there have been interesting connections to other notions of stability such as global stability [24] and differential privacy [16]. Various extensions and generalizations such as list-replicability [24] and approximate replicability [25] have also been considered.

Lastly, a very recent paper from Hopkins et al. [25] studies agnostic PAC learning under various weaker notions of *approximate* replicability. Notably, for "pointwise" replicable algorithms, they show upper and lower bounds on the sample complexity which are linear in the VC dimension in the realizable setting. Hence, our lower bound shows a separation between fully replicable PAC learning and PAC learning with this weaker notion of replicability.

Hopkins et al. [21] show a direct way to prove a lower bound on replicable mean estimation. Their approach considers a d-dimensional cube consisting of all combinations of d biased coins where each coordinate corresponds to the bias of a coin. By picking the biases randomly, they show that the algorithm has to change output distribution appropriately in order to be replicable most of the time, meaning it will need a lot of samples. This approach is similar to our approach in the

lower bound proof except that we consider a graph instead of a cube. Unfortunately, one cannot directly apply their result to get good lower bounds for realizable PAC learning, since they rely on the fact that changing the bias of a coin slightly doesn't change the output distribution of any algorithm that much. This is not necessarily the case for the instances we construct in Section 2.

1.3 Notation

We briefly introduce some notation. Fix numbers $k, d \in \mathbb{N}$ where k is prime, a space $\mathcal{X}$, hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and $\mathcal{D}$ a distribution on $\mathcal{X}$. Then we denote by $\mathbb{Z}_k$ the group of integers with addition mod k. We identify the elements of $\mathbb{Z}_k$ with their representatives in $\{0, \ldots, k-1\}$, yielding a total ordering on $\mathbb{Z}_k$. We also consider the vector space $\mathbb{Z}_k^d$ over the field $\mathbb{Z}_k$ and equip it with the mod k inner product denoted $\langle \cdot, \cdot \rangle$. Furthermore, for hypotheses $h_1, h_2 \in \mathcal{H}$, let the error of h_2 with respect to h_1 be defined as

$$\mathrm{er}_{h_1}(h_2) = \Pr_{x \sim \mathcal{D}}[h_2(x) \neq h_1(x)].$$

For a sample $S = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $h(S) = (h(x_1,) \ldots, h(x_n))$. We denote by log the natural logarithm and $\log_b$ the base-b logarithm. For a positive integer m, we define the set $[m] = \{0, \ldots, m-1\}$.

2 Technical Overview

In this section, we present the high-level ideas of our new lower bound for replicable realizable PAC learning, followed by the main ideas in a near-matching upper bound for the same data distribution and hypothesis set.

In both our lower and upper bounds, we consider the input domain $\mathcal{X} = [d] \times \mathbb{Z}_k$ for a prime k. Our hypothesis set $\mathcal{H}$ contains a hypothesis h_i for every d-tuple $i = (i_0, \ldots, i_{d-1}) \in \mathbb{Z}_k^d$. For a point $(a, b) \in \mathcal{X}$, we have $h_i((a,b)) = 1$ if $i_a \leq b < i_a + \lfloor k/2 \rfloor$ or if $b < i_a + \lfloor k/2 \rfloor < i_a$. Otherwise $h_i((a,b)) = 0$. Each h_i thus corresponds to outputting 1 on d intervals of length $\lfloor k/2 \rfloor$ with wrap-around, where the a^{th} interval starts at i_a. See Figure 1 for an illustration of these intervals. Note that $|\mathcal{H}| = k^d$.

Let $\mathcal{D}$ be the uniform distribution on $\mathcal{X}$. Let $h^\star = h_{i^\star} \in \mathcal{H}$ be an unknown target function and assume training samples are drawn by sampling $S = x_1, \ldots, x_n$ i.i.d. from $\mathcal{D}$ and constructing the training set $(x_1, h^\star(x_1)), \ldots, (x_n, h^\star(x_n))$.

2.1 Lower Bound

Let $\mathcal{A}$ be a replicable PAC learning algorithm for the hypothesis set $\mathcal{H}$. In our lower bound proof, we let $h^\star$ be drawn uniformly from $\mathcal{H}$ and note that $h^\star$ is unknown to $\mathcal{A}$, except through the labels of training samples.

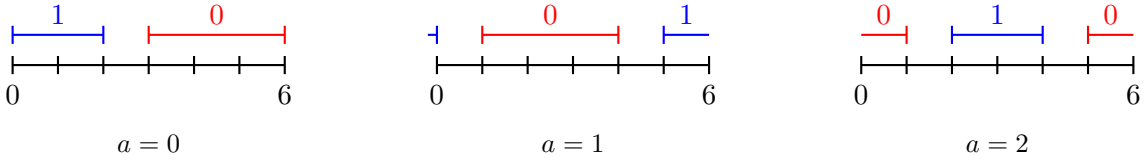

Figure 1: Example with $k = 7$ for hypothesis h_i for $i = (0, 5, 2)$. Each interval shows which values of b will make $h_i((a,b)) = 1$. For instance, if $a = 1$ then $h_i((a,b)) = 1$ for $b \in \{0, 5, 6\}$.

If $(S, h^\star(S))$ denotes a training set labeled by $h^\star$, then we use $\mathcal{A}(S, h^\star(S); r)$ to denote the output of $\mathcal{A}$ on the random string r and training set $(S, h^\star(S))$. Let ε be the error parameter of $\mathcal{A}$, let δ be the failure probability, and let ρ be the replicability parameter. That is, for any target function $h^\star \in \mathcal{H}$, it holds with probability at least $1-\delta$ over the random choice of a training set $S \sim \mathcal{D}^n$ and r that

$$\mathrm{er}_{\mathcal{D}}(\mathcal{A}(S, h^\star(S); r)) := \Pr_{x \sim \mathcal{D}}[\mathcal{A}(S, h^\star(S); r)(x) \neq h^\star(x)] \leq \varepsilon.$$

Our goal is to lower bound n as a function of ε and $|\mathcal{H}| = k^d$, assuming δ, ρ are sufficiently small constants.

Similarly to previous lower bounds in replicable learning, see e.g., [3, 21], we start by fixing the internal randomness r to obtain a deterministic algorithm. Observe that for a fixed $r^\star$, we can define the *mode* of a hypothesis $h \in \mathcal{H}$, denoted $\mathrm{mode}(h)$, as the most frequently reported hypothesis $\hat{h} \in \{0,1\}^{\mathcal{X}}$ on a training set $(S, h(S))$ with $S \sim \mathcal{D}^n$. That is,

$$\mathrm{mode}(h) = \arg\max_{\hat{h} \in \{0,1\}^{\mathcal{X}}} \Pr_{S \sim \mathcal{D}^n}[\mathcal{A}(S, h(S); r^\star) = \hat{h}].$$

For δ, ρ sufficiently small constants, we can show via Markov's inequality and a union bound that there is a fixed value $r^\star$ of the randomness, such that for $(99/100)|\mathcal{H}|$ of the hypotheses h in $\mathcal{H}$, it must be the case that $\Pr_{S \sim \mathcal{D}^n}[\mathcal{A}(S, h(S); r^\star) = \mathrm{mode}(h)] \geq 99/100$, and at the same time, $\mathrm{er}_h(\mathrm{mode}(h)) \leq \varepsilon$. Here $\mathrm{er}_h(\mathrm{mode}(h))$ denotes $\Pr_{x \sim \mathcal{D}}[\mathrm{mode}(h)(x) \neq h(x)]$. We fix an arbitrary such $r^\star$ and let $\bar{\mathcal{H}}$ be the set of at least $(99/100)|\mathcal{H}|$ *good* hypotheses satisfying these two properties. The good hypotheses h thus often output their mode, and the mode has a high accuracy when data is labeled with h.

Our lower bound proof now consists of three main steps. In the first step, we show that if n is small, then most pairs of hypotheses h_1, h_2 that label $\mathcal{X}$ *nearly identically* (in a sense to be made precise later) have the same mode. In the second step, we show that this implies that there are many hypotheses in $\bar{\mathcal{H}}$ with the same mode. Finally, we argue that when many hypotheses in $\bar{\mathcal{H}}$ have the same mode, then there must be a pair $h_1, h_2 \in \bar{\mathcal{H}}$ with the same mode, but h_1 and h_2 are so different that it is not possible for $\mathrm{mode}(h_1) = \mathrm{mode}(h_2)$ to simultaneously satisfy $\mathrm{er}_{h_1}(\mathrm{mode}(h_1)) \leq \varepsilon$ and $\mathrm{er}_{h_2}(\mathrm{mode}(h_1)) \leq \varepsilon$. By definition of $\bar{\mathcal{H}}$, this gives a contradiction (to n being small). We now elaborate on the three steps.

Step 1. Recall that all hypotheses in $\mathcal{H}$ may be represented by a vector $(i_0, \ldots, i_{d-1}) \in \mathbb{Z}_k^d$, each giving the starting point of an interval of length $\lfloor k/2 \rfloor$ in $\mathbb{Z}_k$. We can thus identify each hypothesis h_u in $\mathcal{H}$ with a vector $u \in \mathbb{Z}_k^d$. Let $Z = \{-1, 0, 1\}^d$, and consider the graph G having one node for each hypothesis/vector $u \in \mathbb{Z}_k^d$ and an edge between nodes u and v if $u + z = v \pmod k$ for some $z \in Z$. Note that this is an undirected graph (since $e \in Z$ if and only if $-e \in Z$). The familiar reader may notice that G is in fact the Cayley graph corresponding to the group $\mathbb{Z}_k^d$ with generator Z. We wish to show that if we sample u uniformly in $\mathbb{Z}_k^d$ and let $v = u + z$ for uniform $z \in Z$, then the corresponding hypotheses h_u, h_v are both in $\bar{\mathcal{H}}$, *and* have the same mode with probability at least $96/100$.

To prove this, observe first that if a training set $S \sim \mathcal{D}^n$ contains no samples that are labeled differently by two hypotheses h_u and h_v, i.e. $h_u(S) = h_v(S)$, then since $\mathcal{A}$ is deterministic (we fixed the randomness $r^\star$), we must have that $\mathcal{A}(S, h_u(S); r^\star) = \mathcal{A}(S, h_v(S); r^\star)$. Now instead of letting $v = u + z$ for z uniform in Z, consider first sampling u uniformly, sampling $S \sim \mathcal{D}^n$ and then picking v among all hypotheses with $v - u \in Z$ satisfying $h_u(S) = h_v(S)$. Then we have that $\mathcal{A}(S, h_u(S); r^\star) = \mathcal{A}(S, h_v(S); r^\star)$. If we require $k \geq cn/d$ for large enough constant $c > 0$, then notice that for each

coordinate $a \in [d]$, if we let $S_a \subseteq S$ be the subset of samples in S of the form (a, b) for some $b \in \mathbb{Z}_k$, then with large constant probability (growing with c), we have $h_u(S_a) = h_{u+e_a}(S_a) = h_{u-e_a}(S_a)$ where e_a is the a^{th} standard unit vector and summation is over $\mathbb{Z}_k^d$. This follows simply from the fact that S_a must contain one of the points $\{(a, u_a), (a, u_a - 1), (a, u_a + \lfloor k/2 \rfloor), (a, u_a + \lfloor k/2 \rfloor - 1)\}$ for the label assignments to be distinct. By picking v in a careful randomized way among all hypotheses with $h_u(S) = h_v(S)$ and $v - u \in Z$, we can now ensure that $v - u$ is precisely uniform in Z. This argument crucially needs that $h_u(S_a) = h_{u+e_a}(S_a) = h_{u-e_a}(S_a)$ for each coordinate with large constant probability. We thus obtain a distribution over triples (u, v, S) so that u is uniform in $\mathbb{Z}_k^d$, $v - u$ is uniform in Z, $S \sim \mathcal{D}^n$ and $\mathcal{A}(S, h_u(S); r^\star) = \mathcal{A}(S, h_v(S); r^\star)$.

Next, observe that u is chosen independently of S. Thus with probability at least $98/100$, we have $\mathcal{A}(S, h_u(S); r^\star) = \text{mode}(h_u)$ and $h_u \in \bar{\mathcal{H}}$. A careful argument also shows that if we consider the distribution of the pair (S, v), then v is uniform and independent of S. It is only through the variable u that dependencies between S and v are introduced. Another union bound gives us that with probability at least $96/100$, we have that $h_u, h_v \in \bar{\mathcal{H}}$ and $\text{mode}(h_u) = \mathcal{A}(S, h_u(S); r^\star) = \mathcal{A}(S, h_v(S); r^\star) = \text{mode}(h_v)$.

Step 2. We now want to leverage the result of Step 1. to show that there are many pairs $h_u, h_v \in \bar{\mathcal{H}}$ that are assigned the same mode. For this, consider partitioning the nodes u of G based on the modes $\text{mode}(h_u)$. That is, for every $f \in \{0,1\}^{\mathcal{X}}$ that appears as a mode, we let $C_f \subseteq \bar{\mathcal{H}}$ denote the subset of nodes $u \in \bar{\mathcal{H}}$ so that $\text{mode}(h_u) = f$. We will show that there must be a large C_f. For this argument, notice that Step 1. implies that

$$\begin{aligned} 96/100 &\leq \sum_f \Pr[u \in C_f \wedge u + z \in C_f] \\ &= \sum_f \Pr[u + z \in C_f \mid u \in C_f] \Pr[u \in C_f]. \end{aligned}$$

This implies the existence of an $f^\star$ so that $\Pr[u + z \in C_{f^\star} \mid u \in C_{f^\star}] \geq 96/100$. We claim that this property implies that $C_{f^\star}$ is large. To see this, notice that conditioning on $u \in C_{f^\star}$ simply means that u is uniform in $C_{f^\star}$. Since z is uniform in Z and independent of u, this further implies that $(u, u + z)$ is uniformly random among all (directed) edges incident to the nodes in $C_{f^\star}$. We thus have that $C_{f^\star}$ is a set of nodes with *small expansion* in the Cayley graph G. That is, at most a $4/100 = 1/25$ fraction of the directed edges $\{(u, v) : u \in C_{f^\star}, (u, v) \in E(G)\}$ has $v \notin C_{f^\star}$.

We thus proceed to show that every small set of nodes T in G expands a lot, i.e., many of the incident edges have an end point not in T. For this, we consider the adjacency matrix A of G and let $\mathbf{1}_T \in \{0,1\}^{k^d}$ be an indicator vector for the nodes in T, taking the value 1 in coordinates corresponding to nodes $u \in T$ and 0 elsewhere. If only a $1/25$ fraction of the edges incident to nodes in T leave T, then we must have $\langle A\mathbf{1}_T, \mathbf{1}_T \rangle \geq (24/25)|T||Z|$. This follows since $\langle A\mathbf{1}_T, \mathbf{1}_T \rangle$ counts precisely the number of directed edges (u, v) with both $u, v \in T$. Furthermore, there is a total of $|T||Z|$ directed edges incident to T since all nodes have degree $|Z|$.

The eigenvectors and eigenvalues of the adjacency matrix of a Cayley graph are well understood. In particular, A has an eigenvector corresponding to each vector $u \in \mathbb{Z}_k^d$. Let us denote this eigenvector by χ_u and the corresponding (real valued) eigenvalue by λ_u. We then have that

$$\langle A\mathbf{1}_T, \mathbf{1}_T \rangle = \sum_{u \in \mathbb{Z}_k^d} \lambda_u |\langle \chi_u, \mathbf{1}_T \rangle|^2.$$

One can show that all entries of χ_u are bounded by $k^{-d/2}$ in magnitude, yielding $|\langle \chi_u, \mathbf{1}_T \rangle|^2 \leq |T|^2 k^{-d}$. Furthermore, the eigenvectors form an orthonormal basis and thus $\sum_u |\langle \chi_u, \mathbf{1}_T \rangle|^2 =$

$\|\mathbf{1}_T\|^2 = |T|$. Finally, since G is a $|Z|$-regular graph, the largest eigenvalue of A is $\lambda_{\mathbf{0}} = |Z|$. To exploit these properties, let $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{k^d} = |Z|$ be the eigenvalues λ_v in sorted order and χ_i the eigenvector corresponding to μ_i. Combining our observations, we get that for any set T, we can upper bound $\langle A\mathbf{1}_T, \mathbf{1}_T\rangle$ as

$$\begin{aligned}\langle A\mathbf{1}_T, \mathbf{1}_T\rangle &\leq \mu_{k^d} \cdot \sum_{i=k^d-k^d/(2|T|)+1}^{k^d} |\langle \chi_i, \mathbf{1}_T\rangle|^2 + \mu_{k^d-k^d/(2|T|)} \cdot \sum_{i=1}^{k^d-k^d/(2|T|)} |\langle \chi_i, \mathbf{1}_T\rangle|^2 \\ &\leq |Z| \cdot \frac{k^d}{2|T|} \cdot \frac{|T|^2}{k^d} + \mu_{k^d-k^d/(2|T|)} \cdot \left(|T| - \frac{k^d}{2|T|} \cdot \frac{|T|^2}{k^d}\right) \\ &= \Big(|Z|/2 + \mu_{k^d-k^d/(2|T|)}/2\Big)|T|.\end{aligned}$$

Now if T satisfies $\langle A\mathbf{1}_T, \mathbf{1}_T\rangle \geq (24/25)|T||Z|$, we conclude that we must have $\mu_{k^d-k^d/(2|T|)} \geq (46/50)|Z|$. We thus proceed to bound the $k^d/(2|T|)$'th largest eigenvalue of the adjacency matrix A. Here we argue that for each $u \in \mathbb{Z}_k^d$, we have

$$\lambda_u = |Z| - 2\sum_{z\in Z} \sin^2(\pi\langle u, z\rangle/k).$$

To get a feel for this expression, consider a fixed non-zero $z \in Z$ and let u be drawn uniformly from $\mathbb{Z}_k^d$. Then the inner product $\langle u, z\rangle$ is uniform in $\mathbb{Z}_k$. In particular, with probability close to $1/2$, the inner product lies in $\{\lfloor k/4\rfloor + 1, \ldots, \lfloor k/4\rfloor + \lfloor k/2\rfloor\}$, yielding $\sin^2(\pi\langle u, z\rangle/k) \geq \sin^2(\pi/4) = 1/2$. So at least in expectation over a uniform u, we have $\mathbb{E}[\lambda_u] \leq |Z| - |Z|/2 \ll (46/50)|Z|$. To show that $|T|$ must be large, we thus need to argue that almost all λ_u are close to this expectation. We do this via a probabilistic argument. In particular, we notice that for λ_u to satisfy $\lambda_u \geq (46/50)|Z|$, there can be no more than $(8/50)|Z|$ values of z for which $\langle u, z\rangle \in \{\lfloor k/4\rfloor + 1, \ldots, \lfloor k/4\rfloor + \lfloor k/2\rfloor\}$.

We now let u be chosen uniformly from $\mathbb{Z}_k^d$ and define indicator random variables X_z taking the value 1 if $\langle u, z\rangle \notin \{\lfloor k/4\rfloor + 1, \ldots, \lfloor k/4\rfloor + \lfloor k/2\rfloor\}$ and 0 otherwise. If p denotes the probability that $\sum_{z\in Z} X_z \geq (42/50)|Z|$, then for any $j > pk^d$, we have $\mu_{k^d-j} < (46/50)|Z|$. If we can give a good upper bound on p, we can now conclude that $k^d/(2|T|) \leq pk^d \Rightarrow |T| \geq p^{-1}/2$.

To bound the probability that $\sum_{z\in Z} X_z \geq (42/50)|Z|$, we consider the moment

$$\mathbb{E}\left[\left(\sum_{z\in Z} X_z - \mathbb{E}[X_z]\right)^r\right] = \sum_{Y\in Z^r} \mathbb{E}\left[\prod_{y_i\in Y} (X_{y_i} - \mathbb{E}[X_{y_i}])\right],$$

for an even $r \geq 2$. Here we notice that for any $Y \in Z^r$, if just one of the vectors $y_i \in Y$ is linearly independent of the remaining as vectors over $\mathbb{Z}_k^d$ (remember, k is prime), then the variable X_{y_i} is independent, and the whole monomial $\mathbb{E}\Big[\prod_{y_i\in Y}(X_{y_i} - \mathbb{E}[X_z])\Big]$ is zero. If β denotes the fraction of tuples $Y \in Z^r$ where *every* y_i in Y can be written as a linear combination of the remaining y_j, then we may bound

$$\mathbb{E}\left[\left(\sum_{z\in Z} X_z - \mathbb{E}[X_z]\right)^r\right] \leq \beta|Z|^r.$$

By Markov's inequality, we then have

$$\begin{aligned} p = \Pr\left[\sum_{z\in Z} X_z \geq (42/50)|Z|\right] &\leq \Pr\left[\left|\sum_{z\in Z}(X_z - 1/2)\right| \geq (17/50)|Z|\right] \\ &= \Pr\left[\left(\sum_{z\in Z}(X_z - 1/2)\right)^r \geq (17/50)^r|Z|^r\right] \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{z\in Z}(X_z - 1/2)\right)^r\right]}{(17/50)^r|Z|^r} \\ &\leq \beta(50/17)^r. \end{aligned}$$

That is, we get $|T| \geq p^{-1}/2 \geq \beta^{-1}(17/50)^r/2$. We thus seek a small upper bound on β. Here we again use a probabilistic argument. Consider drawing a set $Y \in Z^r$ uniformly at random, i.e. each y_i in Y is drawn independently and uniformly from $Z = \{-1, 0, 1\}^d$. Then β is precisely the probability that every y_i in Y can be written as a linear combination of the remaining y_j. Giving a tight upper bound on this probability turns out to be rather involved. In particular, the fact that every y_i can be written as a linear combination of the remaining does not imply that $\dim(\text{span}(Y))$ is small. It could be that for instance $\mathbf{0} = \sum_i y_i$. The fact that the rank is at most $r-1$ only gives something like $\beta \leq c^{-d}$ for a constant $c \leq 3$. This turns out to be insufficient for our lower bound (as we shall see later). Instead, we first show that $\dim(\text{span}(Y)) \geq r - \log d$, except for a $d^{-d/2}$ fraction of $Y \in Z^r$. Assuming $\dim(\text{span}(Y)) \geq r - \log d$, we show that this implies that there must be a linear combination $y_i = \sum_{j\in J} \alpha_j y_j$ with $i \notin J$, $\alpha_j \neq 0$ and $|J| \geq r/\log d$. We then argue that such linear combinations involving many y_j's are very unlikely when k is large enough. In particular, the fact that $\mathbf{0} = y_i - \sum_{j\in J} \alpha_j y_j$ implies that in each coordinate $a \in [d]$, we have that $0 = y_i(a) - \sum_{j\in J} \alpha_j y_j(a)$. By definition of Z and drawing Y uniformly, we get that all the $y_i(a)$ and $y_j(a)$ are independent and uniform in $\{-1, 0, 1\}$. Using ideas by Golovnev et al. [26] in a paper on data structure lower bounds in the group model, we can now use a Littlewood-Offord type anti-concentration result to show that $0 = y_i(a) - \sum_{j\in J} \alpha_j y_j(a)$ only with probability roughly $k^{-1} + \sqrt{1/(|J|+1)}$. To get a feel for this claim, observe that it morally says that when k is large enough, a sum $\sum_{j\in J} y_j\alpha_j$ with each y_j uniform in $\{-1, 0, 1\}$ and each α_j non-zero, takes on any particular value mod k with probability no larger than the probability that the random sum $\sum_{j\in J} y_j$ over the integers takes a particular value. The random sum $\sum_{j\in J} y_j$ has standard deviation $\Theta(\sqrt{|J|})$ and is near-uniform within one standard deviation of 0, thus taking on any particular value with probability at most $O(1/\sqrt{|J|})$. Ignoring the dependence on k for simplicity (recall we needed $k \geq cn/d$), we have $k^{-1} + \sqrt{1/(|J|+1)} \approx \sqrt{\log(d)/r}$. Using independence across the d choices for a finally bounds β as roughly $(\log(d)/r)^{d/2}$. Picking $r \approx d$ gives $|T| \geq (cd/\log(d))^{d/2}$ for some constant $c > 0$. Note that this is much stronger than the first bound on β of c^{-d} that one obtains simply from $\dim(\text{span}(Y)) \leq r - 1$.

Step 3. From Step 2. we have concluded that there is a set $C_{f^\star}$ with $|C_{f^\star}| \geq (cd/\log(d))^{d/2}$. Recall that $C_{f^\star}$ is defined as all nodes u so that $h_u \in \bar{\mathcal{H}}$ and $\text{mode}(h_u) = f^\star$. These nodes u thus correspond to hypotheses h_u with the same mode, and where by definition of $\bar{\mathcal{H}}$, this mode $f^\star$ has error at most ε when $\mathcal{X}$ is labeled by h_u. Since $\mathcal{D}$ is uniform over $\mathcal{X}$ and $|\mathcal{X}| = dk$, this implies that any two hypotheses h_u, h_v with $u, v \in C_{f^\star}$ can disagree on the label of at most $2\varepsilon dk$ points $(a, b) \in \mathcal{X}$. When viewed as vectors $u, v \in \mathbb{Z}_k^d$, this basically corresponds to $\|u - v\|_1 \leq 2\varepsilon kd$ (if we ignore wrap-around for simplicity). But an ℓ_1 ball of radius $2\varepsilon kd$ has at most $\sum_{i=0}^{2\varepsilon kd} 2^d\binom{d+i-1}{d-1}$ points with integer coordinates inside it. Let us for simplicity assume $\varepsilon kd > 2d$, then this number of points is

roughly $(c'\varepsilon k)^d$ for some constant $c' > 0$. We therefore must have $(cd/\log(d))^{d/2} \leq |C_{f^\star}| \leq (c'\varepsilon k)^d$. Recall again that we needed to set $k \geq c''n/d$ for a constant $c'' > 0$. Inserting $k = c''n/d$ and taking d'th root finally gives

$$k = \Omega\Big(\varepsilon^{-1}\sqrt{d/\log(d)}\Big) \Rightarrow n = \widetilde{\Omega}(\varepsilon^{-1}d^{3/2}).$$

If we instead state the lower bound as a function of $|\mathcal{H}| = k^d$ and use $k = c''n/d$, we have $d = \log|\mathcal{H}|/\log k = \widetilde{\Omega}(\log|\mathcal{H}|)$ and the lower bound is $n = \widetilde{\Omega}(\varepsilon^{-1}(\log|\mathcal{H}|)^{3/2})$, where $\widetilde{\Omega}$ hides logarithmic factors in $\log|\mathcal{H}|$ and $1/\varepsilon$.

2.2 Upper Bound

We also present an upper bound for the same input domain $\mathcal{X} = [d] \times \mathbb{Z}_k$, hypothesis set $\mathcal{H}$ and distribution $\mathcal{D}$ considered in the lower bound. Recall here that $\mathcal{H}$ contains a hypothesis h_i for every d-tuple $(i_0, \ldots, i_{d-1}) \in \mathbb{Z}_k^d$ that for a point $(a, b) \in \mathcal{X}$ returns 1 if $i_a \leq b < i_a + \lfloor k/2 \rfloor$ or if $b < i_a + \lfloor k/2 \rfloor < i_a$. Each h_i thus corresponds to outputting a 1 on d intervals of length $\lfloor k/2 \rfloor$ with wrap-around. The distribution $\mathcal{D}$ is simply the uniform distribution on $\mathcal{X}$. We let the unknown target function $h_{i^\star} \in \mathcal{H}$ be arbitrary.

Our replicable algorithm $\mathcal{A}$ is quite natural. From a sample $S \sim \mathcal{D}^n$, we start by estimating each i_a. Since we expect to see n/d samples of the form (a, b) for each $a \in [d]$, we can estimate each $i_a^\star$ to within additive $\widetilde{O}(kd/n)$. Call these estimates b_a^S. Since $\mathcal{D}$ is uniform, if we output the hypothesis h_{b^S} with $b^S = (b_0^S, \ldots, b_{d-1}^S)$, then $\mathrm{er}_{\mathcal{D}}(h_{b^S})$ is essentially equal to $\|b^S - i^\star\|_1 d^{-1}$, except that the wrap-around may reduce the error further. To be slightly more formal, define for any $u, v \in \mathbb{Z}_k^d$ the metric $\nu : \mathbb{Z}_k^d \times \mathbb{Z}_k^d \to \{0, \ldots, \lfloor k/2 \rfloor\}$ defined by $\nu(u, v) = \sum_{i=0}^{d-1} \min(u_i - v_i, v_i - u_i)$. Note here that we are working over the finite field $\mathbb{Z}_k$ and thus we are implicitly taking $\bmod k$ in $u_i - v_i$ and $v_i - u_i$. We can thus interpret $\nu(u, v)$ as the wrap-around ℓ_1 distance between u and v. For convenience, we will also use $\nu(u_i, v_i) = \min(u_i - v_i, v_i - u_i)$ to denote the one-dimensional version of ν.

With this notation, we see that $\mathrm{er}_{\mathcal{D}}(h_i) = 2\nu(i, i^\star)/|\mathcal{X}| = 2\nu(u, v)/(kd)$. It follows that if $n = \widetilde{\Omega}(d\varepsilon^{-1})$ then $\nu(b^S, i^\star) = \widetilde{O}(kd^2/n) \leq \varepsilon kd/4$ and thus $\mathrm{er}_{\mathcal{D}}(h_{b^S}) \leq \varepsilon/2$ as desired. Unfortunately it is not replicable to simply output h_{b^S}.

Instead we need to randomly round b^S using shared randomness, such that for another i.i.d. sample $S' \sim \mathcal{D}^n$, the rounding of b^S and $b^{S'}$ are equal with probability at least $1 - \rho$. Our goal is to round b^S and $b^{S'}$ to a hypothesis h_i with $\nu(i, b^S), \nu(i, b^{S'}) \leq \varepsilon kd/4$. The triangle inequality would then give $\nu(i, i^\star) \leq \nu(i, b^S) + \nu(b^S, i^\star) \leq \varepsilon kd/4 + \varepsilon kd/4 = \varepsilon kd/2$, implying $\mathrm{er}_{\mathcal{D}}(h_i) \leq \varepsilon$.

For the randomized rounding, we use shared randomness to shuffle all hypotheses in $\mathcal{H}$ uniformly at random. From sample S, $\mathcal{A}$ now outputs the *first* hypothesis h_i in the shuffled order which satisfies $\nu(i, b^S) \leq \varepsilon kd/4$. Correctness is thus guaranteed from the arguments above. The tricky part is to show that for two samples S and S', the first such hypothesis h_i is the same with probability at least $1 - \rho$.

For this analysis, let h_i be the first hypothesis satisfying $\nu(i, b^S) \leq \varepsilon kd/4$. We will argue that with probability at least $1 - \rho/2$, we also have $\nu(i, b^{S'}) \leq \varepsilon kd/4$. This implies that if $h_{i'}$ is the first hypothesis satisfying $\nu(i, b^{S'}) \leq \varepsilon kd/4$, then $h_{i'}$ is no later than h_i in the random ordering of $\mathcal{H}$. A symmetric argument and a union bound shows that at the same time, h_i is no later in the random ordering than $h_{i'}$ implying $h_i = h_{i'}$.

Now observe that if we fix an S and S' and condition on the event that h_i is the first hypothesis in the shuffled $\mathcal{H}$ with $\nu(i, b^S) \leq \varepsilon kd/4$, then the distribution of h_i is uniform random among all hypotheses with $\nu(i, b^S) \leq \varepsilon kd/4$. To analyze $\nu(i, b^{S'})$, observe that in every coordinate $a \in [d]$, we

have with probability $1/2$ (over i) that

$$\nu(i_a, b_a^{S'}) \leq \max\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\} - \min\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\}.$$

and with probability $1/2$, we have

$$\nu(i_a, b_a^{S'}) \leq \max\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\} + \min\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\}.$$

The two cases corresponds to whether i_a is in the direction towards $b_a^{S'}$ from b_a^S or in the opposite direction. If we have $n = \widetilde{O}(\beta^{-1}d)$ for a parameter $\beta > 0$ to be fixed, then we can ensure that $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$ for each coordinate a. Since we sample h_i satisfying $\nu(i, b^S) \leq \varepsilon kd/4$, we would expect most coordinates a to have $\nu(i_a, b_a^S) \approx \varepsilon k/4$. If we let $\beta \ll \varepsilon$, then for most a, the min above is $\nu(b_a^S, b_a^{S'})$. Let us for simplicity assume that this is always the minimum. Then there are signs $\sigma_a \in \{-1, 1\}$ such that

$$\begin{aligned}
\nu(i, b^{S'}) &\leq \sum_{a=0}^{d-1} \max\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\} + \sigma_a \min\{\nu(i_a, b_a^S), \nu(b_a^S, b_a^{S'})\} \\
&= \sum_{a=0}^{d-1} \nu(i_a, b_a^S) + \sigma_a \nu(b_a^S, b_a^{S'}) \\
&= \nu(i, b^S) + \sum_{a=0}^{d-1} \sigma_a \nu(b_a^S, b_a^{S'})
\end{aligned}$$

Noting that the signs σ_a are uniform and independent, and that $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$, we get from Hoeffding's inequality that the contribution from $\sum_{a=0}^{d-1} \sigma_a \nu(b_a^S, b_a^{S'})$ is bounded by $\widetilde{O}(\beta k\sqrt{d})$ with high probability (like $1 - \rho/4$).

What remains is thus to show that $\nu(i, b^S)$ is somewhat smaller than $\varepsilon kd/4$ with probability $1 - \rho/4$. This amounts to a counting/probabilistic argument where we show that a uniform random i satisfying $\nu(i, b^S) \leq \varepsilon kd/4$ has $\nu(i, b^S) \leq \varepsilon kd/4 - \Omega(\varepsilon k\rho)$ with probability $1 - \rho/4$. In some sense, this is showing that the ℓ_1 "wrap-around" ball in $\mathbb{Z}_k^d$ has most points somewhat in the interior of the ball.

We now have that with probability $1 - \rho/2$, it holds that $\nu(i, b^{S'}) \leq \varepsilon kd/4 + \widetilde{O}(\beta k\sqrt{d}) - \Omega(\varepsilon k\rho)$. Picking $\beta = \widetilde{O}(\varepsilon\rho/\sqrt{d})$ gives $\nu(i, b^{S'}) \leq \varepsilon kd/4$ and thus h_i is also a valid output on S' and we conclude $h_{i'}$ is no later than h_i in the random order.

For the above, we needed $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$. Since $\nu(b_a^S, i_a) = \widetilde{O}(dk/n)$ we also have $\nu(b_a^S, b_a^{S'}) = \widetilde{O}(dk/n)$. It is thus sufficient to pick an n satisfying $n = \widetilde{O}(\beta^{-1}d) = \widetilde{O}(\varepsilon^{-1}\rho^{-1}d^{3/2})$. Since $|\mathcal{H}| = k^d$, we have $d^{3/2} = O((\log|\mathcal{H}|)^{3/2})$ and we have the claimed upper bound.

In our full proof, we also have to deal with the fact that not all coordinates a have $\nu(i_a, b_a^S) \geq \nu(b_a^S, b_a^{S'})$. This complicates the analysis somewhat, but the overall intuition and strategy remains the same.

In summary, the two main observations are that, 1., the uniform h_i among all hypotheses with $\nu(i, b^S) \leq \varepsilon kd/4$ actually has $\nu(i, b^S)$ somewhat smaller than $\varepsilon kd/4$ with good probability, and 2., for every coordinate a, there is a probability $1/2$ that i_a is towards $b_a^{S'}$ from b_a^S and thus $\nu(i_a, b_a^{S'})$ is actually less than $\nu(i_a, b_a^S)$. This probability of $1/2$ is exactly what yields the $\sqrt{d}$ behavior (a sum of d random signs has a standard deviation of $\sqrt{d}$).

3 Proof of the Lower Bound

In this section, we will prove our sample complexity lower bound for replicable realizable PAC learning. For convenience, we restate the theorem here.

Theorem 1.2 (Replicable Learning Lower Bound)**.** *For any integer $d \geq 10^{11}$, and positive reals $\varepsilon, \delta, \rho \leq 10^{-4}$, there exists a domain $\mathcal{X}$, a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ with VC-dimension d, such that for any algorithm $\mathcal{A}$ there is a distribution $\mathcal{D}$ over $\mathcal{X}$ for which $\mathcal{A}$ needs at least*

$$n = \widetilde{\Omega}\left(\frac{(\log|\mathcal{H}|)^{3/2}}{\varepsilon}\right)$$

labeled samples from $\mathcal{D}$ in order to be a ρ-replicable PAC learner for $\mathcal{H}$ with error ε and failure probability δ. Here $\widetilde{\Omega}$ hides logarithmic factors in $\log|\mathcal{H}|$ and $1/\varepsilon$.

The instance we will prove is hard is the one described in Section 2. So, let $\mathcal{X} = [d] \times \mathbb{Z}_k$ for prime k to be determined in a moment, and let $\mathcal{H}$ contain the hypotheses h_i for every d-tuple $i = (i_0, \ldots, i_{d-1}) \in \mathbb{Z}_k^d$ given by

$$h_i((a,b)) = \begin{cases} 1, & \text{if } i_a \leq b < i_a + \lfloor k/2 \rfloor \text{ or } b < i_a + \lfloor k/2 \rfloor < i_a \\ 0, & \text{otherwise} \end{cases}$$

Lastly, we will pick $\mathcal{D}$ to be the uniform distribution over $\mathcal{X}$. We will pick k to be a prime satisfying

$$\max\left\{\frac{2}{\varepsilon}, \frac{\sqrt{d}}{\log d}, \frac{4n}{\log(81/80)d}\right\} \leq k \leq 2 \cdot \max\left\{\frac{2}{\varepsilon}, \frac{\sqrt{d}}{\log d}, \frac{4n}{\log(81/80)d}\right\}.$$

Remark that such a prime always exists due to the Bertrand-Chebyshev theorem. To prove the lower bound, we assume that there exists a randomized algorithm $\mathcal{A}$ which is a replicable PAC learner for $\mathcal{H}$ with sample complexity n. The goal is then to show that n must be large. By replicability and correctness of $\mathcal{A}$, we have that for $S, S' \sim \mathcal{D}^m$ and for any $h \in \mathcal{H}$ it holds that

$$\Pr_{S,S',r}[\mathcal{A}(S, h(S); r) = \mathcal{A}(S', h(S'); r)] \geq 1 - \rho, \tag{1}$$

$$\Pr_{S,r}[\mathrm{er}_h(\mathcal{A}(S, h(S); r) \leq \varepsilon] \geq 1 - \delta. \tag{2}$$

Specifically, (1) and (2) hold if we pick h uniformly at random in $\mathcal{H}$. Now, we would like to de-randomize $\mathcal{A}$. Thus, we apply Markov's inequality and a union bound to get that with probability at least $1/3$ we pick a randomness $r^\star$ such that both

$$\Pr_{S,S',h}[\mathcal{A}(S, h(S); r^\star) = \mathcal{A}(S', h(S'); r^\star)] \geq 1 - 3\rho, \tag{3}$$

$$\Pr_{S,h}[\mathrm{er}_h(\mathcal{A}(S, h(S); r^\star) \leq \varepsilon] \geq 1 - 3\delta. \tag{4}$$

Therefore, consider a deterministic version of $\mathcal{A}$, which uses a fixed randomness satisfying properties (3) and (4). Remark that the sample complexity of the deterministic $\mathcal{A}$ will lower bound the sample complexity of the randomized $\mathcal{A}$. For the rest of the proof, we will therefore consider this deterministic version of $\mathcal{A}$ and lower bound its sample complexity. For ease of notation, we will not write the fixed randomness used by $\mathcal{A}$ explicitly in the rest of the proof. After de-randomizing $\mathcal{A}$, we can now define the mode of a hypothesis with respect to $\mathcal{A}$.

Definition 3.1 (Mode)**.** Let $h \in \mathcal{H}$. Then we define the *mode* of h as

$$\mathrm{mode}(h) = \arg\max_{f \in \{0,1\}^{\mathcal{X}}}\left\{\Pr_{S \sim \mathcal{D}^m}[\mathcal{A}(S, h(S)) = f]\right\}.$$

In words, the mode of h is just the most likely output of $\mathcal{A}$ when the samples are labeled by h. To make the mode uniquely determined, we break ties by choosing the lexicographically smallest f in the arg max.

Now, we arrange all the hypotheses in a graph $G = (V, E)$. Here, the vertices are the vectors associated to the hypotheses in the graph. That is $V = \mathbb{Z}_k^d$, so we have one node for each hypothesis. Now, let $Z = \{-1, 0, 1\}^d$. Then, for each $u, v \in \mathbb{Z}_k^d$, E contains the edge (u, v) if and only if $u - v \in Z$. We can therefore think of G as a d-dimensional grid with diagonal edges and wrap-around. This can also be described as the Cayley graph on the group $\mathbb{Z}_k^d$ with generating set Z.

Now, for each possible output f of $\mathcal{A}$ we create a subset of nodes u with $\mathrm{mode}(h_u) = f$ for which the error f with respect to h_u is less than ε. Formally, for each $f \in \{0, 1\}^{\mathcal{X}}$ we define such subset as

$$C_f = \{u \in \mathbb{Z}_k^d \mid \mathrm{mode}(h_u) = f,\ \mathrm{er}_{h_u}(f) \leq \varepsilon\}.$$

We then have the following lemma about these subsets.

Lemma 3.2. *There exists a function $f^\star \in \{0, 1\}^{\mathcal{X}}$ such that*

$$\#\{(u, v) \in E \mid u, v \in C_{f^\star}\} \geq \frac{24}{25}|C_{f^\star}||Z|.$$

Remark that $|C_f||Z|$ is the total number of edges incident to nodes in C_f. Consequently, Lemma 3.2 implies that for at least one $f^\star$, the set $C_{f^\star}$ has low expansion. We defer the proof of this lemma to Section 3.1. For now, assume such a subset exists and consider the following theorem.

Theorem 3.3. *Let $T \subseteq \mathbb{Z}_k^d$ be a subset of the vertices. If $\#\{(u, v) \in E \mid u, v \in T\} \geq (24/25)|T||Z|$ then $|T| \geq \left(\frac{\sqrt{d}}{450 \log d}\right)^d$.*

We will defer the proof of this theorem to Section 3.2. We can now combine this with Lemma 3.2 to lower bound the size of $C_{f^\star}$ as $|C_{f^\star}| \geq \left(\frac{\sqrt{d}}{450 \log d}\right)^d$. Now, to upper bound the size of $C_{f^\star}$, we have the following lemma.

Lemma 3.4. *For any $f \in \{0, 1\}^{\mathcal{X}}$ we have $|C_f| \leq (6\varepsilon k)^d$.*

The proof of this lemma is deferred to Section 3.3. We can then combine the upper and lower bound on $|C_{f^\star}|$ to get

$$\left(\frac{\sqrt{d}}{450 \log d}\right)^d \leq |C_{f^\star}| \leq (6\varepsilon k)^d \tag{5}$$

Remember that we chose k such that $k \leq \max\left\{\frac{4}{\varepsilon}, \frac{2\sqrt{d}}{\log d}, \frac{8n}{\log(81/80)d}\right\}$. We will now see that the first two terms in the max cannot be the maximum. Indeed, if the first term was the maximum, then inequality (5) would imply that

$$\frac{\sqrt{d}}{\log d} < 10800$$

which contradicts the assumption that $d \geq 10^{11}$. If the second term was the maximum, then inequality (5) imply that

$$\frac{1}{5400} \leq \varepsilon$$

which contradicts the fact that $\varepsilon \leq 10^{-4}$. Thus, it must be the case that $k \leq \frac{8n}{\log(81/80)d}$. Substituting this value of k into inequality (5) gives us

$$n \geq 10^{-7} \frac{d^{3/2}}{\varepsilon \log d}.$$

We will now write the lower bound in terms of $|\mathcal{H}|$. Using that $d = \log|\mathcal{H}|/\log k \geq \log|H|/(7\log n)$, and $d \leq \log|\mathcal{H}|$, we have

$$n \geq 10^{-9} \frac{(\log|\mathcal{H}|)^{3/2}}{\varepsilon \log\log|\mathcal{H}| \log n}.$$

We still have a dependence of n on the right side of the inequality. However, a standard trick using a small proof by contradiction (see Lemma A.2), one can show that the above implies

$$n \geq 10^{-9} \frac{(\log|\mathcal{H}|)^{3/2}}{\varepsilon(\log\log|\mathcal{H}|)^2 \log(\log(|\mathcal{H}|)/\varepsilon)}.$$

This is the claim of the lower bound, since we only need to remove logarithmic factors in $\log|\mathcal{H}|$ and $1/\varepsilon$. What remains is therefore to show Lemma 3.2, Theorem 3.3, and Lemma 3.4. Each of these will be proved in the following sections.

3.1 Random Step Approach

In this section, we will prove Lemma 3.2. For this, we will need the following lemma about random steps in the graph G.

Lemma 3.5 (Random step)**.** *Let h_u be a uniformly random hypothesis in $\mathcal{H}$, and let h_v be a hypothesis chosen by starting at h_u and taking a uniformly random step in the graph G (that is, $v = u + z$ for z uniform on Z). Then, we have*

$$\Pr_{h_u,h_v}[\mathrm{mode}(h_u) = \mathrm{mode}(h_v)] \geq 1 - 6\rho.$$

Furthermore, with probability $1 - 6(\delta + \rho)$ over the choice of h_u, h_v, it holds that

$$\mathrm{er}_{h_u}(\mathrm{mode}(h_u)) \leq \varepsilon,$$
$$\mathrm{er}_{h_v}(\mathrm{mode}(h_v)) \leq \varepsilon.$$

First, we will show how Lemma 3.5 implies Lemma 3.2. Then we will prove Lemma 3.5. For convenience, we also restate Lemma 3.2 here.

Lemma 3.2. *There exists a function $f^\star \in \{0,1\}^{\mathcal{X}}$ such that*

$$\#\{(u,v) \in E \mid u, v \in C_{f^\star}\} \geq \frac{24}{25}|C_{f^\star}||Z|.$$

Proof of Lemma 3.2. Consider h_u, h_v chosen as in Lemma 3.5. Using that $\rho, \delta \leq \frac{1}{450}$, we have from Lemma 3.5 with probability at least $24/25$ that $\mathrm{mode}(h_u) = \mathrm{mode}(h_v)$, $\mathrm{er}_{h_u}(\mathrm{mode}(h_u)) \leq \varepsilon$, and $\mathrm{er}_{h_v}(\mathrm{mode}(h_v)) \leq \varepsilon$. Now, these three statements together imply that there exists an $f \in \{0,1\}^{\mathcal{X}}$ such that $u, v \in C_f$. Also, define $f^* = \arg\max_f \Pr[v \in C_f \mid u \in C_f]$. Since all C_f are disjoint, we can write this out as

$$\frac{24}{25} \leq \sum_f \Pr[v \in C_f \wedge u \in C_f] = \sum_f \Pr[v \in C_f \mid u \in C_f] \Pr[u \in C_f] \leq \Pr[v \in C_{f^\star} \mid u \in C_{f^\star}] \quad (6)$$

Remark that since u is chosen uniform in $\mathbb{Z}_k^d$ independently of everything else, conditioning on $u \in C_{f^\star}$ just means that u is uniform in $C_{f^\star}$. Now, since z is uniform in Z, this implies that (u, v) is a uniformly random among all edges incident to nodes in $C_{f^\star}$. Since there are $|C_{f^\star}||Z|$ of such edges, we have

$$\Pr[v \in C_{f^\star} \mid u \in C_{f^\star}] = \frac{\#\{(u,v) \in E \mid u, v \in C_{f^\star}\}}{|C_{f^\star}||Z|}$$

Combining this with (6) gives us the desired result. □

What remains in this section is to prove Lemma 3.5. However, to do so, we will need two additional lemmas. The first one states that the mode is output with high probability. The second one tells us that there is a way to pick a uniformly random neighbor in G while being indistinguishable with respect to samples S.

Lemma 3.6 (Mode is likely)**.** *Let $S \sim \mathcal{D}^n$ be n i.i.d. samples from $\mathcal{D}$, and let $h \sim \mathcal{H}$ be a uniformly random hypothesis in $\mathcal{H}$ independent of S. Then,*

$$\Pr_{S,h}[\mathcal{A}(S, h(S)) = \text{mode}(h)] \geq 1 - 3\rho.$$

Proof. Consider another sample $S' \sim \mathcal{D}^n$, independent of both S and h. We can then use the replicability property of $\mathcal{A}$ (see inequality (3)) to get that

$$\begin{aligned}
1 - 3\rho &\leq \mathop{\mathbb{E}}_{h}\left[\Pr_{S,S'}[\mathcal{A}(S, h(S) = \mathcal{A}(S', h(S')]\right] = \mathop{\mathbb{E}}_{h}\left[\sum_{f \in \mathcal{H}} \Pr_{S}[\mathcal{A}(S, h(S) = f]^2\right] \\
&\leq \mathop{\mathbb{E}}_{h}\left[\Pr_{S}[\mathcal{A}(S, h(S)) = \text{mode}(h)] \sum_{f \in \mathcal{H}} \Pr_{S}[\mathcal{A}(S, h(S) = f]\right] \\
&= \mathop{\mathbb{E}}_{h}\left[\Pr_{S}[\mathcal{A}(S, h(S)) = \text{mode}(h)]\right] = \Pr_{S,h}[\mathcal{A}(S, h(S)) = \text{mode}(h)].
\end{aligned}$$

□

Lemma 3.7. *Let u be a uniformly random node in the graph G and let $S \sim \mathcal{D}^n$ be independent of u. Then, there exists a way to pick v a neighbor of u such that $h_u(S) = h_v(S)$ and v is a uniformly random neighbor of u. Furthermore, v and S are independent.*

Let us make some remarks on Lemma 3.7. In particular the dependencies among the random variables are delicate. If we consider the whole triple of random variables $(u, u-v, S)$ then they are *not* mutually independent. This should be clear from the fact that the requirement $h_u(S) = h_v(S)$ disallows some choices of v given u and S. Instead, the random variables $(u, u-v, S)$ are *pair-wise independent*. So any two of the variables behave as a uniform random and independently chosen pair. This pair-wise independence is critical for our proof.

We defer the proof of Lemma 3.7 to later in this section. For now, we show how these lemmas imply Lemma 3.5.

Proof of Lemma 3.5. The proof will be divided into two parts. Part 1 will show the first inequality of the lemma, and part 2 will show the second and third inequality.

Part 1. Let $S \sim \mathcal{D}^n$ be independent of u. Then, we invoke Lemma 3.7 to take a uniformly random step from u to a node v while making sure $h_u(S) = h_v(S)$. Now, since S is independent of u and v (although they are not all three independent of each other), and h_u, h_v are both uniformly random hypotheses in $\mathcal{H}$, we can use Lemma 3.6 on both u and v, to get that

$$\Pr_{S,h_u,h_v}[\mathcal{A}(S,h_u(S)) = \text{mode}(h_u)] = \Pr_{S,h_u}[\mathcal{A}(S,h_u(S)) = \text{mode}(h_u)] \geq 1 - 3\rho,$$
$$\Pr_{S,h_u,h_v}[\mathcal{A}(S,h_v(S)) = \text{mode}(h_v)] = \Pr[\mathcal{A}(S,h_v(S)) = \text{mode}(h_v)] \geq 1 - 3\rho.$$

Now, using the fact that $h_u(S) = h_v(S)$, a union bound tells us that

$$\Pr_{h_u,h_v}[\text{mode}(h_u) = \text{mode}(h_v)] \geq 1 - 6\rho.$$

Part 2. For the second part, we will need the correctness property (4) of $\mathcal{A}$ and the fact that h_u is a uniformly random hypothesis in $\mathcal{H}$. First, let $h \sim \mathcal{H}$ be a uniformly random hypothesis in $\mathcal{H}$, and let $S \sim \mathcal{D}^n$ be n i.i.d. samples drawn from $\mathcal{D}$, independently of h. The correctness property of $\mathcal{A}$ requires that

$$\Pr_{h,S}[\text{er}_h(\mathcal{A}(S,h(S))) \leq \varepsilon] \geq 1 - 3\delta. \tag{7}$$

In words, this just says that the error of the classifier produced by $\mathcal{A}$ can be more than ε with probability at most 3δ. We can then use the law of total probability to get that

$$\begin{aligned}
1 - 3\delta &\leq \Pr_{h,S}[\text{er}_h(\mathcal{A}(S,h(S)) \leq \varepsilon, \mathcal{A}(S,h(S)) = \text{mode}(h)] \\
&\quad + \Pr_{h,S}[\text{er}_h(\mathcal{A}(S,h(S))) \leq \varepsilon, \mathcal{A}(S,h(S)) \neq \text{mode}(h)] \\
&\leq \Pr_h[\text{er}_h(\text{mode}(h)) \leq \varepsilon] + \Pr_{h,S}[\mathcal{A}(S,h(S)) \neq \text{mode}(h)] \\
&\leq \Pr_h[\text{er}_h(\text{mode}(h)) \leq \varepsilon] + 3\rho,
\end{aligned}$$

where the last inequality follows from Lemma 3.6. This implies that

$$\Pr_h[\text{er}_h(\text{mode}(h)) \leq \varepsilon] \geq 1 - 3(\delta + \rho).$$

Now, since both h_u and h_v are uniformly random hypotheses in $\mathcal{H}$, the second part of the lemma follows by a union bound over the above inequality instantiated for these two. □

Finally, to prove Lemma 3.7 we will need one last lemma.

Lemma 3.8. *Given values $x_0, \ldots, x_d$ and $y_0, \ldots, y_d$ with the following properties:*

- *(Non-negative). $x_0, \ldots, x_d \geq 0$ and $y_0, \ldots, y_d \geq 0$,*
- *(Equal sum). $\sum_{i=0}^d x_i = \sum_{j=0}^d y_j$,*
- *(Dominating). For all $k \in \{0, \ldots, d\}$, we have $\sum_{i=0}^k x_i \leq \sum_{j=0}^k y_j$.*

Then there exist values $p_{i,j}$ where $0 \leq j \leq i \leq d$, such that

1. *for all $0 \leq j \leq i \leq d$, we have $0 \leq p_{i,j} \leq 1$,*
2. *for all $i \in \{0, \ldots, d\}$, we have $\sum_{j=0}^i p_{i,j} = 1$,*

3. for all $j \in \{0, \dots, d\}$, we have $\sum_{i=j}^{d} x_i \cdot p_{i,j} = y_j$.

We will note that an almost identical version of this lemma is proved by Li [27, Lemma 1]; Bruno and Vaccaro [28, Theorem 3.4]. The only difference is that they require x and y to be ordered. This means they can compute the values $p_{i,j}$ more efficiently. However, in this lower bound, we only need the existence of such values, and we cannot make sure that x, y are ordered. For completeness, we will therefore include a proof of this version in Appendix A. We are now ready to prove Lemma 3.7.

Proof of Lemma 3.7. We will describe a different way of taking a step in the graph and then show that this has the same distribution as taking a uniformly random step.

First, let $\sigma_0, \dots, \sigma_{d-1} \in \{-1, 1\}$ be d uniformly random signs, and let

$$P = \{i \in [d] \mid h_u(S) = h_{u+\sigma_i e_i}(S)\}, \tag{8}$$

where e_i is the i^{th} standard basis vector. That is, P is the set of all axis-aligned directions in which we could take a step from h_u to $h_{u+\sigma_i e_i}$ without seeing any change in labels over the sample S. The way we will take a step is then to pick a subset $P' \subseteq P$ and compute $z' = \sum_{i \in P'} \sigma_i e_i$. We then take a step in direction z'. We therefore prove that z' is uniform on Z if we pick the subset P' in the right way. Let z be a uniformly random element of Z. Notice since the directions σ_i are uniformly random, it is sufficient to show that P' has the same distribution as $Q = \{i \in [d] \mid z_i \neq 0\}$. We will first argue that we can make $|P'|$ have the same distribution as $|Q|$.

To do this, we can invoke lemma 3.8 with $x_i = \Pr[|P| = i]$ and $y_i = \Pr[|Q| = i]$. Then $x := x_0, \dots, x_d$ and $y := y_0, \dots, y_d$ clearly satisfy the non-negativity and equal sum requirements of the lemma; what remains to be shown is that $\Pr[|P| \leq t] \leq \Pr[|Q| \leq t]$ for all $t \in \{0, \dots, d\}$. If this is indeed the case, then the lemma statement implies that there exists a distribution which tells us how many elements to remove from P to make $|P'|$ have the same distribution as $|Q|$. Namely, if one samples P with $|P| = i$, then one can obtain P' by removing $i - j$ elements from P with probability $p_{i,j}$, for the $p_{i,j}$ values given by the lemma. We can then choose the elements to remove from P uniformly at random. Because of symmetry in the contents of P and Q, this would make sure that the elements that *remain* in P' are uniformly random, and thus P' will have the same distribution as Q. We will now show that $\Pr[|P| \leq t] \leq \Pr[|Q| \leq t]$ for every $t \in \{0, \dots, d\}$.

First, note that $|Q| \sim \text{Binomial}\big(d, \frac{2}{3}\big)$. We therefore have

$$\Pr[|Q| \leq t] = \sum_{i=0}^{t} \binom{d}{i} \left(\frac{2}{3}\right)^i \left(\frac{1}{3}\right)^{d-i} \geq \sum_{i=0}^{t} \binom{d}{i} 3^{-d}. \tag{9}$$

Now before looking at $|P|$, we introduce some notation. For any h_u, and for a specific direction i, $h_u(x) \neq h_{u+\sigma_i e_i}(x)$ for exactly two points $x \in \{x_i^1, x_i^2\}$. This follows directly from the definition of h_u, since we can only distinguish h_u and $h_{u+\sigma_i e_i \bmod k}$ in the endpoints of the interval it induces. Therefore, the event $P = \{l_1, \dots, l_i\}$ implies

$$\bigcap_{j \notin P} \{x_j^1 \in S \vee x_j^2 \in S\}.$$

In words, this just means that S contained either endpoint of the interval for all directions $j \notin P$. Notice that the events in the intersection above (across different values of j) are *negatively correlated*; namely, given that one of x_j^1 or x_j^2 is in S, it is less likely for $x_{j'}^1$ or $x_{j'}^2$ to also be in S. Furthermore, by symmetry, $\Pr[P = \{l_1, \dots, l_i\}] = \Pr[P = \{1, \dots, i\}]$. Using these observations, we

can bound $\Pr[|P| \le t]$ as:

$$\begin{aligned}
\Pr[|P| \le t] &= \sum_{i=0}^{t} \Pr[|P| = i] = \sum_{i=0}^{t} \binom{d}{i} \Pr[P = \{1, \dots, i\}] \\
&\le \sum_{i=0}^{t} \binom{d}{i} \Pr[\cap_{j=i+1}^{d} \{x_j^1 \in S \vee x_j^2 \in S\}] \\
&\le \sum_{i=0}^{t} \binom{d}{i} \prod_{j=i+1}^{d} \Pr[x_j^1 \in S \vee x_j^2 \in S] && \text{(negative correlation)} \\
&= \sum_{i=0}^{t} \binom{d}{i} \prod_{j=i+1}^{d} \left(1 - \Pr[x_j^1 \notin S \wedge x_j^2 \notin S]\right) \\
&= \sum_{i=0}^{t} \binom{d}{i} \prod_{j=i+1}^{d} \left(1 - \left(1 - \frac{2}{kd}\right)^n\right) \\
&= \sum_{i=0}^{t} \binom{d}{i} \left(1 - \left(1 - \frac{2}{kd}\right)^n\right)^{d-i} \\
&\le \sum_{i=0}^{t} \binom{d}{i} \left(1 - \exp\left(\frac{-4n}{kd}\right)\right)^{d-i} && (1 - x \ge e^{-2x} \text{ for } x \in [0, 3/4]) \\
&\le \sum_{i=0}^{t} \binom{d}{i} \left(\frac{1}{81}\right)^{d-i}. && (\text{since } k \ge \tfrac{4n}{\log(81/80)d})
\end{aligned}$$

Now note that the i'th term in the sum above is smaller than the i'th term in (9) as long as $t \le 3d/4$. For $d > t > 3d/4$, we can instead bound the complementary event.

$$\begin{aligned}
\Pr[|P| \le t] = 1 - \Pr[|P| > t] \le 1 - \Pr[|P| = d] = 1 - \left(1 - \frac{2d}{kd}\right)^n \\
\le 1 - \exp(-4n/k) \le 1 - \left(\frac{80}{81}\right)^{d/10} \le 1 - (0.997)^d.
\end{aligned}$$

For $|Q|$, we can do a Chernoff bound, using $t > 3d/4$, to get that

$$\Pr[|Q| > t] \le \Pr[|Q| > 3d/4] \le \exp\left(\frac{-d}{288}\right) \le (0.997)^d$$

and thus, for $t > 3d/4$, we have

$$\Pr[|P| \le t] \le 1 - (0.997)^d \le 1 - \Pr[|Q| > t] = \Pr[|Q| \le t].$$

This finishes the down-sampling part of the proof.

We now move on to proving independence of v and S. We have already shown for any fixed $a \in \mathbb{Z}_k^d$ that $\Pr[v = a] = k^{-d}$ without fixing S. To show independence of v and S, it is therefore enough to argue that $\Pr[v = a \mid S] = k^{-d}$. Therefore, fix the sample S. Then, suppose for any $b \in \mathbb{Z}_k^d$, the following equality holds:

$$\Pr[v = a \mid S, u = b] = \Pr[v = b \mid S, u = a]. \tag{10}$$

Then we can see that

$$\begin{aligned}
\Pr[v = a \mid S] &= \sum_{b \in \mathbb{Z}_k^d} \Pr[v = a \mid S, u = b] \Pr[u = b \mid S] \\
&= k^{-d} \sum_{b \in \mathbb{Z}_k^d} \Pr[v = a \mid S, u = b] && (u \text{ and } S \text{ independent}) \\
&= k^{-d} \sum_{b \in \mathbb{Z}_k^d} \Pr[v = b \mid S, u = a] \\
&= k^{-d}.
\end{aligned}$$

We therefore just need to prove equation (10).

Note that by definition of the sampling of v, both $\Pr[v = a \mid S, u = b]$ and $\Pr[v = b \mid S, u = a]$ are 0 when $h_a(S) \neq h_b(S)$. Hence, we will restrict our attention to values of a and b that satisfy $h_a(S) = h_b(S)$. In this case, let us further condition on the value of $\sigma \in \{-1, 1\}^d$ and recall that σ is drawn uniformly at random, and independently of both u and S. Additionally, for any σ, define σ' as

$$\sigma' = \left(\sigma_i \cdot (-1)^{\mathbf{1}\{a_i \neq b_i\}}\right)_{i=1}^{d}. \tag{11}$$

That is, σ' flips the sign of σ_i for all i where $a_i \neq b_i$. We will argue that

$$\Pr[v = a \mid S, \sigma, u = b] = \Pr[v = b \mid S, \sigma', u = a]. \tag{12}$$

We can then conclude that

$$\begin{aligned}
\Pr[v = a \mid S, u = b] &= \sum_{\sigma \in \{-1,1\}^d} \Pr[v = a \mid S, \sigma, u = b] \Pr[\sigma | S, u = b] \\
&= 2^{-d} \sum_{\sigma \in \{-1,1\}^d} \Pr[v = a \mid S, \sigma, u = b] && (\sigma \text{ independent of both } u \text{ and } S) \\
&= 2^{-d} \sum_{\sigma \in \{-1,1\}^d} \Pr[v = b \mid S, \sigma', u = a] && (\text{equation 12}) \\
&= 2^{-d} \sum_{\sigma' \in \{-1,1\}^d} \Pr[v = b \mid S, \sigma', u = a] \\
&= \sum_{\sigma' \in \{-1,1\}^d} \Pr[v = b \mid S, \sigma', u = a] \Pr[\sigma' | S, u = a] \\
&= \Pr[v = b \mid S, u = a],
\end{aligned}$$

which is the desired equality. To show equation 12, notice first that by definition of the sampling of v, $\Pr[v = a \mid S, \sigma, u = b] = 0$ if σ does not satisfy:

$$\exists D \subseteq [d] : a = \left(b + \sum_{i \in D} \sigma_i e_i\right) \bmod k. \tag{13}$$

Note how this implies, by virtue of how σ' is defined, that if σ does not satisfy equation 13, then it is also the case that $\Pr[v = b \mid S, \sigma', u = a] = 0 = \Pr[v = a \mid S, \sigma, u = b]$.

Now, let us consider a value of σ that satisfies equation 13. If we condition on σ, along with S and $u = b$, this completely determines the random variable P (defined in equation 8) to be some set

$P_1 \subseteq [d]$ (and in fact, the set D which witnesses equation 13 is some subset of P_1). Furthermore, in this case, $\Pr[v = a \mid S, \sigma, u = b] = p_{i,j} \cdot \binom{i}{\|a-b\|_1}^{-1}$, where $i = |P_1|$ and $j = \|a - b\|_1$, for the $p_{i,j}$ values given by lemma 3.8.

We then claim that, when we condition on S, but with $u = a$ and the signs being σ' instead, the value P_2 that the random variable P gets determined to be is *still equal to* P_1. To see this, first consider any $i \in [d] \setminus P_1$. For such an i, it holds that $a_i = b_i$, and hence $\sigma_i = \sigma'_i$. Since i was not included in P_1, $h_b(S) \neq h_{b+\sigma_i e_i \bmod k}(S)$, which also means that $h_a(S) \neq h_{a+\sigma'_i e_i \bmod k}(S)$. So, $i \neq P_2$.

Now, consider any $i \in P_1 \setminus D$. Again, for any such i, $a_i = b_i$ and hence $\sigma_i = \sigma'_i$. Consider any point $(x_1, x_2) \in S$ with $x_1 = i$; since $i \in P_1$, it holds that $h_b((x_1, x_2)) = h_{b+\sigma_i e_i \bmod k}((x_1, x_2))$, and hence, $h_a((x_1, x_2)) = h_{a+\sigma'_i e_i \bmod k}((x_1, x_2))$. Since these are the only points affected, we have that $h_a(S) = h_{a+\sigma'_i e_i \bmod k}(S)$, meaning that $i \in P_2$.

Finally, consider any $i \in D$. We have that $a_i = b_i + \sigma_i \bmod k$, meaning that $\sigma'_i = -\sigma_i$. Consider any point $(x_1, x_2) \in S$ with $x_1 = i$; since $i \in P_1$, it holds that $h_b((x_1, x_2)) = h_{b+\sigma_i e_i \bmod k}((x_1, x_2))$. But this immediately also means that $h_{a+\sigma'_i e_i \bmod k}((x_1, x_2)) = h_a((x_1, x_2))$. Since these are the only points affected, we have that $h_a(S) = h_{a+\sigma'_i e_i \bmod k}(S)$, meaning that $i \in P_2$.

In summary, we have argued that $\Pr[v = b \mid S, \sigma', u = a]$ is also equal to $p_{i,j} \cdot \binom{i}{\|a-b\|_1}^{-1}$, where $i = |P_2| = |P_1|$ and $j = \|b - a\|_1$; but this is also the value of $\Pr[v = a \mid S, \sigma, u = b]$, concluding the proof. □

3.2 Expansion Property of G

In this section, we will prove Theorem 3.3. For convenience, we restate the theorem here.

Theorem 3.3. *Let $T \subseteq \mathbb{Z}_k^d$ be a subset of the vertices. If $\#\{(u, v) \in E \mid u, v \in T\} \geq (24/25)|T||Z|$ then $|T| \geq \left(\frac{\sqrt{d}}{450 \log d}\right)^d$.*

Throughout this section, we use the notation $R(Y)$ for the rank of any matrix Y and remind the reader that $\langle \cdot, \cdot \rangle$ denotes the mod k inner product on $\mathbb{Z}_k^d$ considered as a vector space over the field $\mathbb{Z}_k$. Also, let A denote the adjacency matrix of G. The proof is split into several parts.

3.2.1 Relating Expansion Property to the Spectrum of A

Let $\mathbf{1}_T \in \{0, 1\}^{k^d}$ denote the indicator vector on T. Then one easily checks that

$$\#\{(u, v) : u, v \in T\} = \langle A\mathbf{1}_T, \mathbf{1}_T \rangle.$$

In order to better understand this number, we compute the eigenvectors and eigenvalues of A. The spectral theory of Cayley graphs on Abelian groups is well understood [29, 30], but we provide all proofs here for completeness.

Lemma 3.9 (Eigenvectors and eigenvalues of A)**.** *For any $v \in \mathbb{Z}_k^d$, let χ_v[2] be the vector indexed by $\mathbb{Z}_k^d$ with entries $\chi_v(w) = k^{-d/2} \exp(2\pi i \langle v, w \rangle / k)$. Then χ_v is an eigenvector of A with corresponding eigenvalue*

$$\lambda_v = |Z| - 2 \sum_{z \in Z} \sin^2(\pi \langle v, z \rangle / k).$$

Furthermore, the collection $(\chi_v)_{v \in \mathbb{Z}_k^d}$ is an orthonormal basis of $\mathbb{C}^{k^d}$.

[2]While the letter χ is usually used for the characters of the group, we emphasize here that our χ_v are the characters scaled by $k^{-d/2}$ to ensure unit norm.

Proof. Let Z^+ denote the set of vectors in Z where the first non-zero coordinate is 1. observe that

$$\begin{aligned}
(A\chi_v)(w) &= \sum_{z\in Z} \chi_v(w+z) \\
&= \chi_v(w+(0)^d) + \sum_{z\in Z^+} (\chi_v(w+z) + \chi_v(w-z)) \\
&= \chi_v(w) + k^{d/2}\chi_v(w) \sum_{z\in Z^+} (\chi_v(z) + \chi_v(-z)) \\
&= \chi_v(w) + \chi_v(w) \sum_{z\in Z^+} (\exp(2\pi i\langle v,z\rangle/k) + \exp(-2\pi i\langle v,z\rangle/k)) \\
&= \chi_v(w) + \chi_v(w) \sum_{z\in Z^+} 2\cos(2\pi\langle v,z\rangle/k) \\
&= \chi_v(w) + \chi_v(w) \sum_{z\in Z^+} (2 - 4\sin^2(\pi\langle v,z\rangle/k)) \\
&= |Z|\chi_v(w) - 2\chi_v(w) \sum_{z\in Z} \sin^2(\pi\langle v,z\rangle/k).
\end{aligned}$$

This shows that χ_v is indeed an eigenvector with the claimed eigenvalue. We now show that they are an orthonormal basis

$$\langle \chi_v, \chi_w\rangle_{\mathbb{C}} = \sum_{u\in\mathbb{Z}_k^d} \chi_v(u)\overline{\chi_w(u)} = \sum_{u\in\mathbb{Z}_k^d} \chi_v(u)\chi_{-w}(u) = \sum_{u\in\mathbb{Z}_k^d} \chi_{v-w}(u).$$

If $v = w$, each term in the sum is equal to k^{-d}, making the whole expression 1. If instead $v \neq w$, there is at least one coordinate j such that $v_j - w_j \neq 0$. Hence

$$k^{-d} \sum_{u\in\mathbb{Z}_k^d} \chi_{v-w}(u) = k^{-d} \sum_{u_1\in\mathbb{Z}_k} e^{2\pi i(v_1-w_1)u_1/k} \cdots \sum_{u_d\in\mathbb{Z}_k} e^{2\pi i(v_d-w_d)u_d/k}.$$

Looking at just the j'th sum:

$$\sum_{u_j\in\mathbb{Z}_k} e^{2\pi i(v_j-w_j)u_j/k} = \sum_{\ell=0}^{k-1} \Big(e^{2\pi i(v_j-w_j)/k}\Big)^\ell = \frac{\big(e^{2\pi i(v_j-w_j)/k}\big)^k - 1}{\big(e^{2\pi i(v_j-w_j)/k}\big) - 1} = 0.$$

Hence, in this case $\langle \chi_v, \chi_w\rangle_{\mathbb{C}} = 0$. Thus, the eigenvectors have unit length and are orthogonal, so they comprise an orthonormal basis □

We now show how to bound $\langle A\mathbf{1}_T, \mathbf{1}_T\rangle$ in terms of the eigenvalues. First, let $\mu_1, \ldots, \mu_{k^d}$ be the eigenvalues of A sorted in increasing order and χ_i be the eigenvector associated with μ_i.

Lemma 3.10. *It holds that*

$$\langle A\mathbf{1}_T, \mathbf{1}_T\rangle \leq |T|\Big(|Z|/2 + \mu_{k^d-k^d/(2|T|)}\Big)$$

Proof. By expanding $\mathbf{1}_T$ in the eigenvector basis and using the fact that χ_v is an eigenvector of A, we can then write

$$\langle A\mathbf{1}_T, \mathbf{1}_T\rangle = \sum_{v\in\mathbb{Z}_k^d} |\langle \mathbf{1}_T, \chi_v\rangle|^2 \lambda_v.$$

From the definition of χ_v, $|\langle \mathbf{1}_T, \chi_v \rangle|^2 \leq |T|^2 k^{-d}$. Hence

$$\sum_{i=k^d - k^d/(2|T|)+1}^{k^d} |\langle \mathbf{1}_T, \chi_i \rangle|^2 \leq |T|/2.$$

Also, it follows from Lemma 3.9 that $\lambda_v \leq |Z|$ for all v and also that $\lambda_0 = |Z|$. Hence, the largest eigenvalue is $\mu_{k^d} = |Z|$. This implies that

$$\sum_{i=k^d - k^d/(2|T|)+1}^{k^d} \mu_i |\langle \mathbf{1}_T, \chi_i \rangle|^2 \leq 3^d |T|/2.$$

Furthermore, $\|\mathbf{1}_T\|^2 = |T|$ meaning that

$$\sum_{i=1}^{k^d - k^d/(2|T|)} \mu_i |\langle \mathbf{1}_T, \chi_i \rangle|^2 \leq \mu_{k^d - k^d/(2|T|)} |T|.$$

Combining all this yields

$$\langle A\mathbf{1}_T, \mathbf{1}_T \rangle \leq |T| \Big(|Z|/2 + \mu_{k^d - k^d/(2|T|)} \Big).$$

□

Using lemma 3.10, we can bound the size of $|T|$ in terms of the tail probabilities of a suitable binomial random variable related to the eigenvalues. For this, let $\mathcal{I} = \{\lfloor k/4 \rfloor + 1, \ldots, \lfloor k/4 \rfloor + \lfloor k/2 \rfloor\}$ and note that $|\mathcal{I}| = \lfloor k/2 \rfloor$ and let u be a uniform random variable on $\mathbb{Z}_k^d$. For each $z \in Z$, define an indicator X_z taking the value 1 if $\langle u, z \rangle \notin \mathcal{I}$ and 0 otherwise.

Lemma 3.11. *Define*

$$p = \Pr\left[\sum_{z \in Z} X_z \geq (42/50)|Z| \right].$$

Then $|T| \geq p^{-1}/2$.

Proof. Consider a $u \in \mathbb{Z}_k^d$ such that

$$\sum_{z \in Z} \mathbf{1}\{\langle u, z \rangle \notin \mathcal{I}\} < (42/50)|Z|$$

Then,

$$\begin{aligned}
\lambda_u &= |Z| - 2 \sum_{z \in Z} \sin^2(\pi \langle u, z \rangle / k) \\
&\leq |Z| - 2 \sum_{z \in Z} \mathbf{1}\{\langle u, z \rangle \in \mathcal{I}\} \sin^2(\pi \langle u, z \rangle / k) \\
&< |Z| - 2 \sum_{z \in Z} \mathbf{1}\{\langle u, z \rangle \in \mathcal{I}\} \sin^2(\pi/4) \\
&\leq |Z| - (8/50)|Z| = (42/50)|Z| \leq (46/50)|Z|.
\end{aligned}$$

Negating the implication we just showed, we then have that

$$\lambda_u > (46/50)|Z| \Rightarrow \sum_{z\in Z} \mathbf{1}\{\langle u, z\rangle \notin \mathcal{I}\} \geq (42/50)|Z|).$$

This then means that over uniformly random $u \in \mathbb{Z}_k^d$

$$\Pr[\lambda_u > (46/50)|Z|] \leq \Pr\left[\sum_{z\in Z} X_z \geq (42/50)|Z|\right] = p.$$

Hence, for any $j > pk^d$, we have $\mu_{k^d-j} < (46/50)|Z|$. In particular, if $1/(2|T|) \geq p$ it must be the case that $\mu_{k^d-k^d/(2|T|)} < (46/50)|Z|$. However, recalling Lemma 3.10 and our assumptions, we have

$$(24/25)|T||Z| \leq \langle A\mathbf{1}_T, \mathbf{1}_T\rangle \leq |T|\Big(|Z|/2 + \mu_{k^d-k^d/(2|T|)}/2\Big),$$

such that $\mu_{k^d-k^d/(2|T|)} \geq (46/50)|Z|$. Thus, it must be the case that $\frac{1}{2|T|} < p$ or equivalently

$$|T| > p^{-1}/2. \qquad \square$$

3.2.2 Bounding p

The strategy is now to bound p using Markov's inequality with sufficient control over the central moments. Specifically, for $r \leq d$ we bound

$$\mathbb{E}\left[\left|\sum_{z\in Z}(X_z - \mathbb{E}[X_z])\right|^r\right] = \sum_{Y\in Z^r} \mathbb{E}\left[\prod_{i=0}^{r-1}(X_{y_i} - \mathbb{E}[X_{y_i}])\right],$$

where we identify Z^r with the set of matrices $\{-1,0,1\}^{d\times r}$. Since k is the power of a prime, it holds that X_{y_i} is independent of $(y_j)_{j\neq i}$ if y_i is linearly independent of $(y_j)_{j\neq i}$ over $\mathbb{F}_k^d$ (see Lemma A.1). Hence,

$$\sum_{Y\in Z^r} \mathbb{E}\left[\prod_{i=0}^{r-1}(X_{y_i} - \mathbb{E}[X_{y_i}])\right] = \sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j : i\neq j)\forall i}} \mathbb{E}\left[\prod_{i=0}^{r-1}(X_{y_i} - \mathbb{E}[X_{y_i}])\right].$$

We must then bound the number of $Y = (y_0, \ldots, y_{r-1}) \in Z^r$ that satisfy $y_i \in \mathrm{span}(y_j : i \neq j)$ for all $i \in [r]$. We split this into two parts depending on the rank of the matrix Y. We begin with a simple lemma.

Lemma 3.12. *Let V be an r-dimensional subspace of $\mathbb{F}_k^d$. Then $|V \cap \{-1,0,1\}^d| \leq 3^r$.*

Proof. Let $v_0, \ldots, v_{r-1}$ be an arbitrary basis of V and consider the matrix A with v_i's as rows. Since the column rank is equal to the row rank, there is a set R of r linearly independent columns. If $x \in \mathbb{F}_k^r$ is a vector so that $xA \in \{-1,0,1\}^d$, then in particular, $xA_R \in \{-1,0,1\}^r$. There are 3^r choices for xA_R, and any such choice forces r linearly independent constraints on x, each resulting in a unique choice of x. It follows that $|V \cap \{-1,0,1\}^d| \leq |\{x \in \mathbb{F}_k^r : xA_R \in \{-1,0,1\}^r\}| \leq 3^r$. $\square$

With this lemma in hand, we can bound on the number of $Y \in Z^r$ with small rank.

Lemma 3.13. *Assume $r \le d/2$. Then the number of $Y \in Z^r$ with rank $R(Y) \le r - \log_3(d)$ is at most $d^{-d/2}/d|Z|^r$.*

Proof. Consider drawing Y one vector y_i at a time, each obtained by sampling each coordinate independently, taking the values $\{-1, 0, 1\}$ uniformly. Now for all $\ell \in [r]$, let W_ℓ be a subspace of $\mathbb{F}_k^d$ of dimension $\ell - 1$ having maximal intersection with Z (which is less than $3^{\ell-1} \le 3^{r-1}$ by Lemma 3.12). Then letting $s = \log_3(d)$,

$$\begin{aligned}
\Pr[\dim \operatorname{span}(y_0, \ldots, y_{r-1}) \le r - s] &= \Pr[\exists i_0, \ldots, i_{s-1} \in [r] : y_{i_\ell} \in \operatorname{span}\{y_{i_0}, \ldots, y_{i_\ell - 1}\}\, \forall \ell \in [s]] \\
&\le \Pr[\sum_{\ell=0}^{r-1} \mathbf{1}\{y_\ell \in W_\ell\} \ge s] = \sum_{j=s}^{r} \Pr[\sum_{\ell=0}^{r-1} \mathbf{1}\{y_\ell \in W_\ell\} = j] \\
&\le \sum_{j=s}^{r} \binom{r}{j} \left(3^{(r-1-d)}\right)^j \le 2^r 3^{s(r-d-1)} = 2^r d^{r-d-1} \\
&\le 2^{d/2} d^{-d/2}/d = \left(\frac{d}{2}\right)^{-d/2}/d
\end{aligned}$$

Multiplying with $|Z|^r$ to get the total number then yields the result. □

We now move on to bound the number of large rank matrices with the additional property that each column lies in the span of the other columns.

Lemma 3.14. *Let $y_0, \ldots, y_{r-1}$ be vectors in $\mathbb{F}_k^d$ for prime k with the property that every $y_i \in \operatorname{span}(y_j : i \ne j)$ and $\dim(\operatorname{span}(Y)) = r - s$ with $s \ge 1$. Then there is a subset of indices $S \subseteq [r]$ with $|S| \ge r/s$ so that $\dim(\operatorname{span}(\{y_i\}_{i \in S})) = |S| - 1$ but for any subset of $S' \subseteq S$ with $|S'| = |S| - 1$, the corresponding vectors are linearly independent.*

Proof. Assume for the sake of contradiction that the claim is false. Let M be the matrix having the y_i's as rows.

Now consider the following process for constructing a basis $b_0, \ldots, b_{s-1}$ for the left nullspace of M, one vector at a time. Initialize $T = \emptyset$ as the set of indices j so that at least one b_i among already constructed basis vectors has non-zero j'th coordinate. For $i = 0, \ldots, r-1$, pick a vector y_j with $j \notin T$ and write it as a linear combination $y_j = \sum_{h \ne j} \alpha_h y_h$. If there are multiple such linear combinations, pick one with the smallest number of non-zero coefficients α_h, breaking ties arbitrarily. Then the vector b_i with h'th coordinate α_h for $h \ne j$ and j'th coordinate -1 is in the left nullspace of M. Furthermore, it is linearly independent of $b_0, \ldots, b_{i-1}$ since these vectors all have 0 in their j'th coordinate by definition of T. We thus add it as the i'th basis vector and update $T \leftarrow T \cup \{h : \alpha_h \ne 0\} \cup \{j\}$.

We now bound $|S|$ with $S = \{h : \alpha_h \ne 0\} \cup \{j\}$. We now show that the vectors $\{y_h\}_{\alpha_h \ne 0}$ are linearly independent. Indeed, if there was a linear dependence, we could write $\sum_{h \ne j} \alpha_h y_h$ as a linear combination of a smaller number of vectors, contradicting the fact that we picked a linear combination with the smallest number of non-zero coefficients. Thus $\dim(\operatorname{span}(\{y_i\}_{i \in S})) = |S| - 1$ and the subset $S' = S \backslash \{j\}$ is linearly independent. Finally, consider any subset S' with $|S'| = |S| - 1$ excluding any other $h \ne j$. Again, if there was a linear dependency, we could again write y_j as a linear combination of fewer vectors, contradicting the choice of coefficients α_h. Hence, since we assumed the theorem is not true, it must be that case that $|S| < r/s$, meaning that each b_i adds fewer than r/s indices to T. Thus upon selecting b_i, we have $|T| < (i-1)r/s$. The process can thus continue until $i = s + 1$. This contradicts that the nullspace has dimension s. □

Before proceeding, we need a variant of the classical Littlewood-Offord lemma.

Lemma 3.15. *Let $\mathbb{F}_k$ be a prime field. Let $s \geq 1$, $x_0, \ldots, x_{s-1} \in \mathbb{F}_k \setminus \{0\}$ and $y \in \mathbb{F}_k$. Then for ε_i sampled independently and uniformly in $\{-1, 0, 1\}$ we have*

$$\Pr\left[\sum_{i=0}^{s-1} \varepsilon_i x_i = y\right] \leq \min\left\{\frac{1}{2}, \frac{1}{k} + \exp(-s/8) + \sqrt{\frac{32}{s}}\right\}.$$

Proof. We use Lemma 2.5 in Golovnev et al. [26] to conclude that the probability is at most

$$\begin{aligned}
&\sum_{z=0}^{s} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} \min\left\{\frac{1}{2}, \left(\frac{1}{k} + \sqrt{\frac{8}{z}}\right)\right\} \\
&\leq \frac{1}{2} \sum_{z=0}^{31} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} + \sum_{z=32}^{s} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} \min\left\{\frac{1}{2}, \left(\frac{1}{k} + \sqrt{\frac{8}{z}}\right)\right\} \\
&\leq \min\left\{\frac{1}{2} \sum_{z=0}^{s} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z}, \frac{1}{2} \sum_{z=0}^{31} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} + \sum_{z=32}^{s} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} \left(\frac{1}{k} + \sqrt{\frac{8}{z}}\right)\right\} \\
&\leq \min\left\{\frac{1}{2}, \frac{1}{k} + \sum_{z=0}^{s/4} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} + \sum_{z=s/4}^{s} \binom{s}{z} \left(\frac{2}{3}\right)^z \left(\frac{1}{3}\right)^{s-z} \sqrt{\frac{8}{z}}\right\}.
\end{aligned}$$

By the Chernoff bound, we have $\sum_{z=0}^{s/4} \binom{s}{z} (2/3)^z (1/3)^{s-z} \leq \exp(-s/8)$. Hence the sum is at most

$$\min\left\{\frac{1}{2}, \frac{1}{k} + \exp(-s/8) + \sqrt{\frac{32}{s}}\right\}. \qquad \square$$

Lemma 3.16. *Assume that r is an even number in the interval $[d/(\log_3(d)/2), d/\log_3(d)]$. Then the fraction of matrices $Y \in Z^r$ with rank $R(Y) \geq r - \log_3(d)$ that have at least one column independent of the other columns is at least*

$$d 2^{2d} (1/k + 16\sqrt{\log_3^2(d)/d})^{d - d/\log_3(d)}.$$

Proof. Let Y be a uniformly random matrix in Z^r with columns $y_0, \ldots, y_{r-1}$. Due to Lemma 3.14, we have

$$\Pr[\{y_i \in \operatorname{span}(\{y_j\}_{j \neq i}) \, \forall i\} \cap \{R(Y) = r - s\}] \leq \Pr[B_s].$$

Here B_s is the event that there exists a subset of indices $S \subseteq [r]$ with $|S| \geq r/s$ such that $\dim(\operatorname{span}(\{y_i\}_{i \in S})) = |S| - 1$ but for any subset of $S' \subseteq S$ with $|S'| = |S| - 1$, the corresponding vectors are linearly independent. Note that $\Pr[B_i] \leq \Pr[B_{i+1}]$ for any $i \in [r-1]$. Thus

$$\begin{aligned}
&\Pr[\{y_i \in \operatorname{span}(\{y_j\}_{j \neq i}) \, \forall i\} \cap \{R(Y) > r - \log_3(d)\}] \\
&= \sum_{s=1}^{\log_3(d)} \Pr[\{y_i \in \operatorname{span}(\{y_j\}_{j \neq i}) \, \forall i\} \cap \{R(Y) = r - s\}] \\
&\leq \sum_{s=1}^{\log_3(d)} \Pr[B_s] \leq \log_3(d) \Pr[B_{\log_3(d)}].
\end{aligned}$$

We must then bound $\Pr[B_{\log_3(d)}]$. On $B_{\log_3(d)}$ there exists a set S of size $|S| \geq r/\log_3(d)$ with the properties mentioned above. Notice that the event $B_{\log_3(d)}$ also implies the existence of a subset of rows T with $|T| = |S| - 1$ so that the submatrix $Y_{T,S}$ has full row rank, i.e. rank $R(Y_{T,S}) \geq r/\log_3(d) - 1$. Now, for any fixed $S \subseteq [r], T \subseteq [d]$ where $|S| \geq r/\log_3(d)$ and $|T| = \log_3(d) - 1$, define $E_{S,T}$ as the event that $Y_{T,S}$ has full row rank, and for every subset $S' \subseteq S$ with $|S'| = |S| - 1$, the matrix $Y_{S'}$ has full column rank. Then by the above remarks,

$$\Pr[B_{\log_3(d)}] \leq \Pr\left[\bigcup_{\substack{(S,T)\subseteq[r]\times[d]\\|S|\geq r/\log_3(d)\\|T|=|S|-1}} E_{S,T}\right] \leq \sum_{\substack{(S,T)\subseteq[r]\times[d]\\|S|\geq r/\log_3(d)\\|T|=\log_3(d)-1}} \Pr[E_{S,T}]$$

Now, observe that for $E_{S,T}$ to occur, we must have that the dimension of the right nullspace of Y_S is 1. Let $b \in \mathbb{Z}_k^{|S|}$ be the smallest vector in the lexicographical ordering spanning this space and observe that all its entries are non-zero as otherwise deleting a column corresponding to a zero would result in a set of $|S| - 1$ vectors with a linear dependence. Any row x of Y_S with index in $[d] \setminus T$ must satisfy $\langle x, b\rangle = 0$ for b to be in the right nullspace of Y_S. Now, observe that b is fixed when conditioning on $Y_{T,S}$ and conditional on $Y_{T,S}$, any row of Y_S with index in $[d] \setminus T$ is uniform on $\{-1, 0, 1\}^{|S|}$. Thus, since $\mathbb{Z}_k$ is a prime field, Lemma 3.15 implies that

$$\Pr[\langle x, b\rangle = 0] = \mathbb{E}[\Pr[\langle x, b\rangle = 0 | Y_{T,S}]] \leq \min\{1/2, 1/k + \exp(-|S|/8) + \sqrt{32/|S|}\}$$

where x is a row of Y_S with index in $[d]\setminus T$. Since we assumed $r \geq d/\log_3(d)$ and $|S| \geq r/\log_3(d)$, this probability is at most $1/k + 16\sqrt{\log_3^2(d)/d}$ for d sufficiently large. Since all such x are independent, we conclude that

$$\begin{aligned}
\Pr[E_{S,T}] &\leq \Pr[\langle x, b\rangle = 0, \forall x \text{ row in } Y_{S,[d]\setminus T}] \\
&\leq (1/k + 16\sqrt{\log_3^2(d)/d})^{d-(|S|-1)} \\
&\leq (1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}
\end{aligned}$$

We can now go back and bound

$$\Pr[\{y_i \in \text{span}(\{y_j\}_{j\neq i})\, \forall i\} \cap \{R(Y) = r - s\}] \leq \sum_{\substack{(S,T)\subseteq[r]\times[d]\\|S|\geq r/\log_3(d)\\|T|=|S|-1}} \Pr[E_{S,T}]$$

$$\leq \sum_{\substack{(S,T)\subseteq[r]\times[d]\\|S|\geq r/\log_3(d)\\|T|=|S|-1}} \Pr[E_{S,T}](1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}$$

$$= \sum_{s=d/\log_3^2(d)}^{r} \binom{d - d/\log_3(d)}{s-1}\binom{d}{s}(1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}.$$

Bounding all binomial coefficients by 2^d, we finally get that the probability is at most

$$d2^{2d}(1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}.$$

□

With all our lemmas, we are now ready to bound the moments.

Lemma 3.17 (Moment bound)**.** *Let r an even number in the interval $[d/(2\log_3(d)), d/\log_3(d)]$. Then it holds that*

$$\mathbb{E}\left[\left|\sum_{z\in Z}(X_z - \mathbb{E}[X_z])\right|^r\right] \leq 113^d |Z|^r \log^d(d) d^{-d/2}$$

Proof.

$$\mathbb{E}\left[\left|\sum_{z\in Z}(X_z - 1/2)\right|^r\right] = \sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j:i\neq j)\forall i}} \mathbb{E}\left[\prod_{i=0}^{r-1}|X_{y_i} - \mathbb{E}[X_{y_i}]|\right]$$

$$= \sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j:i\neq j)\forall i \\ R(Y)\geq r-\log_3(d)}} \mathbb{E}\left[\prod_{i=0}^{r-1}|X_{y_i} - \mathbb{E}[X_{y_i}]|\right] + \sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j:i\neq j)\forall i \\ R(Y)< r-\log_3(d)}} \mathbb{E}\left[\prod_{i=0}^{r-1}|X_{y_i} - \mathbb{E}[X_{y_i}]|\right]$$

We bound the second sum in the following way

$$\sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j:i\neq j)\forall i \\ R(Y)< r-\log_3(d)}} \mathbb{E}\left[\prod_{i=0}^{r-1}|X_{y_i} - \mathbb{E}[X_{y_i}]|\right] \leq \#\{Y\in Z^r : R(Y) < r - \log_3(d)\}$$

$$\leq \frac{(d/2)^{-d/2}}{d}|Z|^r,$$

by Lemma 3.13. Now for the other sum,

$$\sum_{\substack{Y\in Z^r \\ y_i\in \mathrm{span}(y_j:i\neq j)\forall i \\ R(Y)\geq r-\log_3(d)}} \mathbb{E}\left[\prod_{i=0}^{r-1}|X_{y_i} - \mathbb{E}[X_{y_i}]|\right]$$

$$\leq \#\{Y\in Z^r : y_i \in \mathrm{span}(y_j : i\neq j)\forall i \text{ and } R(Y)\geq r-\log_3(d)\}$$

$$\leq d2^{2d}(1/k + 12\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}|Z|^r$$

In conclusion the moments are bounded as follows:

$$\mathbb{E}\left[\left(\sum_{z\in Z}(X_z - \mathbb{E}[X_z])\right)^r\right]$$

$$\leq |Z|^r\left(\frac{(d/2)^{-d/2}}{d} + d2^{2d-r}(1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}\right)$$

$$\leq 5^d|Z|^r(1/k + 16\sqrt{\log_3^2(d)/d})^{d-d/\log_3(d)}$$

$$\leq 5^d|Z|^r\left(17\log(d)/\sqrt{d}\right)^{d(1-1/\log_3(d))} \qquad \text{(since } k\geq \sqrt{d}/\log(d)\text{)}$$

$$\leq 150^d|Z|^r\log^d(d)d^{-d/2}$$

where we in the last inequality, we use the fact that $d^{\frac{d}{2\log_3(d)}} = (d^{\frac{1}{\log_3(d)}})^{d/2} = 3^{d/2}$. □

We can now bring everything together and prove Theorem 3.3.

Proof of Theorem 3.3. Applying Markov's inequality

$$\begin{aligned} p := \Pr\left[\sum_{z\in Z} X_z \geq (42/50)|Z|\right] &= \Pr\left[\left|\sum_{z\in Z}(X_z - \mathbb{E}[X_z])\right|^r \geq (17/50)^r |Z|^r\right] \\ &\leq (50/17)^r \frac{150^d |Z|^r \log^d(d) d^{-d/2}}{|Z|^r} \\ &\leq 450^d \log^d(d) d^{-d/2}. \end{aligned}$$

We can thus plug this bound on p into Lemma 3.11 and conclude that

$$|T| \geq \left(\frac{\sqrt{d}}{450\log(d)}\right)^d. \qquad \square$$

3.3 Upper Bounding $|C_f|$

In this section, we will prove Lemma 3.4. For this, we will need the following upper bound on the size of discrete ℓ_1-balls.

Lemma 3.18. *Denote the discrete ℓ_1-ball with center in the origin as $B_r = \{u \in \mathbb{Z}^d \mid \|u\|_1 \leq r\}$. Then, for $r \geq 1$ it holds that $|B_r| \leq 6^d r^2$.*

Proof. To compute the exact volume of such ℓ_1-ball, we can sum over all points with distance $i \leq r$ from the origin. This can be done with the stars and bars formula, and then multiplying with the possible number of signs. Using Bernoulli's inequality, we can bound this in the following way.

$$\begin{aligned} |B_r| &\leq \sum_{i=0}^{r} 2^d \binom{d+i-1}{d-1} \leq 2^d \sum_{i=0}^{r} \left(\frac{e(d+i-1)}{d-1}\right)^{d-1} \\ &= 2^d e^{d-1} \sum_{i=0}^{r} \left(1 + \frac{i}{d-1}\right)^{d-1} \leq 2^d e^{d-1} \sum_{i=0}^{r} (1+i) \leq 2^d e^{d-1} \left(r + 1 + \frac{r(r+1)}{2}\right) \\ &\leq 6^{d-1}(r+2)(r+1) = 6^{d-1}(r^2+3r+2) \leq 6^d r^2. \end{aligned} \qquad \square$$

We are now ready to prove Lemma 3.4. For convenience, we restate the lemma here.

Lemma 3.4. *For any $f \in \{0,1\}^{\mathcal{X}}$ we have $|C_f| \leq (6\varepsilon k)^d$.*

Proof. Fix an $f \in \{0,1\}^{\mathcal{X}}$ and for any $u, v \in \mathbb{Z}_k^d$ let $\nu : \mathbb{Z}_k^d \times \mathbb{Z}_k^d \to \{0, \ldots, \lfloor k/2 \rfloor\}$ be the metric defined by $\nu(u,v) = \sum_{i=0}^{d-1} \min(u_i - v_i, v_i - u_i)$ where the minimum is determined by the representatives in $[k]$. Remark that $\nu(u,v)$ can be interpreted as the shortest wrap-around ℓ_1 distance between u, v. Now, note that the number of of $x \in \mathcal{X}$ where $h_u(x) \neq h_v(x)$ is exactly $2 \cdot \nu(u,v)$. This means that $\mathrm{er}_{h_u}(h_v) = 2\nu(u,v)/|\mathcal{X}| = 2\nu(u,v)/(kd)$. Now, fix some $u \in C_f$, and set $r = \varepsilon k d$. Then, denote the discrete wrap-around ball with radius r centered in u as $B_{[r]}(u) = \{v \in \mathbb{Z}_k^d \mid \nu(u,v) \leq r\}$. Also, denote the discrete ℓ_1 ball with radius r centered in u as $B_r(u) = \{v \in \mathbb{Z}^d \mid \|u - v\|_1 \leq r\}$. Remark that $|B_{[r]}(u)| \leq |B_r(u)|$ since the wrap-around ball starts overlapping with itself when $r \geq \lceil k/2 \rceil$.

Now, assume for sake of contradiction there is a point $v \in C_f$ such that $\nu(u,v) > r$. This would mean that

$$\mathrm{er}_{h_u}(h_v) \cdot kd = 2\nu(u,v) > 2r = 2\varepsilon kd \implies \mathrm{er}_{h_u}(h_v) > 2\varepsilon.$$

However, we know from the definition of C_f, that

$$\begin{aligned}
\mathrm{er}_{h_u}(h_v) &= \Pr_{x\sim\mathcal{D}}[h_u(x) \neq h_v(x)] \\
&= \Pr_{x\sim\mathcal{D}}[(h_u(x) \neq f(x) \wedge h_v(x) = f(x)) \vee (h_u(x) = f(x) \wedge h_v(x) \neq f(x))] \\
&\leq \Pr_{x\sim\mathcal{D}}[h_u(x) \neq f(x)] + \Pr_{x\sim\mathcal{D}}[h_v(x) \neq f(x)] \\
&\leq 2\varepsilon
\end{aligned}$$

giving us a contradiction. Therefore, for every $v \in C_f$ we know that $\nu(u,v) \leq r$. This means that $C_f \subseteq B_{[r]}(x)$, meaning that $|C_f| \leq |B_{[r]}(x)| \leq |B_r(x)| \leq 6^d r^2$ where the last inequality follows from Lemma 3.18. Note, by the definition of k, we have $\varepsilon k \geq 2$. Therefore, when we plug in the value of r, we get

$$|C_f| \leq 6^d(\varepsilon kd)^2 \leq 6^d(\varepsilon k)^d 2^{2-d} d^2 \leq (6\varepsilon k)^d$$

where the last inequality holds for $d \geq 8$. □

4 Proof of the Upper Bound

In this section, we will establish the upper bound on the sample complexity for replicably PAC learning the hypothesis class from above in the realizable setting, which nearly matches our lower bound. For convenience, we restate the theorem here.

Theorem 1.3 (Replicable Learning Upper Bound)**.** *There exists an algorithm $\mathcal{A}$ such that for every instance $(\mathcal{X}, \mathcal{H}, \mathcal{D})$ shown to be hard in the proof of Theorem 1.2, and for every $\varepsilon, \delta, \rho \in (0,1)$, $\mathcal{A}$ is a ρ-replicable PAC learner on this instance with sample complexity*

$$n = \widetilde{O}\left(\frac{(\log|\mathcal{H}|)^{3/2}}{\rho\varepsilon}\right).$$

We recall that the input domain is $\mathcal{X} = [d] \times \mathbb{Z}_k$ and the distribution $\mathcal{D}$ is uniform over $\mathcal{X}$. Our hypothesis set $\mathcal{H}$ contains a hypothesis h_i for every d-tuple $i = (i_0, \ldots, i_{d-1}) \in \mathbb{Z}_k^d$. For a point $(a,b) \in \mathcal{X}$, we have $h_i((a,b)) = 1$ if $i_a \leq b < i_a + \lfloor k/2 \rfloor$ or if $b < i_a + \lfloor k/2 \rfloor < i_a$. Otherwise $h_i((a,b)) = 0$.

Let $h^\star = h_{i^\star} \in \mathcal{H}$ be an unknown target function and assume training samples are drawn by sampling $x_1, \ldots, x_n$ i.i.d. from $\mathcal{D}$ and constructing the training set $(x_1, h^\star(x_1)), \ldots, (x_n, h^\star(x_n))$. Let a desired accuracy $0 < \varepsilon < 1$, replicability parameter $0 < \rho < 1$ and failure probability $0 < \delta < 1$ be given. If $k = O(\varepsilon^{-1}\rho^{-1}\sqrt{d})$, then with $n = \widetilde{O}(dk) = \widetilde{O}(\varepsilon^{-1}\rho^{-1}d^{3/2})$ samples, we will see every point in the input domain with probability at least $1 - \delta$. We can thus output the unique $h^\star$. This is also replicable. So we assume $k \geq C\varepsilon^{-1}\rho^{-1}\sqrt{d}$ for sufficiently large constant $C > 0$.

Our learning algorithm is as follows: On samples $S = \{(x_i, h^\star(x_i))\}_{i=1}^n$ with $x_i = (a_i, B_i^z)$, partition S into d pieces $S_0, \ldots, S_{d-1}$ such that S_a contains all samples (a_i, B_i^z) with $a_i = a$. For each S_a, sort the samples by B_i^z and remove duplicates. Let b_a^S denote the value B_i^z of the point whose predecessor b_j (possibly with wrap around) has $h^\star((a, b_j)) = 0$ while $h^\star((a, B_i^z)) = 1$. If

some S_a contains no 0's or no 1's, we simply let $b_a^S = 0$. We have that b_a^S serves as an estimate of $i_a^\star$.

Now, as also introduced in the technical overview section, we define $\nu(a,b)$ for $a,b \in \mathbb{Z}_k$ as the distance between a and b with wrap-around, i.e. $\nu(a,b) = \min\{a-b, b-a\}$, where the minimum is determined when treating a,b as elements of $\mathbb{Z}_k$. That is, we apply $\bmod k$ before taking minimum. Let $0 < \beta < \varepsilon/4$ be a parameter to be determined. Shuffle all hypotheses in $\mathcal{H}$ (using the shared randomness) and return the first hypothesis $h_S := h_i$ satisfying $\sum_{a=0}^{d-1} \nu(b_a^S, i_a) \le \varepsilon k d/4$.

Correctness. First observe that h_S and the hypothesis h_{b^S} with $b^S = (b_0^S, \ldots, b_{d-1}^S)$ disagree in the predictions of at most $\varepsilon k d/2$ points. Since $\mathcal{D}$ is uniform over dk points, we have $|\operatorname{er}_{\mathcal{D}}(h_S) - \operatorname{er}_{\mathcal{D}}(h_{b^S})| \le \varepsilon/2$. Next we show that b_a^S is close to $i_a^\star$ for all a:

Lemma 4.1. *For any $0 < \alpha < 1/2$ and any $0 < \delta < 1$, if $n \ge \alpha^{-1} d \ln(2d/\delta)$ then it holds with probability at least $1-\delta$ that $\nu(b_a^S, i_a^\star) \le \alpha k$ for all $a \in [d]$.*

Proof. Fix a coordinate a. If S_a contains at least one point with label 0 and at least one point of the form (a, B_i^z) with $B_i^z \in \{i_a^\star, \ldots, i_a^\star + \alpha k\}$ then $\nu(b_a^S, i_a^\star) \le \alpha k$. The probability that S_a contains no points with label 0 is at most $(1 - 1/(2d))^n \le \exp(-n/(2d))$. The probability that S_a contains no points (a, B_i^z) with B_i^z of the above form is at most $(1 - (\alpha k + 1)/(dk))^n \le \exp(-\alpha n/d)$. For $n \ge \alpha^{-1} d \ln(2d/\delta)$ we can union bound over all d choices of a to conclude $\nu(b_a^S, i_a^\star) \le \alpha k$ for all a. □

If we pick $\alpha = \varepsilon/4$ and require $n = \Omega(\varepsilon^{-1} d \ln(d/\delta))$, we get with probability at least $1-\delta$ that $\sum_{a=0}^{d-1} \nu(b_a^S, i_a^\star) \le \varepsilon d k/4$. This implies $\operatorname{er}_{\mathcal{D}}(h_{b^S}) = \operatorname{er}_{\mathcal{D}}(h_{b^S}) - \operatorname{er}_{\mathcal{D}}(h^\star) \le \varepsilon/2$. Using the triangle inequality, we conclude $\operatorname{er}_{\mathcal{D}}(h_S) \le \varepsilon$ with probability $1-\delta$.

Replicability. For the replicability guarantee, let r characterize the random shuffle of all hypotheses in $\mathcal{H}$. For a fixed r, for two arbitrary samples S, S', we say that S (respectively S') *accepts* the i'th hypothesis if the i'th hypothesis in the shuffled $\mathcal{H}$ satisfies $\sum_{a=0}^{d-1} \nu(b_a^S, i_a) \le \varepsilon k d/4$. Abusing notation slightly, let the i^{th} hypothesis in $\mathcal{H}$ (when $\mathcal{H}$ is shuffled according to randomness r) be h_i, corresponding to the tuple $i = (i_0, \ldots, i_{d-1})$.

Let $A_{i,r}$ (respectively $A'_{i,r}$) denote the event that the i'th hypothesis (in the shuffle) is the *first* hypothesis in the shuffled $\mathcal{H}$ to be accepted by S (respectively S'). Similarly, let $B_{i,r}$ (respectively $B'_{i,r}$) denote the event that the i'th hypothesis in the shuffle is accepted by S (respectively S') (note that $A_{i,r} \subseteq B_{i,r}$ but not necessarily vice-versa). Let E denote the event that b^S and $b^{S'}$ satisfy $\nu(b_a^S, b_a^{S'}) \le \beta k/2$ for all a. Note that this event is defined solely from S and S' and is independent of the randomness r.

By relating both b^S and $b^{S'}$ to $i^\star$ and instantiating the triangle inequality, the correctness proof above (Lemma 4.1) gives us such b^S and $b^{S'}$ with probability $1 - \rho/4$ when $n \ge c\beta^{-1} d \ln(d/\rho)$ for large enough constant c.

Now consider two arbitrary samples S, S'. The event $\mathcal{A}(S;r) \ne \mathcal{A}(S';r)$ implies that $\exists i \in \{1, 2, \ldots, |\mathcal{H}|\}$ such that the event $A_{i,r}$ occurs but $B'_{i,r}$ does not, *or* the event $A'_{i,r}$ occurs but $B_{i,r}$

does not. Therefore, we have

$$\begin{aligned}
\Pr_{S,S',r}[\mathcal{A}(S,r) \neq \mathcal{A}(S',r)] &\leq \Pr_{S,S',r}[\mathcal{A}(S,r) \neq \mathcal{A}(S',r) \mid E] + \rho/4 \\
&\leq \Pr\big[\exists i \in \{1,\dots,|\mathcal{H}|\} : (A_{i,r} \wedge \neg B'_{i,r}) \vee (A'_{i,r} \wedge \neg B_{i,r}) \mid E\big] + \rho/4 \\
&\leq \sum_{i=1}^{|\mathcal{H}|} \Pr\big[A_{i,r} \wedge \neg B'_{i,r} \mid E\big] + \sum_{i=1}^{|\mathcal{H}|} \Pr\big[A'_{i,r} \wedge \neg B_{i,r} \mid E\big] + \rho/4 \\
&= \sum_{i=1}^{|\mathcal{H}|} 2\Pr\big[A_{i,r} \wedge \neg B'_{i,r} \mid E\big] + \rho/4 \\
&= \sum_{i=1}^{|\mathcal{H}|} 2 \Pr_{S,S',r}[A_{i,r} \mid E] \Pr[\neg B'_{i,r} \mid A_{i,r}, E] + \rho/4.
\end{aligned}$$

In the following, we will show that for any pair of samples S, S' satisfying the conditions in the event E (i.e. $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$ for all a), we have $\Pr_r[\neg B'_{i,r} \mid A_{i,r}] \leq 3\rho/8$ for all i. Combining this with $\sum_i \Pr[A_{i,r} \mid E] = 1$ gives the required replicability guarantee. So fix an arbitrary such pair of samples S, S'.

Conditioning on the event $A_{i,r}$ implies that $\mathcal{A}(S,r) = h_i$ and also that the distribution of i is uniform random among all i satisfying $\sum_{a=0}^{d-1} \nu(b_a^S, i_a) \leq \varepsilon kd/4$. For each coordinate $a \in [d]$, let $\sigma_a \in \{-1, 1\}$ be so that $i_a + \sigma_a \nu(b_a^S, i_a)k = b_a^S$ (picking σ_a uniformly in case of ties).

Let us now consider the distance $\sum_{a=0}^{d-1} \nu(b_a^{S'}, i_a)$. We want to show that this distance is no more than $\varepsilon kd/4$ with large probability, i.e. S' also accepts h_i.

Recall that for every coordinate a, we have that $i_a = b_a^S - \sigma_a \nu(b_a^S, i_a)$. It follows that if σ_a is such that $b_a^S - \sigma_a \nu(b_a^S, b_a^{S'}) = b_a^{S'}$, then

$$\nu(b_a^{S'}, i_a) = \max\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} - \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\}.$$

Otherwise, we have

$$\nu(b_a^{S'}, i_a) \leq \nu(b_a^S, b_a^{S'}) + \nu(b_a^S, i_a) = \max\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} + \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\}.$$

Since either of these cases happens with probability $1/2$ each, and using $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$ under the event E, we have

$$\begin{aligned}
&\Pr\left[\sum_{a=0}^{d-1} \nu(b_a^{S'}, i_a) > \varepsilon kd/4\right] \\
&\leq \Pr\left[\sum_{a=0}^{d-1} \max\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} + \tau_a \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} > \varepsilon kd/4\right] \\
&\leq \Pr\left[\sum_{a=0}^{d-1} \max\{\beta k/2, \nu(b_a^S, i_a)\} + \tau_a \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} > \varepsilon kd/4\right] \\
&= \Pr\left[\sum_{a=0}^{d-1} \tau_a \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} > \varepsilon kd/4 - \sum_{a=0}^{d-1} \max\{\beta k/2, \nu(b_a^S, i_a)\}\right].
\end{aligned}$$

where the τ_a's are uniformly random signs that may be sampled independently of everything.

Note that if we condition on everything but the signs τ_a, then by Hoeffding's inequality and the fact that $\nu(b_a^S, b_a^{S'}) \leq \beta k/2$, we have

$$\Pr\left[\sum_{a=0}^{d-1} \tau_a \min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\} > t\right] < \exp\left(\frac{-2t^2}{\sum_{a=0}^{d-1} 4\min\{\nu(b_a^S, b_a^{S'}), \nu(b_a^S, i_a)\}^2}\right) \leq \exp\left(\frac{-2t^2}{d\beta^2 k^2}\right).$$

We will pick $t = \varepsilon kd/4 - \sum_{a=0}^{d-1} \max\{\beta k/2, \nu(b_a^S, i_a)\}$ and therefore we set out to upper bound $\sum_{a=0}^{d-1} \max\{\beta k/2, \nu(b_a^S, i_a)\}$. Our goal is to show that $t \geq \beta k\sqrt{d\ln(2/\rho)}$ with high probability. When this is the case, it holds with probability at least $1 - \rho/4$ over the signs τ_a that $\sum_{a=0}^{d-1} \nu(b_a^{S'}, i_a) \leq \varepsilon kd/4$.

Let Δ denote the vector with coordinates $\nu(b_a^S, i_a)\sigma_a$ and observe that Δ is uniform random among all vectors in $\mathbb{Z}^d$ with $\|\Delta\|_\infty \leq k/2$ (due to wrap around) and $\|\Delta\|_1 \leq \varepsilon kd/4$. Our goal is to show that $\sum_{a=0}^{d-1} \max\{\beta k/2, |\Delta_a|\}$ is noticeably smaller than $\varepsilon kd/4$ with high probability. Here we first observe that

$$\sum_{a=0}^{d-1} \max\{\beta k/2, |\Delta_a|\} \leq \|\Delta\|_1 + |\{a : |\Delta_a| < \beta k/2\}| \cdot \beta k/2.$$

We bound the two terms below and arrive at the following two technical results.

Lemma 4.2. *Assume $k \geq 384\varepsilon^{-1}\rho^{-1}$. Then for any $\rho/4 \leq \gamma \leq 1/2$, it holds with probability at least $1 - \gamma$ that $\|\Delta\|_1 \leq \varepsilon dk/4 - \varepsilon\gamma k/96$.*

Lemma 4.3. *If $\beta \geq 2/k$ and $k \geq 384\varepsilon^{-1}\rho^{-1}$, then it holds with probability at least $1 - \rho/4$ that $|\{a : |\Delta_a| < \beta k/2\}| \leq 2304d\beta\varepsilon^{-1} + 3\ln(4/\rho)$.*

Invoking Lemma 4.2 with $\gamma = \rho/4$ and using Lemma 4.3, we get that with probability at least $1 - \rho/2$, we have

$$\sum_{a=0}^{d-1} \max\{\beta k/2, |\Delta_a|\} \leq \|\Delta\|_1 + |\{a : |\Delta_a| < \beta k/2\}| \cdot \beta k/2 \leq \varepsilon dk/4 - \varepsilon\rho k/(4 \cdot 96) + (\beta k/2)(2304d\beta\varepsilon^{-1} + 3\ln(4/\rho)).$$

Let us now set $\beta = c\min\{\varepsilon\rho/\sqrt{d\ln(2/\rho)}, \varepsilon\rho/\ln(4/\rho)\}$ for sufficiently small constant $c > 0$. For c small enough, we then have $\beta^2 kd\varepsilon^{-1}/1152 \leq \varepsilon\rho k/(16 \cdot 96)$ and $(3/2)\beta k\ln(4/\rho) \leq \varepsilon\rho k/(16 \cdot 96)$. This implies that

$$\sum_{a=0}^{d-1} \max\{\beta k/2, |\Delta_a|\} \leq \varepsilon dk/4 - \varepsilon\rho k/(8 \cdot 96).$$

Recall from above that we picked $t = \varepsilon kd/4 - \sum_{a=0}^{d-1} \max\{\beta k/2, |\Delta_a|\}$. We therefore have $t \geq \varepsilon\rho k/(8 \cdot 96)$. We needed this to satisfy $t \geq \beta k\sqrt{d\ln(2/\rho)}$. This is indeed satisfied whenever $\varepsilon\rho/(8 \cdot 96) \geq \beta\sqrt{d\ln(2/\rho)}$. Our choice of β satisfies this for c small enough.

From earlier, we had that the sample complexity was $n = O(\beta^{-1}d\ln(d/\rho))$. Inserting β finally gives a sample complexity of

$$n = \widetilde{O}\left(\varepsilon^{-1}\rho^{-1}d^{3/2}\right) = \widetilde{O}\left(\varepsilon^{-1}\rho^{-1}(\log|\mathcal{H}|)^{3/2}\right).$$

Bounding ℓ_1-Norm. In the following we prove Lemma 4.2. Define events $F_0, \ldots, F_d$, where F_z is the event that Δ has precisely z entries that are zero. Then for any t

$$\Pr[\|\Delta\|_1 \le t] = \sum_{z=0}^{d} \Pr[\|\Delta\|_1 \le t \mid F_z] \Pr[F_z]$$

We will first bound $\Pr[\|\Delta\|_1 \le t \mid F_z]$. To simplify this analysis, we will relate the distribution of $\|\Delta\|_1$ conditioned on F_z to another random variable Δ^z with a slightly simpler distribution. Concretely, let Δ^z be sampled uniformly among all vectors v in $\mathbb{Z}^d$ with $\|v\|_1 \le \varepsilon kd/4$ and precisely z entries that are 0. That is, we drop the requirement $|v_a| \le k/2$ compared to the distribution of Δ conditioned on F_z. We claim that

Lemma 4.4. *For any t, we have* $\Pr[\|\Delta\|_1 \le t \mid F_z] \ge \Pr[\|\Delta^z\|_1 \le t]$.

Proof. Let $p_i = \Pr[\|\Delta^z\|_\infty \le k/2 \mid \|\Delta^z\|_1 = i]$. The p_i's are monotonically decreasing in i. To see this, for any $i \le \varepsilon kd/4$, let $B_i^z := \{v : \|v\|_1 = i, \|v\|_0 = d - z\}$, where $\|v\|_0$ denotes the number of non-zero entries of v. Then, we have that $\Pr_{\Delta^z}[\Delta^z = v \mid \|\Delta^z\|_1 = i] = 1/|B_i^z|$. Let μ_i be the uniform distribution over B_i^z. Notice also that $|B_i^z| = 2^{d-z} \cdot \binom{d}{z} \cdot \binom{i-(d-z)+d-z-1}{d-z-1} = 2^{d-z} \cdot \binom{d}{z} \cdot \binom{i-1}{d-z-1}$: this is the number of ways we can pick z entries to be zero, a sum of $d-z$ integers, each of which is at least 1, to obtain the sum i, and then assigning signs to all the integers.

Now consider the randomized map $\rho : \mathbb{Z}^d \to \mathbb{Z}^d$ that acts as follows: on any input v, ρ first samples a non-zero coordinate j with probability $\frac{|v_j|-1}{\|v\|_1-(d-z)}$, and then outputs $\tilde{v}$, where,

$$\tilde{v}_{j'} = \begin{cases} v_{j'} & \forall j' \ne j \\ |v_j| - 1 \text{ with probability } 1/2, \text{ and } -(|v_j|-1) \text{ with probability } 1/2 & \text{for } j' = j. \end{cases}$$

That is, $\tilde{v}$ is equal to v at all coordinates other than j, where its absolute value is one smaller than $|v_j|$, so that $\|\tilde{v}\|_1 = \|v\|_1 - 1$. Then, if we first sample $v \sim \mu_i$, and then obtain $\tilde{v} = \rho(v)$, observe that the probability of obtaining a particular $\tilde{v} \in B_{i-1}^z$ is precisely the chance that we sampled v that satisfies $|v_j| = |\tilde{v}_j| + 1$ for some j with $v_j \ne 0$ and $v_{j'} = \tilde{v}_{j'}$ for all $j' \ne j$, and thereafter sampled the coordinate j and set $\tilde{v}_j$ as required: this is equal to

$$\sum_{j:\tilde{v}_j \ne 0} \frac{1}{|B_i^z|} \cdot \frac{|\tilde{v}_j| + 1 - 1}{i - (d-z)} \cdot \frac{1}{2} \cdot 2 = \frac{i-1}{|B_i^z|(i-(d-z))} = \frac{1}{|B_{i-1}^z|},$$

where we used the expression $|B_i^z| = 2^{d-z} \cdot \binom{d}{z} \cdot \binom{i-1}{d-z-1}$. We have thus argued that we can obtain a sample from μ_{i-1} by first sampling $v \sim \mu_i$, and then applying $\rho(v)$. We can then conclude

$$\begin{aligned} p_{i-1} &= \Pr[\|\Delta^z\|_\infty \le k/2 \mid \|\Delta^z\|_1 = i-1] = \Pr_{\Delta' \sim \mu_{i-1}}[\|\Delta'\|_\infty \le k/2] \\ &= \Pr_{\Delta' \sim \mu_i}[\|\rho(\Delta')\|_\infty \le k/2] \ge \Pr_{\Delta' \sim \mu_i}[\|\Delta'\|_\infty \le k/2] \\ &= \Pr[\|\Delta^z\|_\infty \le k/2 \mid \|\Delta^z\|_1 = i] = p_i, \end{aligned}$$

where the inequality above follows because $\|\Delta'\|_\infty \le k/2 \implies \|\rho(\Delta')\|_\infty \le k/2$. This establishes that the p_i' s are non-increasing.

Now observe that the distribution of Δ^z conditioned on $\|\Delta^z\|_\infty \le k/2$ equals the distribution of Δ conditioned on F_z. We thus have

$$\begin{aligned}
\Pr[\|\Delta\|_1 \le t \mid F_z] &= \sum_{i=0}^{t} \Pr[\|\Delta\|_1 = i \mid F_z] \\
&= \frac{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_i}{\Pr[\|\Delta^z\|_\infty \le k/2]} \\
&= \frac{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_i}{\sum_{i=0}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_i} \\
&\ge \frac{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_i}{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_i + \sum_{i=t+1}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_t} \\
&= 1 - \frac{\sum_{i=t+1}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_t}{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_i + \sum_{i=t+1}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_t} \\
&\ge 1 - \frac{\sum_{i=t+1}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_t}{\sum_{i=0}^{t} \Pr[\|\Delta^z\|_1 = i] p_t + \sum_{i=t+1}^{\varepsilon kd/4} \Pr[\|\Delta^z\|_1 = i] p_t} \\
&= 1 - \Pr[\|\Delta^z\|_1 > t] \\
&= \Pr[\|\Delta^z\|_1 \le t].
\end{aligned}$$

□

In light of Lemma 4.4 we set out to show that $\|\Delta^z\|_1$ is somewhat smaller than $\varepsilon dk/4$ with high probability.

We see for $t \ge 2d$ that

$$1 \le \frac{B^z_{t+1}}{B^z_t} = \frac{\binom{t+1}{d-z-1}}{\binom{t}{d-z-1}} = \frac{t+1}{t+1-(d-z-1)} = 1 + \frac{d-z-1}{t+2-(d-z)} \le 1 + \frac{2d}{t}.$$

Here the last inequality follows from $d - z \le d \le t/2$ (assuming $t \ge 2d$). Assume now that $k \ge 384\varepsilon^{-1}\rho^{-1}$. For any γ satisfying $\rho/4 \le \gamma \le 1/2$, define $q = \varepsilon\gamma k/96$ (which is at least 1 by our requirement on k). Note that this choice of q also satisfies $\varepsilon kd/4 - 2q\gamma^{-1} \ge \varepsilon kd/8 \ge 2d$. We now have that the probability that $\|\Delta^z\|_1 > \varepsilon kd/4 - q$ is at most

$$\begin{aligned}
\frac{\sum_{t=\varepsilon kd/4-q+1}^{\varepsilon kd/4} B^z_t}{\sum_{t=0}^{\varepsilon kd/4} B^z_t} &\le \frac{\sum_{t=\varepsilon kd/4-q+1}^{\varepsilon kd/4} B^z_t}{\sum_{t=\varepsilon kd/4-2q\gamma^{-1}}^{\varepsilon kd/4} B^z_t} \\
&\le \frac{q \cdot B_{\varepsilon kd/4-q+1}(1+2d/(\varepsilon kd/8))^q}{(2\gamma^{-1}q) \cdot B_{\varepsilon kd/4-q+1}(1+2d/(\varepsilon kd/8))^{-2\gamma^{-1}q}} \\
&\le \frac{(1+2d/(\varepsilon kd/8))^{q+2\gamma^{-1}q}}{2\gamma^{-1}} \\
&\le \frac{\exp(48\gamma^{-1}q/(\varepsilon k))}{2\gamma^{-1}} \\
&\le e^{1/2}\gamma/2 \\
&\le \gamma.
\end{aligned}$$

We thus have with probability at least $1-\gamma$ that $\|\Delta^z\|_1 \leq \varepsilon dk/4 - \varepsilon\gamma k/96$. Combined with Lemma 4.4 this also implies

$$\begin{aligned}\Pr[\|\Delta\|_1 \leq \varepsilon dk/4 - \varepsilon\gamma k/96] &= \sum_{z=0}^{d} \Pr[\|\Delta\|_1 \leq \varepsilon dk/4 - \varepsilon\gamma k/96 \mid F_z] \Pr[F_z] \\ &\geq \Pr[\|\Delta^z\|_1 \leq \varepsilon dk/4 - \varepsilon\gamma k/96] \Pr[F_z] \\ &\geq 1-\gamma.\end{aligned}$$

This completes the proof of Lemma 4.2.

Bounding Number of Small Coordinates. We next set out to prove that $|\{a : |\Delta_a| < \beta k/2\}|$ is small with high probability.

Proof of Lemma 4.3. Let X_a denote an indicator random variable for the event $|\Delta_a| \leq \beta k/2$. We now bound $\Pr[X_a = 1]$. For this, consider the set S of all v with $\|v\|_\infty \leq k/2$ and $\|v\|_1 \leq \varepsilon kd/4$. Let S_a be the set of all v with $\|v\|_\infty \leq k/2, \|v\|_1 \leq \varepsilon kd/4$ and $|v_a| \leq \beta k/2$. By definition, we have $|S_a| = \Pr[X_a = 1]|S|$.

Now observe that the events $|v_a| \leq \beta k/2$ and $\|v\|_1 \leq \varepsilon kd/4 - q$ for $q \geq 1$ are positively correlated. Let us now pick $q = \varepsilon k/192$. Then by Lemma 4.2, we have $\Pr[\|\Delta\|_1 \leq \varepsilon dk/4 - q] \geq 1/2$. By the positive correlation, this further implies that the subset $S_a^\star \subseteq S_a$ of vectors v in S_a also satisfying $\|v\|_1 \leq \varepsilon kd/4 - q$ has $|S_a^\star| \geq |S_a|/2 = \Pr[X_a = 1]|S|/2$.

We now relate $|S_a^\star|$ and $|S|$ by considering a bipartite graph where the left side has a node for each $v \in S_a^\star$ and the right side has a node for each $v \in S$. For each $v \in S_a^\star$, add an edge to every $w \in S$ so that $v_j = w_j$ for all $j \neq a$. We argue that every node on the left side has large degree, and every node on the right side has small degree. This eventually bounds the ratio between the number of nodes on the two sides.

So consider a node on the left side, corresponding to a fixed $v \in S_a^\star$. By definition of $S_a^\star$, we have $\|v\|_1 \leq \varepsilon kd/4 - q$ and $|v_a| \leq \beta k/2$. This implies that every vector w with $|w_a| \leq q - \beta k/2$ and $w_j = v_j$ for $j \neq a$ has $\|w\|_1 \leq \varepsilon kd/4$ and thus any such w is in S. The degree of v in $S_a^\star$ is hence at least $2(q - \beta k/2)$.

Consider next a node on the right side, corresponding to a fixed $w \in S$. For any integer z with $|z| \leq \beta k/2$, there is at most one $v \in S_a^\star$ satisfying $w_j = v_j$ for every $j \neq a$ and $v_a = z$. Thus the degree of w is at most $\beta k + 1$.

If E denotes the set of edges in the bipartite graph, we thus have $|E| \geq |S_a^\star|(2q - \beta k)$ and $|E| \leq |S|(\beta k + 1)$. We therefore have $|S_a^\star| \leq |S|(\beta k + 1)/(2q - \beta k)$. Combining this with the inequality $|S_a^\star| \geq \Pr[X_a = 1]|S|/2$ finally yields $\Pr[X_a = 1] \leq 2(\beta k + 1)/(2q - \beta k)$.

If we constrain $\beta k \leq q = \varepsilon k/192$, this probability is at most $2(\beta k + 1)/q = 384(\beta k + 1)\varepsilon^{-1}/k$. If we further require $\beta k \geq 1$, this is again upper bounded by $768\beta\varepsilon^{-1}$. Since the X_a are negatively correlated, we have by a Chernoff bound that $\Pr[\sum_a X_a > 2304 d\beta\varepsilon^{-1} + 3\ln(4/\rho)] \leq \rho/4$. □

A Deferred Proofs

In this section, we will prove all the lemmas that were skipped in the main text. For convenience, we restate the lemmas.

Lemma A.1 (Independence and linear independence)**.** *Let k be prime power, $d \in \mathbb{N}$ and consider the finite field with k elements $\mathbb{F}_k$. Then let $y_1, \ldots, y_r$ be vectors from $\mathbb{F}_k^d$ and let v be uniform on $\mathbb{F}_k^d$. If $y_1 \notin \mathrm{Span}\{y_2, \ldots, y_r\}$, then $\langle y_1, v\rangle$ is independent of $(\langle y_2, v\rangle, \ldots, \langle y_r, v\rangle)$.*

Proof. For a finite set S, let $\mathcal{U}(S)$ denote the uniform distribution on S. Also, for a matrix M, let $r(M)$ be the rowspace of M. Let A denote the matrix with $y_1, \ldots, y_r$ as rows and B the matrix with $y_2, \ldots, y_r$ as rows. Note then that $Av = (\langle y_1, v\rangle, \ldots, \langle y_r, v\rangle)$. We first show that Av is uniform on $r(A)$, the row space of A. Let $x, y \in r(A)$. Then there is $w \in \mathbb{F}_k^d$ such that $Aw = x - y$, and hence

$$\Pr[Av = x] = \Pr[A(v + w) = x] = \Pr[Av = x - Aw] = \Pr[Av = y],$$

which shows the first claim, using that v is uniform. Now, by linear independence, we must have (for example by a dimension-argument) that $r(A) = \mathbb{F}_k \times r(B)$. Thus, we have the joint distribution:

$$Av = (\langle y_1, v\rangle, Bv) \sim \mathcal{U}(\mathbb{F}_k \times r(B)) = \mathcal{U}(\mathbb{F}_k) \otimes \mathcal{U}(r(B)).$$

In particular $\langle y_1, v\rangle$ is independent of $Bv = (\langle y_2, v\rangle, \ldots, \langle y_r, v\rangle)$. □

Lemma A.2. *Let $x > 0$ and $\alpha \geq e$ be positive reals. Then $x \geq \alpha/\log(x)$ implies $x \geq \alpha/\log(\alpha)$.*

Proof. Assume for sake of contradiction that $x < \alpha/\log\alpha$. Then

$$x \geq \alpha/\log x > \frac{\alpha}{\log(\alpha/\log(\alpha))} \geq \alpha/\log\alpha$$

giving us the desired contradiction. □

Lemma 3.8. *Given values $x_0, \ldots, x_d$ and $y_0, \ldots, y_d$ with the following properties:*

- *(Non-negative). $x_0, \ldots, x_d \geq 0$ and $y_0, \ldots, y_d \geq 0$,*
- *(Equal sum). $\sum_{i=0}^{d} x_i = \sum_{j=0}^{d} y_j$,*
- *(Dominating). For all $k \in \{0, \ldots, d\}$, we have $\sum_{i=0}^{k} x_i \leq \sum_{j=0}^{k} y_j$.*

Then there exist values $p_{i,j}$ where $0 \leq j \leq i \leq d$, such that

1. *for all $0 \leq j \leq i \leq d$, we have $0 \leq p_{i,j} \leq 1$,*
2. *for all $i \in \{0, \ldots, d\}$, we have $\sum_{j=0}^{i} p_{i,j} = 1$,*
3. *for all $j \in \{0, \ldots, d\}$, we have $\sum_{i=j}^{d} x_i \cdot p_{i,j} = y_j$.*

Proof. The proof goes by induction in d. For the base case $d = 0$, one can just set $p_{0,0} = 1$; the first and second properties are immediately satisfied, while the third property follows from the fact that $x_0 = y_0$.

For the inductive step, we can assume, that one can always find such values $p_{i,j}$ for instances of size $d - 1$. Now assume we are given an instance $x_0, \ldots, x_d$ and $y_0, \ldots, y_d$ of size d. Remark that, if $x_d = 0$, then it follows from the dominating and equal sum property that $y_d = 0$. Therefore, we can just pick $p_{d,j} = 1/d$ for all $j \in \{0, \ldots, d\}$ and then solve the problem for $x_1, \ldots, x_{d-1}$ and $y_1, \ldots, y_{d-1}$ using the induction hypothesis. We can verify that this satisfies all the three properties required of the $p_{i,j}$ values.

We now move on to the case of $x_d \neq 0$. Here, we will pick the values $p_{d,k}$ in a greedy fashion, where $p_{d,k}$ is chosen in terms of $p_{d,k+1}, \ldots, p_{d,d}$ as seen below:

$$p_{d,d} = \frac{y_d}{x_d}, \qquad \forall k \in \{0, \ldots, d-1\} : p_{d,k} = \min\left\{\frac{y_k}{x_d}, 1 - \sum_{j=k+1}^{d} p_{d,j}\right\}.$$

First, we verify that the $p_{d,k}$ values satisfy Property 1 and 2 above. Note that since $\sum_{i=0}^{d-1} x_i \leq \sum_{j=0}^{d-1} y_j$ and $\sum_{i=0}^{d} x_i = \sum_{j=0}^{d} y_j$, it must be the case that $y_d \leq x_d$, which means that $0 \leq p_{d,d} \leq 1$. For any other $k \neq d$, observe that by definition, $p_{d,k} \leq 1 - \sum_{j=k+1}^{d} p_{d,j} \implies \sum_{j=k}^{d} p_{d,j} \leq 1$. Together, we have that $\sum_{j=k}^{d} p_{d,k} \leq 1$ for every $k \in \{0, \ldots, d\}$. This means that $0 \leq p_{d,k} \leq 1$ for all k.

For Property 2, we claim that there must be some $k \in \{0, \ldots, d-1\}$ for which $p_{d,k} = 1 - \sum_{j=k+1}^{d} p_{d,j}$. Otherwise, if $p_{d,k} = y_k/x_d$ for all $k = \{1, \ldots, d\}$, then

$$\sum_{j=0}^{d} y_j \geq \sum_{i=0}^{d} x_i \geq x_d \implies \frac{y_0}{x_d} \geq 1 - \sum_{j=1}^{d} \frac{y_j}{x_d} = 1 - \sum_{j=1}^{d} p_{d,j},$$

meaning that $p_{d,0} = \frac{y_0}{x_d} = 1 - \sum_{j=1}^{d} p_{d,j}$. So we know that there is some value $k \in \{0, \ldots d-1\}$ where $p_{d,k} = 1 - \sum_{j=k+1}^{d} p_{d,j}$. Then it must be the case that $p_{d,l} = 0$ for all $l < k$, since then $1 - \sum_{j=k}^{d} p_{d,j} = 0$. We can therefore compute the sum in Property 2 as

$$\sum_{j=0}^{d} p_{d,j} = \sum_{j=k}^{d} p_{d,j} = \sum_{j=k+1}^{d} p_{d,j} + \left(1 - \sum_{j=k+1}^{d} p_{d,j}\right) = 1.$$

Now, to choose the rest of the values $p_{i,j}$, we construct a smaller instance of the problem with values $x' := x'_0, \ldots x'_{d-1}$ and $y' := y'_0, \ldots y'_{d-1}$. In this instance, we let $x'_i = x_i$ and $y'_j = y_j - x_d \cdot p_{d,j}$. Remark that $x'_i = x_i \geq 0$ and $y'_j = y_j - x_d \cdot p_{d,j} \geq y_j - x_d \cdot \frac{y_j}{x_d} = 0$, so non-negativity still holds. Also, they still have equal sum since

$$\begin{aligned}\sum_{j=0}^{d-1} y'_j = \sum_{j=0}^{d-1} (y_j - x_d \cdot p_{d,j}) &= -y_d + x_d \cdot p_{d,d} + \sum_{j=0}^{d} y_j - x_d \sum_{j=0}^{d} p_{d,j} \\ &= -y_d + x_d \cdot \frac{y_d}{x_d} + \sum_{i=0}^{d} x_i - x_d = -x_d + \sum_{i=0}^{d} x_i = \sum_{i=0}^{d-1} x'_i.\end{aligned}$$

Finally, we show that x' and y' satisfy the dominating property. For any $k \in \{0, \ldots, d-1\}$ we will consider 2 cases. The first case is that $\sum_{j=0}^{k} p_{d,j} = 0$. Then, we have that

$$\sum_{i=0}^{k} x'_i = \sum_{i=0}^{k} x_i \leq \sum_{j=0}^{k} y_j = \sum_{j=0}^{k} (y'_j + x_d \cdot p_{d,j}) = \sum_{j=0}^{k} y'_j.$$

Now, for the other case $\sum_{j=0}^{k} p_{d,j} \neq 0$, we realize that this must imply that $p_{d,j} = \frac{y_j}{x_d}$ for all $j > k$, since otherwise, the sum would have been 0. This also means that $y'_j = y_j - x_d \cdot p_{d,j} = 0$ for $j > k$. Therefore, we get that

$$\sum_{i=0}^{k} x'_i \leq \sum_{i=0}^{d-1} x'_i = \sum_{j=0}^{d-1} y'_j = \sum_{i=0}^{k} y'_j.$$

We thus conclude that this smaller instance satisfies all three conditions. The induction hypothesis therefore tells us that there exist $p'_{i,j}$ values satisfying Properties 1, 2 and 3 for x', y'. We will use these values in the problem for x, y. That is, we choose $p_{i,j} = p'_{i,j}$ for all $0 \leq j \leq i \leq d-1$. The

$p_{d,k}$ values were already specified above. It remains to show that Properties 1, 2 and 3 are satisfied for x, y, p.

Properties 1 and 2 follow directly from the induction hypothesis, and the justification for the $p_{d,k}$ values given above.

For Property 3, we can see that for any $j \in [d-1]$, by rewriting the sum

$$\sum_{i=j}^{d} x_i \cdot p_{i,j} = x_d \cdot p_{d,j} + \sum_{i=j}^{d-1} x'_i \cdot p'_{i,j} = x_d \cdot p_{d,j} + y'_j = x_d \cdot p_{d,j} + y_j - x_d \cdot p_{d,j} = y_j.$$

And finally, for $j = d$, we also have Property 3 directly from the definition of $p_{d,d}$. □

Acknowledgements

CP was supported by Gregory Valiant's and Moses Charikar's Simons Investigator Awards, and a Google PhD Fellowship. KGL, MEM and CS are supported by the European Union (ERC, TUCLA, 101125203). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Monya Baker. Reproducibility crisis. *nature*, 533(26):353–66, 2016.

[2] Philip Ball. Is ai leading to a reproducibility crisis in science? *Nature*, 624(7990):22–25, 2023.

[3] Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 818–831, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519973. URL https://doi.org/10.1145/3519935.3519973.

[4] Peter Dixon, A. Pavan, Jason Vander Woude, and N. V. Vinodchandran. List and certificate complexities in replicable learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 30784–30806. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/61d0a96d4a73b626367310b3ad32579d-Paper-Conference.pdf.

[5] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[6] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3): 247–261, 1989. ISSN 0890-5401. doi: https://doi.org/10.1016/0890-5401(89)90002-3. URL https://www.sciencedirect.com/science/article/pii/0890540189900023.

[7] Peter Auer. Learning nested differences in the presence of malicious noise. *Theoretical Computer Science*, 185(1):159–175, 1997.

[8] Peter Auer and Ronald Ortner. A new pac bound for intersection-closed concept classes. *Machine Learning*, 66(2):151–163, 2007.

[9] Hans U. Simon. An almost optimal pac algorithm. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1552–1563, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Simon15a.html.

[10] Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

[11] Kasper Green Larsen. Bagging is an optimal pac learner. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 450–468. PMLR, 2023.

[12] Ishaq Aden-Ali, Mikael Møller Høandgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 22–45. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/aden-ali24a.html.

[13] Mikael Høgsgaard Møller. Efficient optimal pac learning. In Gautam Kamath and Po-Ling Loh, editors, *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 578–580. PMLR, 24–27 Feb 2025. URL https://proceedings.mlr.press/v272/hogsgaard-moller25a.html.

[14] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[15] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.

[16] Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023.

[17] Max Hopkins and Shay Moran. The role of randomness in stability. *arXiv preprint arXiv:2502.08007*, 2025.

[18] Hossein Esfandiari, Amin Karbasi, Vahab Mirrokni, Grigoris Velegkas, and Felix Zhou. Replicable clustering. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[19] Alkis Kalavasis, Amin Karbasi, Kasper Green Larsen, Grigoris Velegkas, and Felix Zhou. Replicable learning of large-margin halfspaces. *arXiv preprint arXiv:2402.13857*, 2024.

[20] Saba Ahmadi, Siddharth Bhandari, and Avrim Blum. Replicable online learning. *arXiv preprint arXiv:2411.13730*, 2024.

[21] Max Hopkins, Russell Impagliazzo, Daniel Kane, Sihan Liu, and Christopher Ye. Replicability in high dimensional statistics. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1–8. IEEE, 2024.

[22] Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:74702–74735, 2023.

[23] Ilias Diakonikolas, Jingyi Gao, Daniel Kane, Sihan Liu, and Christopher Ye. Replicable distribution testing. *arXiv preprint arXiv:2507.02814*, 2025.

[24] Zachary Chase, Shay Moran, and Amir Yehudayoff. Replicability and stability in learning. *arXiv preprint arXiv:2304.03757*, 2023.

[25] Max Hopkins, Russell Impagliazzo, and Christopher Ye. Approximate replicability in learning. *arXiv preprint arXiv:2510.20200*, 2025.

[26] Alexander Golovnev, Gleb Posobin, Oded Regev, and Omri Weinstein. Polynomial data structure lower bounds in the group model. *SIAM Journal on Computing*, 53(6):FOCS20–74, 2022.

[27] Cheuk Ting Li. Efficient approximate minimum entropy coupling of multiple probability distributions. *IEEE Transactions on Information Theory*, 67(8):5259–5268, 2021. doi: 10.1109/TIT.2021.3076986.

[28] Roberto Bruno and Ugo Vaccaro. A note on equivalent conditions for majorization. *AIMS Mathematics*, 9(4):8641–8660, 2024. ISSN 2473-6988. doi: 10.3934/math.2024419. URL `https://www.aimspress.com/article/doi/10.3934/math.2024419`.

[29] Luca Trevisan. CS359G: Graph Partitioning and Expanders Lecture 6. `https://theory.stanford.edu/~trevisan/cs359g/lecture06.pdf`, 2011.

[30] Bogdan Nica. *A brief introduction to spectral graph theory*, volume 3. European Mathematical Society Zürich, 2018.