Distribution-free Testing for Halfspaces (Almost) Requires PAC Learning

Xi Chen* Shyamal Patel†

Abstract

It is well known that halfspaces over $\mathbb{R}^n$ and $\{0,1\}^n$ are PAC-learnable with $\Theta(n)$ samples. Recently Blais et al. [4] showed that even the easier task of distribution-free sample-based testing requires $\Omega(n/\log n)$ samples for halfspaces.

In this work we study the distribution-free testing of halfspaces *with queries*, for which we show that the complexity remains to be $\tilde{\Omega}(n)$. Indeed we prove the following stronger tradeoff result: any distribution-free testing algorithm for halfspaces over $\{0,1\}^n$ that receives k samples must make $\exp(\tilde{\Omega}(\sqrt{n/k}))$ queries on the input function, when k satisfies $n^{.99} \leq k \leq O(n/\log^3 n)$. For halfspaces over $\mathbb{R}^n$ we show that any algorithm that makes a finite number of queries must draw $\Omega(n/\log n)$ many samples.

1 Introduction

The fundamental theorem of Statistical Learning [21] shows that the VC dimension of a class essentially captures the number of samples needed for its PAC learning. This implies a tight bound of $\Theta(n)$ for the PAC learning of halfspaces or LTFs (i.e., $f(x) = \mathrm{sgn}(w^T x - b)$ for some $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ over either $x \in \mathbb{R}^n$ or the hypercube $\{\pm 1\}^n$). However, given the pervasiveness of linear models in machine learning, it is natural to ask what one can achieve with fewer than $\Theta(n)$ samples.

An avenue for such investigation is to consider the *testing* of halfspaces under the *distribution-free* model, where an algorithm only needs to solve the easier testing task and can make adaptive queries in addition to drawing samples. Formally, the goal of a testing algorithm is to determine whether an unknown function f is an LTF or far from LTFs with respect to an unknown distribution $\mathcal{D}$ (i.e., $\Pr_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x}) \neq g(\mathbf{x})] \geq \epsilon$ for any LTF g), given query access to f and sampling access to $\mathcal{D}$. The complexity of an algorithm is measured by the number of samples it draws plus the number of queries it makes. Inspired by the PAC learning model [23], the distribution-free testing model was first introduced by Goldreich, Goldwasser and Ron [10] and has been studied extensively [1, 14, 11, 15, 16, 8].

As observed by [10], any proper PAC learning algorithm can be used for distribution-free property testing. The question is whether halfspaces allow more efficient testing algorithms than learning. On the lower bound side, Glasner and Servedio proved that any distribution-free tester for halfspaces must have complexity $\tilde{\Omega}(n^{1/5})$ [11]. This was later improved by Chen and Xie to $\tilde{\Omega}(n^{1/3})$ [7]. On the other hand, recently Blais et al. showed that any sample-based tester must request $\Omega(n/\log n)$ samples [4], by developing a new notion of "lower VC" dimension to characterize the sample complexity of testing problems. Before their work, an $\Omega(n)$ lower bound was obtained by

*Columbia University. Supported by NSF grants CCF-1703925, IIS-1838154, CCF-2106429 and CCF-2107187.

†Columbia University. Supported by NSF grant CCF-1714818 and a NSF Graduate Research Fellowship.

Epstein and Silwal for one-sided sample-based testers [9]. In summary, the state-of-the-art on the testing of halfspaces under the distribution-free model is in sharp contrast with standard testing model (under uniform distribution over $\{\pm 1\}^n$ or Gaussian over $\mathbb{R}^n$), where testing is known to be significantly easier than learning (see discussion after main results).

Our contribution. We come close to resolving the testing versus learning question for halfspaces by showing that distribution-free testing is (almost) as hard as learning. We remark that all the lower bounds we prove apply to adaptive algorithms with two-sided errors.

For $\mathbb{R}^n$, we show that any algorithm that makes a finite number of queries must draw nearly the same number of samples as learning:

THEOREM 1.1. *Suppose* ALG *is a distribution-free testing algorithm for halfspaces over* $\mathbb{R}^n$ *that makes a finite number of queries and, given any* $(f, \mathcal{D})$*, determines with probability at least* $2/3$ *whether* f *is an LTF or is* $\Omega(1)$*-far from LTFs with respect to* $\mathcal{D}$*. Then* ALG *must use* $\Omega(n/\log n)$ *samples.*

For $\{\pm 1\}^n$, we show that any algorithm that draws $k \leq n/\mathsf{polylog}(n)$ samples can make up the difference by only paying a high cost of $\exp(\tilde{\Omega}(\sqrt{n/k}))$ queries:

THEOREM 1.2. *Suppose* ALG *is a distribution-free testing algorithm for halfspaces over* $\{\pm 1\}^n$ *that, given any* $(f, \mathcal{D})$*, determines with probability at least* $2/3$ *whether* f *is an LTF or is* $\Omega(1)$*-far from LTFs with respect to* $\mathcal{D}$*. If* ALG *draws at most* k *samples for some* $k : n^{.99} \leq k \leq O(n/\log^3 n)$*, then it must make at least* $\exp(\tilde{\Omega}(\sqrt{n/k}))$ *queries.*[1]

Our results highlight the difference of power between *samples* and *queries*, giving an exponential tradeoff for the task of testing halfspaces. We believe that it is an interesting direction to understand whether similar tradeoff phenomenons occur in other distribution-free testing problems.

Our results also lead to a strong separation for testing halfspaces between the standard testing model and the distribution-free model. For the standard model, Matulef et al. [19] showed that halfspaces can be tested under the uniform distribution over $\{\pm 1\}^n$ or Gaussian over $\mathbb{R}^n$ with only $\mathrm{poly}(1/\epsilon)$ queries. For sample-based testing under the Gaussian distribution over $\mathbb{R}^n$, Balcan et al. [2] showed that $\tilde{O}(\sqrt{n})$ samples suffice. This was later extended by Harms [13] to show that $\tilde{O}(\sqrt{n})$ samples suffice for any unknown rotationally-invariant distribution $\mathcal{D}$ over $\mathbb{R}^n$.

In addition to testing halfspaces, we believe that our techniques can be straightforwardly applied to prove $\tilde{\Omega}(\ell n)$ lower bounds for the distribution-free testing of the intersection of ℓ halfspaces, which would again match the complexity of learning up to logarithmic factors. We plan to include a proof in the full version of the paper.

1.1 Proof Overview We start with a quick review of the lower bound of [4] on sample-based distribution-free testing of halfspaces. A crucial idea in their lower bound proof is to embed the following *support size distinction* problem [24, 25, 26] (SSD for short):

DEFINITION 1.1. (SUPPORT SIZE DISTINCTION PROBLEM) *Fix any* $n \geq 1$ *and* $0 < \alpha < \beta < 1$*. Let* p *be a hidden distribution supported on* $[n]$ *such that either* $|\mathrm{supp}(p)| \leq \alpha n$ *or* $|\mathrm{supp}(p)| \geq \beta n$*, and in both cases* p *satisfies* $p(i) \geq 1/n$ *for all* $i \in \mathrm{supp}(p)$*. We use* $\mathsf{SSD}(n, \alpha, \beta)$ *to denote the*

[1] We note that the theorem applies to $k < n^{.99}$ as well, though the bound is only exponential in $\tilde{\Omega}(n^{.005})$ instead of $\tilde{\Omega}(\sqrt{n/k})$. We have made no effort to optimize the constant 0.99 in the exponent.

smallest number of samples needed for an algorithm to distinguish the two cases with probability at least 2/3.

The theorem below follows from the proof of [26], but only explicitly appears in [4]:

THEOREM 1.3. ([26]) *For any* $\delta = \Omega\left(\frac{\sqrt{\log(n)}}{n^{1/4}}\right)$ *and* $\alpha, 1-\beta \geq \delta$, *we have*

$$\mathsf{SSD}(n, \alpha, \beta) = \Omega\left(\frac{n\delta^2}{\log n}\right).$$

For the reduction, they make two observations (see proofs in Appendix A for completeness):

LEMMA 1.1. (FOLKLORE) *If* $v^1, ..., v^k \in \mathbb{R}^n$ *are affinely independent, then any function* $f : \{v^1, ..., v^k\} \to \{\pm 1\}$ *is consistent with an LTF.*

LEMMA 1.2. *Let* $v^1, ..., v^k$ *be a set of* $k \geq 100(n+1)$ *points and* $\mathcal{D}$ *be a distribution over them such that every* v^i *has probability* $\Omega(1/k)$. *Then a function* $\boldsymbol{f} : \{v^1, ..., v^k\} \to \{\pm 1\}$ *drawn uniformly at random is* $\Omega(1)$*-far from LTFs with respect to* $\mathcal{D}$ *with probability* $1 - o_n(1)$ *(over* $\boldsymbol{f}$*).*

The lower bound of [4] for $\mathbb{R}^n$ proceeds as follows. Assume for a contradiction that there is a distribution-free sample-based algorithm ALG for testing halfspaces with $o(n/\log n)$ samples and success probability 0.9 [2]. Then a player can use it to solve $\mathsf{SSD}(400n, 1/400, 1/2)$ with $o(n/\log n)$ samples as follows.

Let $m = 400n$ and p be the unknown distribution over $[m]$. The player starts by picking (1) a set $Q = \{q^1, \ldots, q^m\}$ of m points in $\mathbb{R}^n$ such that every n-subset of Q is affinely independent and (2) a random coloring $\boldsymbol{\psi} : [m] \to \{\pm 1\}$ (Q and $\boldsymbol{\psi}$ can both be given to the ALG). Together with the unknown p, they define the following instance $(f, \mathcal{D})$ for distribution-free testing of halfspaces:

1. The distribution $\mathcal{D}$ over Q satisfies $\mathcal{D}(q^i) = p(i)$ for each $i \in [m]$.
2. Given that ALG is sample-based, it suffices to describe f over Q. If p has large support, then $f(q^i) = \boldsymbol{\psi}(i)$ for all $i \in [m]$, which by Lemma 1.2 is far from LTFs with high probability; if p has small support, then f is set to be an LTF that is consistent with $\boldsymbol{\psi}(i)$ at each q^i with $i \in \operatorname{supp}(p)$. The latter is always possible because we have $|\operatorname{supp}(p)| \leq n$ in this case.

While the player does not know p and thus, knows neither f nor $\mathcal{D}$, she can draw samples from p to simulate samples from $\mathcal{D}$. Given that ALG succeeds in determining which case it is with probability at least 0.9, the player solves the SSD problem with probability $0.9 - o_n(1)$ using $o(n/\log n)$ samples (where the loss of $o_n(1)$ is due to the random choice of $\boldsymbol{\psi}$), a contradiction with Theorem 1.3.

Next we describe ideas needed to modify the strategy of [4] to obtain our lower bound for distribution-free LTF testing with queries for $\mathbb{R}^n$. Since our bound will apply to any finite number of queries, we may assume without loss of generality that ALG is nonadaptive. Let's use the version of $\mathsf{SSD}(400n, 1/1600, 1/2)$ with a slightly smaller α. The first thing is that we need to make sure there is no way for ALG to query points in the set Q in which we hide the unknown distribution p.

To this end, the player starts by drawing a random vector $\mathbf{r} \in S^{n-1}$, where S^{n-1} denotes the unit $(n-1)$-sphere in $\mathbb{R}^n$, and uses it to define a random hyperplane $\mathbf{h} = \{x \in \mathbb{R}^n : \mathbf{r}^T x = 0\}$. The

[2] While our lower bounds are for algorithms with error 2/3, it suffices to prove lower bounds against algorithms with error 0.9 as we can always amplify the success probability.

player then draws (1) a sequence of m points $\mathbf{q}^1, \ldots, \mathbf{q}^m$ independently and uniformly at random from $\mathbf{h} \cap S^{n-1}$ to form the new set $\mathbf{Q} = \{\mathbf{q}^1, \ldots, \mathbf{q}^m\}$ (note that these m points are distinct with probability 1) and (2) a coloring $\boldsymbol{\psi} : \mathbf{Q} \to \{\pm 1\}$ uniformly at random. The hidden distribution p over $[m]$ defines our distribution $\mathcal{D}$ over $\mathbf{Q}$ with $\mathcal{D}(\mathbf{q}^i) = p(i)$ for each $i \in [m]$. Before defining the function f, let's consider what ALG is able to do after receiving a sequence of $k = o(n/\log n)$ samples $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ drawn from $\mathcal{D}$. On one hand, ALG can definitely query points in the linear span of $\mathbf{X}$ so we need to make sure points there do not reveal any useful information. On the other hand, it is unlikely for ALG to query any point that is on $\mathbf{h}$ but is not in the span of $\mathbf{X}$, given the randomness of $\mathbf{r}$.

These observations inspire the following construction. First, given the coloring $\boldsymbol{\psi} : \mathbf{Q} \to \{\pm 1\}$, we show how to extend it to a function $g : \mathbb{R}^n \to \{\pm 1\}$ such that g agrees with $\boldsymbol{\psi}$ on $\mathbf{Q}$ and for any subset T of $\mathbf{Q}$ with $|T| \le n/4$, g agrees with an LTF $\operatorname{sgn}(w^T x)$ at every point $x \in \operatorname{span}(T)$ for some w depending only on T and $\boldsymbol{\psi}$ over T. The existence of such an extension follows from linear algebra arguments which we present in Lemma 3.2. Finally we define f as follows:

1. If p has small support, then we know that there is an LTF $\operatorname{sgn}(w^T x)$ that is consistent with g in $\operatorname{span}(\mathbf{q}^i : i \in \operatorname{supp}(p))$ given that $|\operatorname{supp}(p)| \le n/4$. Then we set
$$f(x) = \operatorname{sgn}\left(\mathbf{r}^T x + \delta \cdot w^T x\right),$$
where we assume for intuition that δ is *infinitely*[3] small. In this case f is an LTF.
2. If p has large support, then we set $f(x)$ to be $g(x)$ for $x \in \mathbf{h}$ and $f(x) = \operatorname{sgn}(\mathbf{r}^T x)$ for $x \notin \mathbf{h}$. Given that g is an extension of $\boldsymbol{\psi}$, f is far from LTFs by Lemma 1.2.

As a result, if $(f, \mathcal{D})$ were given as the input instance to ALG, it would be able to tell which case p is. On the other hand, even though the player does not know p, she actually knows the value $f(x)$ for all points x except those that are not in the span of samples $\mathbf{X}$ to ALG but are infinitely close to $\mathbf{h}$. By our discussion earlier it is unlikely for ALG to query these points and thus, the player can simulates ALG running on f correctly with high probability to solve SSD.

For the hypercube $\{\pm 1\}^n$, the player again samples a hyperplane, draws m points from it as $\mathbf{Q}$, and uses it to embed the hidden distribution p from the SSD problem. Unlike in $\mathbb{R}^n$, however, it will be considerably more challenging to argue that our random hyperplane is difficult for ALG to find, especially because the query bound we aim for is superpolynomial. Naively, there are at most $n2^{n^2}$ hyperplanes over the cube, so if each query gives a bit of information we could only prove a query lower bound of $\Omega(n^2)$. One approach to limit the information that the algorithm receives is to take a highly biased hyperplane such that the answer to all of the testing algorithm's queries will be negative with high probability. Unfortunately, a stumbling block is that given even a single sample y, one of the n points within hamming distance 1 of y will likely be positive.

To circumvent this obstacle, our approach to design the distribution of such a hyperplane will have two components. We start by fixing a so-called "restricting" hyperplane h (the existence of such a hyperplane is proved via the probabilistic method). The purpose of this hyperplane will be to select a good subset of points from the hypercube with several favorable properties. In many ways, this subset will look like a random set of points from $\{\pm 1\}^n$. Afterwards, we select a biased hyperplane randomly via a carefully picked distribution. For the final LTF f, $f(x)$ is determined according to the restricting hyperplane h if x is not on it, and is determined according to the biased

[3] This will be one of the technical hurdles we need to overcome in the proof in Section 3.

hyperplane if x is on h. We then argue that it is hard for any (adaptive) algorithm to find many points on h that are also on the positive side of the biased hyperplane.

2 Preliminaries

Notation. We will write $\mathcal{N}_n$ to denote the standard n-dimensional Gaussian distribution $\mathcal{N}(0, I_n)$ over $\mathbb{R}^n$. Given a d-dimensional linear subspace S of $\mathbb{R}^n$, we will write $\mathcal{N}_S$ to denote the standard d-dimensional Gaussian distribution over S. For a set S, we denote by $\binom{S}{k}$ the set of all subsets of S of size k. We also denote $a \in [b-c, b+c]$ by $a = b \pm c$.

Distribution-free Testing. We review the model of distribution-free property testing. Let $f, g : \{\pm 1\}^n \to \{\pm 1\}$ (or $f, g : \mathbb{R}^n \to \{\pm 1\}$) denote two Boolean-valued functions over $\{\pm 1\}^n$ (or $\mathbb{R}^n$), and $\mathcal{D}$ denote a probability distribution over $\{\pm 1\}^n$ (or $\mathbb{R}^n$).

We define the distance between f and g with respect to $\mathcal{D}$ as

$$\mathrm{dist}_{\mathcal{D}}(f, g) = \Pr_{z \in \mathcal{D}}\big[f(z) \neq g(z)\big].$$

Given a class $\mathfrak{C}$ of Boolean functions over $\{\pm 1\}^n$ (or $\mathbb{R}^n$), we define

$$\mathrm{dist}_{\mathcal{D}}(f, \mathfrak{C}) = \inf_{g \in \mathfrak{C}} \Big(\mathrm{dist}_{\mathcal{D}}(f, g) \Big)$$

as the distance between f and $\mathfrak{C}$ with respect to $\mathcal{D}$. We also say that f is ϵ-far from $\mathfrak{C}$ with respect to $\mathcal{D}$ for some $\epsilon \geq 0$ if $\mathrm{dist}_{\mathcal{D}}(f, \mathfrak{C}) \geq \epsilon$. Now we define distribution-free testing algorithms.

Definition 2.1. *Let $\mathfrak{C}$ be a class of Boolean functions over $\{\pm 1\}^n$ (or $\mathbb{R}^n$). A distribution-free testing algorithm ALG for $\mathfrak{C}$ has access to a pair $(f, \mathcal{D})$, where f is an unknown Boolean function $f : \{\pm 1\}^n \to \{\pm 1\}$ and $\mathcal{D}$ is an unknown probability distribution over $\{\pm 1\}^n$, via*

1. *a black-box oracle that returns the value $f(z)$ when $z \in \{\pm 1\}^n$ is queried; and*
2. *a sampling oracle that returns a sample $\mathbf{z} \sim \mathcal{D}$ drawn independently from $\mathcal{D}$ each time.*

The algorithm ALG takes as input a distance parameter $\delta > 0$ and satisfies for any $(f, \mathcal{D})$:

1. *If $f \in \mathfrak{C}$, then T accepts with probability at least $2/3$; and*
2. *If f is δ-far from $\mathfrak{C}$ with respect to $\mathcal{D}$, then T rejects with probability at least $2/3$.*

We say an algorithm is sample-based if it can only draw a sequence of samples $(\mathbf{z}, f(\mathbf{z}))$ with $\mathbf{z} \sim \mathcal{D}$ and cannot make queries. Note that in the definition above every sample comes with the point $\mathbf{z}$ only; this is just to simplify the presentation because the algorithm can always query them later. Finally, we may always assume without loss of generality that an algorithm starts by drawing all samples it needs and then it can make queries only.

3 Warm-Up: Distribution-Free LTF Testing in $\mathbb{R}^n$

3.1 Preparation We start with a few simple geometric lemmas:

Lemma 3.1. *Let $Q \subseteq \mathbb{R}^n$ be a set of vectors such that any k-subset of Q is linearly independent. If $S, T \subseteq Q$ satisfy $|S| \leq |T| \leq k/2$, then we have* $\mathrm{span}(S) \cap \mathrm{span}(T) = \mathrm{span}(S \cap T)$.

Proof. Clearly, $\text{span}(S \cap T) \subseteq \text{span}(S) \cap \text{span}(T)$. To see the opposite inclusion, let $x \in \text{span}(S) \cap \text{span}(T)$. Then there exists coefficients α_s and β_t such that

$$\sum_{s \in S} \alpha_s s = x = \sum_{t \in T} \beta_t t \quad \Longrightarrow \quad \sum_{s \in S} \alpha_s s - \sum_{t \in T} \beta_t t = 0$$

Given that every set of k elements from Q is independent, $S \cup T$ is independent and thus, we must have $\alpha_s = 0$ for all $s \in S \setminus T$ and $\beta_t = 0$ for all $t \in T \setminus S$. As a result, we have $x \in \text{span}(S \cap T)$. ∎

Let $Q \subseteq \mathbb{R}^n$ be a set of points such that every $(n-1)$-subset of Q is linearly independent. The next lemma shows that we can extend any $\psi : Q \to \{\pm 1\}$ to $g_{Q,\psi}$ over $\mathbb{R}^n$ such that $g_{Q,\psi}$ agrees with an LTF in the span of any k-subset of Q with $k \leq n/3$.

LEMMA 3.2. *Let $Q \subseteq \mathbb{R}^n$ be a set of points such that every $(n-1)$-subset of Q is linearly independent. Given any $\psi : Q \to \{\pm 1\}$, there is a function $g_{Q,\psi} : \mathbb{R}^n \to \{\pm 1\}$ such that (i) $g_{Q,\psi}$ is an extension of ψ: $g_{Q,\psi}(x) = \psi(x)$ for all $x \in Q$ and (ii) for any k-subset $\{x^1, \ldots, x^k\}$ of Q with $k \leq n/3$, there exists a $w \in S^{n-1}$ such that $g_{Q,\psi}(x) = \text{sgn}(w^T x)$ for all $x \in \text{span}(\{x^1, ..., x^k\})$ and w only depends on $x^1, ..., x^k$ and $\psi(x^1), ..., \psi(x^k)$.*

Proof. We write g for $g_{Q,\psi}$ for convenience. We only describe how to define g in $\cup_{S \subseteq \binom{Q}{n/3}} \text{span}(S)$; we set $g(x) = 1$ for all other points $x \in \mathbb{R}^n$.

Let $S = \{s^1, \ldots, s^k\}$ be a k-subset of Q with $k \leq n/3$, and let $A : \text{span}(S) \to \mathbb{R}^n$ be the linear transformation such that $As^i = e_i$ for each $i \in [k]$ and let $u \in \mathbb{R}^n$ be the vector such that $u_i = \psi(s^i)$ for each $i \in [k]$ and is 0 elsewhere. Next we define $g_S = \text{sgn}(u^T Ax)$ and take $g(x) = g_S(x)$ for all $x \in \text{span}(S)$. Setting w to be $u^T A$ after normalization (since w cannot be the all-zero vector), we note that w only depends on S and ψ over S.

If well-defined, g satisfies properties (i) and (ii). So it remains to show that g is indeed well-defined. Suppose that $x \in \text{span}(S) \cap \text{span}(T)$ for two subsets S and T of Q both of size at most $n/3$. We'll show that $g_S(x) = g_T(x)$. By Lemma 3.1 we have $x \in \text{span}(S \cap T)$. Denote $S \cap T = \{v^1, ..., v^\ell\}$ and let $x = \alpha_1 v^1 + ... + \alpha_\ell v^\ell$. We then have that

$$g_S(x) = \text{sgn}\left(\textstyle\sum_{i \in [\ell]} \alpha_i \cdot \psi(v^i)\right) = g_T(x).$$

This shows that g is well defined and finishes the proof of the lemma. ∎

3.2 The Hidden Slab Lemma Let $\ell < n-1$ and let Y be an ℓ-subset of S^{n-1}. We write $\mathcal{R}_Y$ to denote the Gaussian distribution over $\text{span}(Y)^\perp$. Given $r \in \mathbb{R}^n$ and $\epsilon > 0$, we define the (r, ϵ)-slab[4] to be

$$\mathsf{slab}(r, \epsilon) := \left\{x \in \mathbb{R}^n : |r^T x| \leq \epsilon \|x\|_2\right\}.$$

The (simple) hidden slab lemma below shows that, when given only an ℓ-subset Y of S^{n-1}, any set of points that are not too close to $\text{span}(Y)$ has little chance of landing in $\mathsf{slab}(\mathbf{r}, \epsilon)$ when $\mathbf{r} \sim \mathcal{R}_Y$.

[4]Looking ahead, the definition is slightly different from slabs in the lower bound proof for $\{\pm 1\}^n$; this is because in the latter all points we query have a fixed ℓ_2 norm $\sqrt{n}$.

LEMMA 3.3. *Let $0 < \epsilon, \delta < 1$ and N be a positive integer such that $\delta \geq nN\epsilon$. Let Y be an ℓ-subset of S^{n-1} with $\ell < n-1$. Let $\{z^1, \ldots, z^N\}$ be a set of points in $\mathbb{R}^n$ such that $\|(z^i)^\perp\|_2 \geq \delta \|z^i\|_2$ for all $i \in [N]$, where we use $(z^i)^\perp$ to denote the component of z^i orthogonal to* $\mathrm{span}(Y)$. *Then*

$$\Pr_{\mathbf{r} \sim \mathcal{R}_Y} \left[\exists i : z^i \in \mathsf{slab}(\mathbf{r}, \epsilon)\right] = o_n(1).$$

Proof. We claim that for every z^i, $z^i \in \mathsf{slab}(\mathbf{r}, \epsilon)$ with probability at most $1/(nN)$. This is because the probability of $z^i \in \mathsf{slab}(\mathbf{r}, \epsilon)$ is the same as the probability of drawing an $\boldsymbol{\alpha} \sim \mathcal{N}(0,1)$ with $\boldsymbol{\alpha} \cdot \|(z^i)^\perp\|_2 \leq \epsilon \|z^i\|_2$ as $\langle \mathbf{r}, z^i \rangle = \langle \mathbf{r}, (z^i)^\perp \rangle$. The claim follows by using the fact that the density of $\mathcal{N}(0,1)$ is at most $1/\sqrt{2\pi}$ point-wise. The lemma follows by a union bound over z^i, $i \in [N]$. ∎

3.3 Lower Bounds for $\mathbb{R}^n$ We start to prove the lower bound for $\mathbb{R}^n$ (although along the way we will need one more subtle technical ingredient in Lemma 3.4). Assume for a contradiction that there is a distribution-free algorithm ALG for testing halfspaces over $\mathbb{R}^n$ such that ALG draws only $k = o(n/\log n)$ samples and makes $N =: N(n)$ queries for any N. Given that we don't care about how big N is, we can assume without loss of generality that ALG is nonadaptive (to simplify the presentation): ALG draws k samples, makes a batch of N queries, and then either accepts or rejects.

Given a sequence of k points $X = (x^1, \ldots, x^k)$, we write $\mathsf{ALG}(X)$ to denote the distribution over N-subsets of $\mathbb{R}^n$ it draws to query; we write $\mathsf{ALG}(X; f) \in \{0,1\}$ to denote the outcome of ALG (as a random variable even when X and f are fixed) when it receives X as its samples and makes its queries on f. $\mathsf{ALG}(X; f) = 1$ means that ALG accepts and $\mathsf{ALG}(X; f) = 0$ means that it rejects. For any $(f, \mathcal{D})$ such that f is an LTF and any $(g, \mathcal{D}')$ such that g is $\Omega(1)$-far from LTFs with respect to $\mathcal{D}'$, we have the following performance guarantee:

$$\Pr_{\mathbf{X} \sim \mathcal{D}} \left[\mathsf{ALG}(\mathbf{X}; f) = 1\right] \geq 0.9 \quad \text{and} \quad \Pr_{\mathbf{X} \sim \mathcal{D}'} \left[\mathsf{ALG}(\mathbf{X}; g) = 0\right] \geq 0.9 \tag{3.1}$$

where we write $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^k) \sim \mathcal{D}$ to denote drawing a sequence of k samples independently from $\mathcal{D}$. We describe how a player can use ALG to solve $\mathsf{SSD}(400n, 1/1600, 1/2)$ (for convenience we just write SSD below to denote SSD with these three parameters, with $m = 400n$).

Before the proof starts, we prove the following lemma showing that given any sequence X of k points from $\mathbb{R}^n$, there is a $\delta_X > 0$ such that it is unlikely for $\mathsf{ALG}(X)$ (recall this is a random N-subset) to contain a point that is close but not in the span of points in X.

LEMMA 3.4. *Let X be a sequence of k points from $\mathbb{R}^n$. There is a $\delta_X > 0$ such that with probability at least $1 - 1/n$, every point $z \in \mathsf{ALG}(X)$ queries is either in* $\mathrm{span}(X)$ *or satisfies $\|z^\perp\|_2 \geq \delta_X \cdot \|z\|_2$, where $z^\perp$ is its orthogonal component to* $\mathrm{span}(X)$.

Proof. Let $\{\mathbf{z}^1, \ldots, \mathbf{z}^N\}$ be the random set of points $\mathsf{ALG}(X)$ queries. Let the random variable

$$\mathbf{Z} := \min_{i : \mathbf{z}_i \notin \mathrm{span}(X)} \frac{\|\mathbf{z}_i^\perp\|_2}{\|\mathbf{z}_i\|_2}.$$

where the minimum is 0 if $z_i \in \mathrm{span}(X)$ for all $i \in [N]$. Let F be the CDF of $\mathbf{Z}$. Since F is right continuous, there must be a $\delta_X > 0$ such that $F(\delta_X) - F(0) \leq 1/n$. This finishes the proof. ∎

We extend the definition of δ to finite sets of points. Given a finite set $Q \subseteq \mathbb{R}^n$, we write $\delta_Q > 0$ to denote the minimum of δ_X over all finitely many sequences X of length at most k from Q. With this, we have everything we need to prove our lower bound.

Proof. [Proof of Theorem 1.1] Let p be an unknown distribution over $[m]$ with $m = 400n$ that is either large ($|\operatorname{supp}(p)| \geq 200n$) or small ($|\operatorname{supp}(p)| \leq n/4$), and $p(i) \geq \Omega(1/n)$ for all $i \in \operatorname{supp}(p)$. The player starts by drawing $\mathbf{r} \sim \mathcal{N}_n$, a sequence of m points $\mathbf{q}^1, \ldots, \mathbf{q}^m$ independently and uniformly from $S^{n-1} \cap \{x \in \mathbb{R}^n : \mathbf{r}^T x = 0\}$, and a coloring $\boldsymbol{\rho} : [m] \to \{\pm 1\}$ uniformly at random. The following two conditions hold with probability 1, which we assume in the rest of the proof:

1. Points $\mathbf{q}^1, \ldots, \mathbf{q}^m$ are distinct; we use $\mathbf{Q} = \{\mathbf{q}^1, \ldots, \mathbf{q}^m\}$ to denote the size-m set they form.
2. Every $(n-1)$-subset of $\mathbf{Q}$ is linearly independent. Letting $\boldsymbol{\psi} : \mathbf{Q} \to \{\pm 1\}$ be the map defined as $\boldsymbol{\psi}(\mathbf{q}^i) = \boldsymbol{\rho}(i)$ for each $i \in [m]$, we can use Lemma 3.2 to define $g_{\mathbf{Q},\boldsymbol{\psi}}$.

Using $\mathbf{r}, \mathbf{Q}$ and $\boldsymbol{\rho}$ together with the unknown distribution p over $[m]$, the player can "implicitly define" the following pair $(f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}, \mathcal{D}_{p,\mathbf{Q}})$:

1. If $|\operatorname{supp}(p)| \leq n/4$, then there is a vector $\mathbf{w} \in S^{n-1}$ such that $g_{\mathbf{Q},\boldsymbol{\psi}}$ satisfies
$$g_{\mathbf{Q},\boldsymbol{\psi}}(x) = \operatorname{sgn}(\mathbf{w}^T x)$$
for all $x \in \operatorname{span}(\mathbf{q}^i : i \in \operatorname{supp}(p))$ (given that the set has size at most $n/4 < n/3$). We set
$$f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}(x) = \operatorname{sgn}\left(\mathbf{r}^T x + \frac{\delta_{\mathbf{Q}}}{nN} \cdot \mathbf{w}^T x\right).$$
2. If $|\operatorname{supp}(p)| \geq 200n$, then we set $f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}(x) = g_{\mathbf{Q},\boldsymbol{\psi}}(x)$ for all x with $\mathbf{r}^T x = 0$, and set
$$f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}(x) = \operatorname{sgn}(\mathbf{r}^T x)$$
for every other point x (with $\mathbf{r}^T x \neq 0$).
3. The distribution $\mathcal{D}_{p,\mathbf{Q}}$ has probability $p(i)$ on $\mathbf{q}^i$ for each $i \in [m]$.

The player then asks ALG to run on $(f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}, \mathcal{D}_{p,\mathbf{Q}})$ even though she does not have the pair in hand because p is hidden to her.

Before moving on, we observe that when p has small support, $f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}$ is an LTF and when p has large support, using Lemma 1.2 the function is $\Omega(1)$-far from LTFs with probability $1 - o_n(1)$. Letting a be the hidden bit that is 1 if p has small support and 0 if p has large support, we have
$$\Pr_{\mathbf{r},\mathbf{Q},\boldsymbol{\rho},\mathbf{X} \sim \mathcal{D}_{p,\mathbf{Q}}}\left[\mathsf{ALG}(\mathbf{X}; f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}) = a\right] \geq 0.9 - o_n(1).$$
Recall that we use $\mathbf{X} \sim \mathcal{D}_{p,\mathbf{Q}}$ to denote a sequence of $k = o(n/\log n)$ samples drawn independently from $\mathcal{D}_{p,\mathbf{Q}}$. As a result, if the player can faithfully simulate running ALG on the pair, she would be able to solve SSD. To this end, the player draws a sequence of k samples $\mathbf{J} = (\mathbf{j}^1, \ldots, \mathbf{j}^k)$ from p and sends $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^k)$ to ALG with $\mathbf{x}^i = \mathbf{q}^{\mathbf{j}^i} \in \mathbf{Q}$ for each i so that $\mathbf{X}$ is distributed correctly. Hence we have
$$\Pr_{\mathbf{r},\mathbf{Q},\boldsymbol{\rho},\mathbf{J}}\left[\mathsf{ALG}(\mathbf{X}; f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}) = a\right] \geq 0.9 - o_n(1).$$

Next ALG randomly picks a set of N points to query from the distribution $\mathsf{ALG}(\mathbf{X})$. Although the player does not know exactly $f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}$, she actually knows its values for most points $x \in \mathbb{R}^n$ (letting $\mathbf{Y}$ be the set of points in $\mathbf{X}$ below):

1. If $x \in \text{span}(\mathbf{Y})$, then in both cases
$$f_{p,\mathbf{r},\mathbf{Q},\boldsymbol{\rho}}(x) = g_{\mathbf{Q},\boldsymbol{\psi}}(x) = \text{sgn}(\mathbf{w}^T x),$$
where $\mathbf{w}^T$ is a vector that the player can compute by herself using $\mathbf{Y}$ and $\boldsymbol{\psi}$;
2. If x satisfies $|\mathbf{r}^T x| > (\delta_{\mathbf{Q}}/nN) \cdot \|x\|_2$, then in both cases the value is $\text{sgn}(\mathbf{r}^T x)$;
3. So the only points x that the player does not know how to answer satisfy both
$$x \notin \text{span}(\mathbf{Y}) \quad \text{and} \quad |\mathbf{r}^T x| \leq \frac{\delta_{\mathbf{Q}}}{nN} \cdot \|x\|_2 \leq \frac{\delta_{\mathbf{Y}}}{nN} \cdot \|x\|_2.$$
We denote this set by $F_{\mathbf{r},\mathbf{Y}}$. In this case the player just returns -1 by default.

The player follows the strategy above to finish simulating ALG and returns what ALG returns. The probability that the player returns the correct bit a is at least
$$0.9 - o_n(1) - \Pr_{\mathbf{r},\mathbf{Q},\boldsymbol{\rho},\mathbf{J}}\Big[\mathsf{ALG}(\mathbf{X}) \text{ overlaps with } F_{\mathbf{r},\mathbf{Y}}\Big].$$
We finish the proof by showing that the last probability is $o_n(1)$. This contradicts with Theorem 1.3 for SSD because the player used only $k = o(n/\log n)$ samples.

We assume for a contradiction that the probability is $\Omega(1)$. Then there is a way to fix $\mathbf{X} = X$ and $\mathbf{Y} = Y$ such that
$$\Pr_{\mathbf{r},\mathbf{Q},\boldsymbol{\rho},\mathbf{J}}\Big[\mathsf{ALG}(\mathbf{X}) \text{ overlaps with } F_{\mathbf{r},\mathbf{Y}} \mid \mathbf{X} = X \wedge \mathbf{Y} = Y\Big] \geq \Omega(1).$$
Let $\ell \leq k = o(n/\log n)$ be the size of Y. We observe that conditioning on $\mathbf{X} = X$ and $\mathbf{Y} = Y$, $\mathbf{r}$ is distributed as $\mathcal{R}_Y$. As a result, the probability above is the same as
$$\Pr_{\mathbf{r}\sim\mathcal{R}_Y}\Big[\mathsf{ALG}(X) \text{ overlaps with } F_{\mathbf{r},Y}\Big] \geq \Omega(1).$$
This and Lemma 3.4 imply there is an N-subset $\{z^1, \ldots, z^N\}$ in the support of $\mathsf{ALG}(X)$ such that

1. Every z^i satisfies either $z^i \in \text{span}(X) = \text{span}(Y)$ or its orthogonal component $(z^i)^\perp$ with respect to $\text{span}(Y)$ satisfies $\|(z^i)^\perp\|_2 \geq \delta_Y \cdot \|z\|_2$;
2. The set $\{z^1, \ldots, z^N\}$ satisfies
$$\Pr_{\mathbf{r}\sim\mathcal{R}_Y}\Big[\{z^1, \ldots, z^N\} \cap F_{\mathbf{r},Y} \neq \emptyset\Big] \geq \Omega(1).$$

However, it follows from the Hidden Slab Lemma that the probability for some z^i to land in the $\mathsf{slab}(\mathbf{r}, \delta_Y/(nN))$ is $o_n(1)$, a contradiction. This finishes the lower bound proof for $\mathbb{R}^n$. ∎

4 Testing Hyperplanes over $\{\pm 1\}^n$ with Few Samples

We now prove our lower bound for $\{\pm 1\}^n$. As in the warm up over $\mathbb{R}^n$, the proof will follow from finding a large set of points on a hyperplane and embedding an instance of the support size distinction problem. To find such a hyperplane, we begin by finding a large restricting hyperplane such that the points on it look roughly random. We then combine this with a biased hyperplane to get a distribution over hard to find hyperplanes that we can use in our reduction.

4.1 The Restricting Hyperplane The goal of this section will be to prove the Good Restricting Hyperplane Lemma. Unfortunately, we will need a few definitions to state this lemma formally. But, it roughly states that there exists a hyperplane h such that (1) the hyperplane contains many points, say $\omega(n^{100})$ and (2) every small set of points from the hyperplane, say of size $O(n/\log(n))$, has various properties we'd expect from a random set of points from $\{\pm 1\}^n$ such as linear independence. In general, the lemma gives a tradeoff between the number of points on the hyperplane and the size of the sets in condition (2). Eventually, when we do our reduction, if an algorithm only uses k samples, we will choose a restricting hyperplane such that every set of $O(k)$ points look random. Later, we'll choose a biased hyperplane such that any algorithm algorithm must query $|h \cap \{\pm 1\}^n|^c$ points for some constant $c > 0$ on the restricting hyperplane before it can find a point that is near the biased hyperplane.

4.1.1 A Distribution over Hyperplanes As one might expect, we prove the existence of large restricting hyperplane whose points from $\{\pm 1\}^n$ look "random" via the probabilistic method. As such, we begin by describing our distribution over hyperplanes $\mathcal{H}_M$, which is parametrized by a real number $M \geq 1$: Choose a direction $\mathbf{v} \sim \mathcal{N}_n$ and random vector $\mathbf{u} \in \{\pm 1\}^n$ *independently.* We then let $\mathbf{w}_i = \lfloor M\mathbf{v}_i + 1/2 \rfloor$ and take our hyperplane to be $\mathbf{h} = \{x : \mathbf{w}^T(x - \mathbf{u}) = 0\}$.

For intuition, we remark that small sets of points on $\mathbf{h}$ behaves very much like a random set where each point appears independently with probability $\frac{1}{M \cdot poly(n)}$. We now move towards formalizing this intuition and proving these hyperplanes have many points. We'll start with a few concentration bounds.

LEMMA 4.1. *Let* $\mathbf{v} \sim \mathcal{N}_n$ *and* $\mathbf{w}_i = \lfloor Mv_i + 1/2 \rfloor$ *where* $M \geq 1$ *then*

$$\Pr_{\mathbf{w}}\left[\|\mathbf{w}\|_2 \geq 3M\sqrt{n}\right] = o_n(1).$$

Proof. Note that we have by chi-square concentration bounds [18]

$$\Pr_{v \sim \mathcal{N}_n}\left[\|\mathbf{v}\|_2 \geq 2\sqrt{n}\right] = o_n(1)$$

The result then follows since $\|\mathbf{w}\| \leq M\|\mathbf{v}\| + \sqrt{n}$ ∎

LEMMA 4.2. *For i.i.d Radamacher random variables* $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n$ *and any* $w \in \mathbb{R}^n$ *we have*

$$\Pr_{\boldsymbol{\epsilon}_1,...,\boldsymbol{\epsilon}_n}\left[\left|\sum_i w_i \boldsymbol{\epsilon}_i\right| \geq t\|w\|_2\right] \leq t^{-2}$$

.

Proof. The proof follows by the second moment method. Let $\mathbf{X} = \sum_i w_i \boldsymbol{\epsilon}_i$. Then

$$\mathbb{E}[\mathbf{X}^2] = \sum_i w_i^2 = ||w||_2^2$$

We then have by Chebychev's inequality that

$$\Pr_{\boldsymbol{\epsilon}_1,...,\boldsymbol{\epsilon}_n}\left[\left|\sum_i w_i \boldsymbol{\epsilon}_i\right| \geq ||w||_2 t\right] \leq \frac{||w||_2^2}{t^2||w||_2^2} = t^{-2}$$

This finishes the proof of the lemma. ∎

Together these imply a lower bound on the number of points that appear on our random hyperplane.

LEMMA 4.3. *We have*

$$\Pr_{\mathbf{h}\sim\mathcal{H}_M}\left[|\mathbf{h}\cap\{\pm 1\}^n| \geq \frac{2^n}{nM}\right] = 1 - o_n(1)$$

Proof. Assume that $\|\mathbf{w}\|_2 \leq 3M\sqrt{n}$ and $|\mathbf{w}^T\mathbf{u}| \leq \|\mathbf{w}\|_2 \log(n)$, which happens with high probability by Lemma 4.1 and Lemma 4.2. If we consider sampling $\mathbf{w}$ first, then $\mathbf{u}$ will be uniformly distributed over the $(1-o_n(1))2^n$ points in the cube satisfying $|\mathbf{w}^T x| \leq \|\mathbf{w}\|_2 \log(n) \leq 6M\sqrt{n}\log(n)$. It then follows that there are at most $6M\sqrt{n}\log(n)\frac{2^n}{nM}$ values in $\{\pm 1\}^n$ for $\mathbf{u}$ that we could take that lead to planes $\mathbf{h}$ with fewer than $\frac{2^n}{nM}$ points. Thus,

$$\Pr_{\mathbf{w},\mathbf{u}}\left[|\mathbf{h}\cap\{\pm 1\}^n| \leq \frac{2^n}{nM}\middle| \|\mathbf{w}\|_2 \leq 3M\sqrt{n}, |\mathbf{w}^T\mathbf{u}| \leq \|\mathbf{w}\|_2\log(n)\right] \leq \frac{6M\sqrt{n}\log(n)\frac{2^n}{nM}}{(1-o_n(1))\cdot 2^n} = \widetilde{O}(n^{-1/2})$$

which finishes the proof of the lemma. ∎

We can also control the probability that multiple vectors appear on the hyperplane. To do so, we need the following standard estimate

LEMMA 4.4. (ODLYZKO [20]) *Let S be an affine subspace of dimension at most k in $\mathbb{R}^n$ then $|\{\pm 1\}^n \cap S| \leq 2^k$.*

For completeness, we include a proof in the appendix.

LEMMA 4.5. *Suppose $M \leq 2^n$, $\mathbf{h}\sim\mathcal{H}_M$, and $v^1,..,v^k \in \mathbb{R}^n$ are affinely independent vectors. Then*

$$\Pr_{\mathbf{h}\sim\mathcal{H}_M}[v^1,...,v^k \in \mathbf{h}] \leq \frac{2}{M^k}$$

Proof. Let $\mathbf{w}$ be the weight vector of $\mathbf{h}$ and $\mathbf{u}$ be the random point in $\{\pm 1\}^n$ selected to define h. We first prove the following claim:

CLAIM 4.1. *Let $u^1,...,u^k$ be independent vectors, then*

$$\Pr_{\mathbf{w}}\left[\forall i \quad \mathbf{w}^T u^i = 0\right] \leq \left(\frac{1}{\sqrt{2\pi}M}\right)^k$$

Proof. Let U be the matrix whose ith row is $(u^i)^T$. Note that by applying row operations to U we can assume that $U = (I_k|R)P$, where I_k is the $k\times k$ identity matrix, $R\in\mathbb{R}^{k\times n-k}$ is a arbitrary matrix, and P is a permutation matrix corresponding to some permutation $\sigma\in S_n$. It then follows that

$$\begin{aligned}\Pr_{\mathbf{w}}[U\mathbf{w}=0] &= \Pr_{\mathbf{w}}[(I_k|R)P\mathbf{w}=0]\\ &= \sum_{w_{k+1},\ldots,w_n}\Pr\left[\mathbf{w}_{\sigma(k+1)}=w_{k+1},\ldots,\mathbf{w}_{\sigma(n)}=w_n\right]\prod_{i=1}^{k}\Pr\left[\mathbf{w}_{\sigma(i)} = -\sum_{j=k+1}^{n} R_{i,j}w_{\sigma(j)}\right]\\ &\leq \left(\frac{1}{\sqrt{2\pi}M}\right)^k\end{aligned}$$

where the final inequality used the fact that $\mathbf{w}_{\sigma(i)}$ takes a value c with probability at most $\frac{1}{\sqrt{2\pi}M}$. ∎

To prove the lemma, we now split into two cases depending on whether $\mathbf{u}$ is in $\text{aff}(u^1, ..., u^k)$. Namely,

$$\Pr_{\mathbf{w},\mathbf{u}}\left[\forall i \quad \mathbf{w}^T(u^i - \mathbf{u}) = 0\right] = \Pr_{\mathbf{w},\mathbf{u}}\left[\forall i \quad \mathbf{w}^T(u^i - \mathbf{u}) = 0 \wedge \mathbf{u} \in \text{aff}(u^1, ..., u^k)\right] + \Pr_{\mathbf{w},\mathbf{u}}\left[\forall i \quad \mathbf{w}^T(u^i - \mathbf{u}) = 0 \wedge \mathbf{u} \notin \text{aff}(u^1, ..., u^k)\right]$$

To bound the first term we note

$$\Pr_{\mathbf{w},\mathbf{u}}\left[\forall i \quad \mathbf{w}^T(u^i - \mathbf{u}) = 0 \wedge \mathbf{u} \in \text{aff}(u^1, ..., u^k)\right] \leq \Pr_{\mathbf{w}}\left[\forall i \quad \mathbf{w}^T(u^i - u^1) = 0\right] \Pr_{\mathbf{u}}\left[\mathbf{u} \in \text{aff}(u^1, ..., u^k)\right] \leq \left(\frac{1}{\sqrt{2\pi}M}\right)^{k-1} \frac{2^{k-1}}{2^n}$$

by the claim and Lemma 4.4.

On the other hand for the second term, we see that

$$\Pr_{\mathbf{w},\mathbf{u}}\left[\forall i \quad \mathbf{w}^T(u^i - \mathbf{u}) = 0 \wedge \mathbf{u} \notin \text{aff}(u^1, ..., u^k)\right] = \sum_{u \notin \text{aff}(u^1,...,u^k)} \frac{1}{2^n} \Pr_{\mathbf{w}}\left[\forall i \quad \mathbf{w}^T(u^i - u) = 0\right] \leq \left(\frac{1}{\sqrt{2\pi}M}\right)^k$$

by the claim. So it follows that

$$\Pr_{\mathbf{h}\sim\mathcal{H}_M}\left[\forall i \quad u^i \in \mathbf{h}\right] \leq \frac{1}{(\sqrt{2\pi}M)^k} + \frac{2^{k-n-1}}{(\sqrt{2\pi}M)^{k-1}} \leq \frac{2}{M^k}$$

∎

With this, we immediately also get an upper bound on the number of points. Namely,

LEMMA 4.6. *Suppose $M \leq 2^n$, then*

$$\Pr_{\mathbf{h}\sim\mathcal{H}_M}\left[|\mathbf{h} \cap \{\pm 1\}^n| \leq \frac{2^n n^2}{M}\right] = 1 - o_n(1)$$

Proof. By Lemma 4.5, we have that for any $x \in \{\pm 1\}^n$

$$\Pr_{\mathbf{h}\sim\mathcal{H}_M}\left[x \in \mathbf{h}\right] \leq \frac{2}{M}$$

So we have that

$$\mathbb{E}\left[|\mathbf{h} \cap \{\pm 1\}^n|\right] \leq \frac{2 \cdot 2^n}{M}$$

Markov's inequality then gives the desired result. ∎

Now the key property that we will need is that unlikely events do not occur among the points on our random hyperplane. Namely, we make the following definition

DEFINITION 4.1. (RARE PROPERTY) *We say that a property $\mathcal{P}$ on sets of vectors $S \subseteq \{\pm 1\}^n$ is rare if there exists a $c < 1$ such that for n sufficiently large and all $k \leq \frac{n}{\log(n)}$ we have that $\Pr(\{\mathbf{y}^1, ..., \mathbf{y}^k\}$ satisfies $\mathcal{P}) \leq c^n$ where $\mathbf{y}^1, ..., \mathbf{y}^k$ are i.i.d are independent Bernoulli vectors.*

The key result about our distribution over hyperplanes will be that

LEMMA 4.7. *Let $\mathcal{P}$ be a rare property, $M \leq 2^n$, and $\mathbf{h} \sim \mathcal{H}_M$, then there exists a constant $\beta > 0$ (independent of dimension) such that with probability $1 - o_n(1)$ every subset of size at most $\frac{\beta n}{\log\left(\frac{2^{n+1}}{M}\right)}$ in $\mathbf{h} \cap \{\pm 1\}^n$ doesn't not satisfy $\mathcal{P}$.*

To prove the lemma, we will need the following result

LEMMA 4.8. *The following property is rare: S is affinely independent and there exists a $y \in \{\pm 1\}^n \setminus S$ such that $S \cup \{y\}$ is affinely dependent.*

We'll prove this in the next section, but assuming it's true for now we can prove Lemma 4.7

Proof. [Proof of Lemma 4.7] We'll show that there are no sets of size $A = A(n)$ on h satisfying $\mathcal{P}$ for some A to be chosen later. We start by making a new property $\mathcal{P}'$: A set S has property $\mathcal{P}'$ if it is affinely independent and either has property $\mathcal{P}$ or there exists a $y \in \{\pm 1\}^n$ such that $y \notin S$ and $y \in \text{aff}(S)$. We claim that it suffices to show that no tuple of size $k \leq A$ on h has property $\mathcal{P}'$. Indeed, suppose this is the case and that $S \subseteq \mathbf{h} \cap \{\pm 1\}^n$ has property $\mathcal{P}$. If S is affinely independent, then we have that S satisfies $\mathcal{P}'$. If S is not affinely independent then there exists $S' \subseteq S$ and a $y \in S \setminus S'$ such that S' is affinely independent and $S \cup \{y\}$ is affinely dependent. Since no subset of size $k \leq A$ has property $\mathcal{P}'$ we have that $S' \not\subseteq \mathbf{h}$ and thus $S \not\subseteq \mathbf{h}$ as desired.

Now note that $\mathcal{P}'$ is also a rare event for n sufficiently large, as the probability a set S satisfies $\mathcal{P}'$ is bounded by the sum of the probability that it satisfies $\mathcal{P}$ and the probability that S is affinely independent and there exists a $y \notin S$ such that $\{y\} \cup S$ is affinely dependent. Since both of these events are rare we have that there exists a $c < 1$ such that the probability of satisfying $\mathcal{P}'$ is bounded by $2c^n$ which is at most c'^n for some $c' < 1$ and n sufficiently large.

Now in a slight abuse of notation we say that $S \in \mathcal{P}'$ if S satisfies $\mathcal{P}'$. Using Lemma 4.5, we now compute the expected number of sets of size at most A satisfying $\mathcal{P}'$

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{A} \sum_{\substack{S \in \mathcal{P}' \\ |S|=k}} 1_{S \subseteq \mathbf{h}}\right] &\leq \sum_{k=1}^{A} c'^n 2^{kn} \frac{2}{M^k} \\
&\leq 2 \sum_{k=1}^{A} c'^n \left(\frac{2^{n+1}}{M}\right)^k \\
&\leq 2c'^n \left(\frac{2^{n+1}}{M}\right)^{A+1}
\end{aligned}$$

Taking $A = \left\lfloor \frac{\beta n}{\log\left(\frac{2^{n+1}}{M}\right)} \right\rfloor - 1$ for $\beta = -\frac{1}{2}\log(c')$ then gives us that the expectation is $o_n(1)$. The result then follows by Markov's inequality. ∎

We can do something similar for pairwise events. Namely, we will use the following result

LEMMA 4.9. *Suppose $M \leq 2^n/n^2$ and $\mathbf{h} \sim \mathcal{H}_M$, then with high probability, for every pair $(\mathbf{x}, \mathbf{y})$ of distinct vectors in $h \cap \{\pm 1\}^n$, we have that $|\mathbf{x}^T\mathbf{y}| \leq O\left(\sqrt{n\log(2^n/M)}\right)$*

Proof. By the Bernstein bound, if x and y are random iid Bernoulli vectors then $\Pr[|\mathbf{x}^T\mathbf{y}| \geq t] \leq 2e^{-t^2/4n}$. It then follows that there are at most $2 \cdot 4^n e^{-t^2/4n}$ pairs with absolute inner product at least t. We then compute using Lemma 4.5 that $\mathbf{h}$ contains at most

$$\frac{4^{n+1}e^{-t^2/4n}}{M^2}$$

such pairs in expectation. Taking $t = 2\sqrt{n\log\left(\frac{4^{n+1}n^2}{M^2}\right)}$ and Markov's inequality then gives the desired result. ∎

4.1.2 Rare Properties and Results from Random Matrix Theory We now define various rare properties that will be useful in our analysis. We will start by proving Lemma 4.8. It will quickly follow from a result of Odlyzko.

THEOREM 4.1. (ODLYZKO [20]) *Suppose that $k \leq n - 10n/\log(n)$ then if $\mathbf{v}^1, ..., \mathbf{v}^k$ are random vectors from $\{\pm 1\}^n$ the probability that $\text{span}(\{\mathbf{v}^1, ..., \mathbf{v}^k\})$ contains a vector different from $\pm\mathbf{v}^j$ is at most $O(k^3(3/4)^n)$.*

We note that stronger statements hold. In particular, we can take $k = n - C$ for some absolute constant C [17]; however, we will not need this.

Proof. [Proof of Lemma 4.8] Note that we have that any set of 3 points on the hypercube is affinely independent. So we assume that $k \geq 3$. Randomly choose vectors $\mathbf{v}^1, ..., \mathbf{v}^k$ from $\{\pm 1\}^n$ uniformly and independently at random. Define $\mathbf{u}^i = \begin{pmatrix} 1 \\ \mathbf{v}^i \end{pmatrix}$. If $\{\mathbf{v}^1, ..., \mathbf{v}^k\}$ satisfy $\mathcal{P}$ then there must exist a vector $\mathbf{u} = \begin{pmatrix} 1 \\ \mathbf{v} \end{pmatrix}$ in the span of $\mathbf{u}^1, ..., \mathbf{u}^k$ such that $\mathbf{u} \neq \mathbf{u}^1, ..., \mathbf{u}^k$. Since we also have that $\mathbf{u} \neq -\mathbf{u}^1, ..., -\mathbf{u}^k$, it follows that if p is the probability that $\{\mathbf{v}^1, ..., \mathbf{v}^k\} \in \mathcal{P}$ then

$$p = \Pr\left[\text{span}\{\mathbf{r}^1, ..., \mathbf{r}^k\} \cap \{-1, 1\}^{n+1} \neq \pm\mathbf{r}^1, ..., \pm\mathbf{r}^k | \mathbf{r}^1_1 = 1, ..., \mathbf{r}^k_1 = 1\right]$$

where $\mathbf{r}^1, ...\mathbf{r}^k$ are random vectors from $\{-1, 1\}^{n+1}$. It then follows that

$$p2^{-k} \leq \Pr\left[\text{span}\{\mathbf{r}^1, ..., \mathbf{r}^k\} \cap \{-1, 1\}^{n+1} \neq \pm\mathbf{r}^1, ..., \pm\mathbf{r}^k\right] \leq O(k^3(3/4)^n)$$

So $p \leq O(k^3(3/4)^n 2^k)$ which is at most $.999^n$ for n sufficiently large and $k \leq n/\log(n)$. ∎

THEOREM 4.2. (BENNETT ET AL. [3]) *There exists absolute constants $c, \lambda > 0$ such that if $Y \in \{\pm 1\}^{n \times k}$ is a matrix with random Bernoulli entries and $k \leq \lambda n$ then $\Pr(\sigma_k(Y) \leq c\sqrt{n}) \leq e^{-cn}$. In other words, the following property is rare: the matrix M whose columns correspond to S has $\sigma_{|S|}(M) \leq c\sqrt{n}$.*

Finally, we have

LEMMA 4.10. (TAO AND VU [22]) *Let W be a fixed subspace of dimension $1 \leq d \leq n-4$ and let $\mathbf{x} \in \{\pm 1\}^n$ be a random Bernoulli vector. We then have that $\mathbb{E}[\operatorname{dist}(\mathbf{x}, W)^2] = n - d$. Moreover,*

$$\Pr_{\mathbf{x}}\left[|\operatorname{dist}(\mathbf{x}, W) - \sqrt{n-d}| \geq t+2\right] \leq 4e^{-t^2/16}$$

which implies

COROLLARY 4.1. *The following is a rare property: There exists a $y \in S$ such that $\|\operatorname{proj}(y, \operatorname{span}(S \setminus \{y\}))\|_2 \geq \sqrt{n}/100$).*

Proof. By a union bound, it suffices to show that if $\mathbf{y}^1, ..., \mathbf{y}^k$ are i.i.d. Bernoulli vectors then $\Pr\left[\|\operatorname{proj}(\mathbf{y}^1, \operatorname{span}(\mathbf{y}^2, ..., \mathbf{y}^k))\| \geq \sqrt{n}/100\right] \leq c^n$ for a constant $c > 0$. But indeed by Lemma 4.10 for any fixed $y^2, ..., y^k$ the probability that $\Pr\left[\|\operatorname{proj}(\mathbf{y}^1, \operatorname{span}(y^2, ..., y^k))\| \geq \sqrt{n}/100\right] \leq c^n$ when n is sufficiently large. ∎

4.1.3 Inference Sets

DEFINITION 4.2. (INFERENCE SETS) *We say a set $S = \{x^1, ..., x^k\}$ is an inference set with respect to a set $T = \{y_1, ..., y_\ell\}$ if for all x^i and for all α_j satisfying $\operatorname{proj}(x^i, \operatorname{span}(S)) = \alpha_1 y^1 + ... + \alpha_\ell y^\ell$ we have that*

$$\sum_{j \in [\ell]} \alpha_j \geq 1/100$$

LEMMA 4.11. *The following is a rare property: $S = \{x^1, ..., x^\ell\}$ has the property that there exists a set $T \subseteq S$ of size $|T| \leq \ell/2$ such that $S \setminus T$ is an inference set with respect to T.*

Proof. The proof follows primarily from the following claim:

CLAIM 4.2. *Let $y^1, ..., y^m \in \{\pm 1\}^n$ and Y be the matrix whose ith column is Y_i. Additionally suppose $\sigma_m(Y) \geq \Omega(\sqrt{n})$ and let $\mathbf{y}$ be a random point in $\{\pm 1\}^n$ with $\boldsymbol{\alpha}_i$ such that $\operatorname{proj}(\mathbf{y}, \{y^1, ..., y^k\}) = \sum_i \boldsymbol{\alpha}_i y^i$ then*

$$\Pr\left[\sum_i \boldsymbol{\alpha}_i \geq 1/100\right] \leq e^{-\Omega(n/m)}$$

Proof. Let $P : \mathbb{R}^n \to \mathbb{R}^k$ be the orthogonal projection onto $\operatorname{span}\{y^1, ..., y^k\}$. We then note that $\sum_i \boldsymbol{\alpha}_i y^i = 1^T (PY)^{-1} P\mathbf{y}$. Indeed, note that $\sum_i \alpha_i$ is a linear function of y, so it suffices to prove equality on a basis. We then observe $1^T (PY)^{-1} P y^i = 1^T (PY)^{-1} PY e_i = 1$ and if $x \perp \operatorname{span}\{y^1, ..., y^k\}$ then $w^T Y (PY)^{-1} P x = 0$ as desired.

Now we note that $1^T (PY)^{-1} P\mathbf{y}$ is a weighted sum of Bernoulli random variables. As such we will now bound $\|1^T (PY)^{-1} P\|_2^2$ so that we may use Hoeffding's inequality. Denote $v^T = 1^T (PY)^{-1} P$. Note that $v^T Y = 1^T$ and thus that

$$m = \|1^T\|_2^2 = \|v^T Y\|_2^2 = v^T Y Y^T v$$

Now observe that for any vector u such that $u^T Y = 0$, we have that $\langle u, y^i \rangle = 0$ for all i and $v^T u = 1^T (PY)^{-1} P u = 0$. Let the eigenvalues of YY^T be a $q_1, ..., q_n$ with corresponding eigenvalues

$\lambda_1, ..., \lambda_n$ where $\lambda_i = 0$ for $m < i \leq n$. Since $v^T u = 0$ for all u such that $u^T Y = 0$ we have that $v = \sum_{i=1}^{m} \beta_i q_i$. But now note that since the smallest non-zero singular value of Y is at least $c\sqrt{n}$ it follows that

$$v^T Y Y^T v = \sum_i \lambda_i \beta_i^2 \geq c^2 n \sum_i \beta_i^2 = c^2 n \|v\|_2^2$$

It then follows that $\|v\|_2^2 \leq O(m/n)$ and that

$$\Pr_{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m} \left[\left| \sum_i \boldsymbol{\alpha}_i \right| \geq t \right] \leq 2e^{-t^2/2\|v\|_2^2} \leq 2e^{-\Omega(nt^2/m)}$$

as claimed. ∎

Now let $\mathbf{S} = \{\mathbf{y}^1, ..., \mathbf{y}^\ell\}$, where each $\mathbf{y}^i$ is independently and uniformly drawn from $\{\pm 1\}^n$ and $\ell \leq n/\log(n)$. By a union bound and symmetry, we note that

$$\begin{aligned}\Pr\left[\exists \mathbf{T} \subseteq \mathbf{S} : \mathbf{S} \setminus \mathbf{T} \text{ inference set with respect to } \mathbf{T} \wedge |\mathbf{T}| \leq \ell/2\right] \\ \leq 2^\ell \max_{k \leq \ell/2} \Pr\left[\{\mathbf{y}^{k+1}, ..., \mathbf{y}^\ell\} \text{ inference set with respect to } \{\mathbf{y}^1, ..., \mathbf{y}^k\}\right]\end{aligned}$$

Fixing $k \leq \ell/2$, letting $\mathbf{y}^i = \sum_{j=1}^{k} \boldsymbol{\alpha}_j^i \mathbf{y}^j$ for $i \geq k+1$, and letting $\mathbf{Y}$ be the random matrix whose columns are $\mathbf{y}^1, ..., \mathbf{y}^k$ we see that

$$\begin{aligned}&\Pr\left[\{\mathbf{y}^{k+1}, ..., \mathbf{y}^\ell\} \text{ inference set with respect to } \{\mathbf{y}^1, ..., \mathbf{y}^k\}\right] \\ &\quad \leq \left[\{\mathbf{y}^{k+1}, ..., \mathbf{y}^\ell\} \text{ inference set with respect to } \{\mathbf{y}^1, ..., \mathbf{y}^k\} \wedge \sigma_k(\mathbf{Y}) \geq c\sqrt{n}\right] + c_1^n \\ &\quad \leq 2 \prod_{i=k+1}^{\ell} \Pr\left[\sum_{j=1}^{k} \boldsymbol{\alpha}_j^i \geq 1/100 \mid \sigma_k(\mathbf{Y}) \geq c\sqrt{n}\right] + c_1^n \\ &\quad \leq e^{-\Omega(n)}\end{aligned}$$

where the first inequality holds for some $c_1 < 1$ by Theorem 4.2 and the final inequality is a result of the claim. It then follows that

$$\Pr\left[\exists \mathbf{T} \subseteq \mathbf{S} : \mathbf{S} \setminus \mathbf{T} \text{ inference set with respect to } \mathbf{T} \wedge |\mathbf{T}| \leq \ell/2\right] \leq 2^\ell e^{-\Omega(n)} \leq c_2^n$$

when n is sufficiently large and where $c_2 < 1$ is a suitably large constant. ∎

4.1.4 Existence of a Good Restricting Hyperplane We now reap the benefits of our work:

LEMMA 4.12. (GOOD RESTRICTING HYPERPLANE LEMMA) *For any n sufficiently large and $\Omega(1) \leq \ell \leq O\left(\frac{n}{\log(n)}\right)$, there exists a hyperplane h_ℓ and absolute constant $\gamma > 0$ such that:*

(i) $\frac{2^{\gamma n/\ell}}{2n} \leq |h_\ell \cap \{\pm 1\}^n| \leq 2^{\gamma n/\ell} n^2$

(ii) For any $k \leq \ell$ and any distinct $y^1, ..., y^k \in h_\ell \cap \{\pm 1\}^n$, the matrix with columns $y^1, ..., y^k$ has smallest singular value $c\sqrt{n}$ for some absolute constant c.

(iii) *For any $k \le \ell$ and any distinct $y^1, ..., y^k \in h_\ell \cap \{\pm 1\}^n$, $\| \operatorname{proj}(y, \{y^1, ..., y^k\}\| \le \sqrt{n}/100$*

(iv) *Let $k \le \ell$, then for any vectors $y^1, ..., y^k \in h_\ell \cap \{\pm 1\}^n$ the largest inference set $T \subseteq h_\ell \cap \{\pm 1\}^n$ satisfies $|T| \le k$.*

(v) *For every distinct $x, y \in h_\ell \cap \{\pm 1\}^n$, $|x^T y| \le O\left(\frac{n}{\sqrt{\ell}}\right)$*

(vi) *Let $k \le \ell$, then for any vectors $y^1, ..., y^k, y \in h_\ell \cap \{\pm 1\}^n$ if $\operatorname{proj}(y, \{y^1, ..., y^k\}) = \sum_i \alpha_i y^i$ then $\sum |\alpha_i| \le \sqrt{n}/c$, where c is the constant from (ii).*

Proof. We proceed via the probabilistic method. Let $\mathbf{h} \sim \mathcal{H}_M$ where $M = 2^{n - \gamma n/\ell + 1}$ for a sufficiently small constant γ. We then have that that $\mathbf{h}$ satisfies the property (i) with probability $1 - o(1)$ by Lemmas 4.3 and 4.6. Similarly, using Lemma 4.7 we get that for a small enough constant absolute constant β, $\mathbf{h}$ satisfies properties (ii), (iii) and (iv) by Theorem 4.2, Corollary 4.1, and Lemma 4.11 for $k \le \frac{\beta n}{2 \log\left(\frac{2^{n+1}}{M}\right)}$. Taking $\gamma = \beta/2$ gives, that these properties hold for $k \le \ell$. Additionally, we get property (v) holds with the bound $O\left(\sqrt{n \log(2^n/M)}\right)$ high probability by Lemma 4.9, which is $O(n/\sqrt{\ell})$ for our choice of M. Finally, we observe that property (vi) is a corollary of property (ii). Namely, let Y be the matrix whose ith column is y^i and α the vector whose ith entry is α_i. Then

$$c\sqrt{n}\|\alpha\|_2 \le \|Y\alpha\|_2 = \|\operatorname{proj}(y, \{y^1, ..., y^k\})\|_2 \le \|y\| = \sqrt{n}$$

where we used the fact that the smallest singular value of Y is $c\sqrt{n}$. Thus, $\|\alpha\|_2 \le 1/c$ which implies $\|\alpha\|_1 \le \sqrt{n}/c$ as desired. ∎

4.2 The Biased Hyperplane For our second hyperplane, we will use a biased Gaussian slab and intersect it with the good restricting hyperplane from the previous section. Before describing the distribution, we set up some notation. We will consider a fixed $\ell \in \left[n^{.99}, O\left(\frac{n}{\log(n)}\right)\right]$, where the upper bound is such that ℓ is a valid parameter in the good restricting hyperplane lemma and $|h_\ell \cap \{\pm 1\}^n| \ge n^{1000}$. Throughout this section, we will think of ℓ as being fixed, as such, we will often write h_ℓ as h. We will denote $H = h \cap \{\pm 1\}^n$ and $N = |H|$.

Our distribution over biased hyperplanes will in fact be a distribution over biased Gaussian slabs and is parametrized by ϵ, t and s^*.

Choices of ϵ and t. We set $\epsilon = \frac{1}{\sqrt{N}}$ and $t = \Theta(\sqrt{n \log N})$ such that

$$\Pr_{\mathbf{x} \sim \mathcal{N}}\left[\mathbf{x} = (t \pm \epsilon)/\sqrt{n}\right] = \frac{2}{N^{3/4}}. \tag{4.2}$$

With ϵ and t fixed, we define the *r-slab* of $r \in \mathbb{R}^n$ as

$$\mathsf{slab}(r) := \left\{a \in H : r^T a = t \pm \epsilon\right\}$$

and the *size* of $r \in \mathbb{R}^n$ as $\mathsf{size}(r) = |\mathsf{slab}(r)|$.

Choice of s^*. We start with some notation. Let

$$\gamma_s = N^{1/4}\left(1 + \frac{1}{n}\right)^s$$

for each $s \geq 0$. We divide $[0 : N]$ into buckets $B_0, B_1, \ldots, B_U$ with $B_0 = [0, \gamma_0)$ and $B_s = [\gamma_{s-1}, \gamma_s)$ for each $s \geq 1$, with $U = O(n \log N)$ to cover $[0 : N]$. We say $r \in \mathbb{R}^n$ is in bucket s if $\mathsf{size}(r) \in B_s$. It will be convenient to set s^* to be the bin that maximizes $\binom{\gamma_s}{\ell/3} \cdot \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}) \in B_s\right]$ among $s \geq 1$, breaking ties arbitrarily. As a short hand, we'll write B^* for B_{s^*} and γ^* for γ_{s^*}.

With these values all set, we let $(\mathbf{r}, \mathbf{Y}) \sim \mathcal{R}_d$ be the distribution where the pair is generated as follows. First we draw $\mathbf{r} \sim \mathcal{N}_n^*$ where we write $\mathcal{N}_n^*$ to denote the distribution of $\mathbf{r} \sim \mathcal{N}_n$ conditioning on $\mathsf{size}(\mathbf{r}) \in B^*$ and then we draw a d-subset $\mathbf{Y}$ from the $\mathbf{r}$-slab (of size in B^*) uniformly at random. We mostly work with with $\mathcal{R}_{\ell/3}$, as such we will simply denote it by $\mathcal{R}$.

We will use $\mathcal{R}_d$ in our reduction to the support size distinction problem in a similar manner to our proof for $\mathbb{R}^n$. Before getting to the reduction, however, we will need to prove that points in $|\mathsf{slab}(r)|$ are hard to find.

4.2.1 Properties of the Biased Hyperplane Distribution We now define some more notation. Given a k-subset $Y = \{y^1, \ldots, y^k\}$ of H, we write I_Y to denote the largest inference set of Y in H (so $I_Y \subseteq H$ and $I_Y \cap Y = \emptyset$); it follows from the property of the restricting hyperplane that $|I_Y| \leq k$ for all Y. We use $S_Y := \mathrm{span}(Y, I_Y)$ to denote the linear space spanned by Y and I_Y of dimension at most $2k$. We write W_Y to denote the set of $w \in S_Y$ such that $w^T y^i = t \pm \epsilon$ for every $i \in [k]$. Similarly we use R_Y to denote the set of $r \in \mathbb{R}^n$ such that $r^T y^i = t \pm \epsilon$ for every $i \in [k]$ (or equivalently, $\mathrm{proj}(r, S_Y) \in W_Y$).

We now formalize what it means for it to be hard to find a point near the slab. Formally, we'll show

Lemma 4.13. (Hidden Slab Lemma) *Suppose $(\mathbf{r}, \mathbf{Y}) \sim \mathcal{R}_d$ for some $d \leq \ell/3$ and ALG is a randomized algorithm that is given $\mathbf{Y}$ and makes at most $N^{1/8}$ (adaptive) queries to the LTF $\mathrm{sgn}(\mathbf{r}^T x - t)$ with points $x \in H$, then $\mathsf{ALG}(\mathbf{Y})$ queries a point in $\{x \in H \setminus \mathbf{Y} : |\mathbf{r}^T x - t| \leq n\epsilon\}$ with probability $O(N^{-1/40})$.*

To prove this, we first prove the following lemma to lift probabilities from a Gaussian distribution to $\mathcal{R}$.

Lemma 4.14. (Gaussian Transfer Lemma) *Let E be any event on $r \in \mathbb{R}^n$ and $Y \subseteq \binom{H}{\ell/3}$ (we write $E(r, Y)$ to denote that E holds on r and Y) then*

$$\Pr_{(\mathbf{r},\mathbf{Y}) \sim \mathcal{R}}\left[\Pr_{(\mathbf{r}',\mathbf{Y}') \sim \mathcal{R}}\left[E(\mathbf{r}', \mathbf{Y}) \mid \mathbf{Y}' = \mathbf{Y}\right] \geq 2N^{1/10} \Pr_{\mathbf{r}' \sim \mathcal{N}_n}\left[E(\mathbf{r}', \mathbf{Y}) | \mathbf{r}' \in R_{\mathbf{Y}}\right]\right] \leq O(N^{-1/20})$$

Proving the Gaussian Transfer Lemma will be the goal of the remainder of this section.

Lemma 4.15. *For each bucket $s \geq 1$ we have*

$$\frac{1}{2}\binom{\gamma_s}{\ell/3} \cdot \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}) \in B_s\right] \leq \sum_{Y \in \binom{H}{\ell/3}} \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathbf{r} \in R_Y \wedge \mathsf{size}(\mathbf{r}) \in B_s\right] \leq \binom{\gamma_s}{\ell/3} \cdot \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}) \in B_s\right].$$

For bucket 0, we have

$$\sum_{Y \in \binom{H}{\ell/3}} \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathbf{r} \in R_Y \wedge \mathsf{size}(\mathbf{r}) \in B_0\right] \leq \binom{\gamma_0}{\ell/3} \cdot \Pr_{\mathbf{r} \sim \mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}) \in B_0\right].$$

Proof. For the first inequality we have

$$\begin{aligned}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r})\in B_s\right] &= \sum_{\gamma\in B_s}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r})=\gamma\right]\\ &= \sum_{\gamma\in B_s}\frac{1}{\binom{\gamma}{\ell/3}}\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge\mathsf{size}(\mathbf{r})=\gamma\right]\\ &\geq \frac{1}{\binom{\gamma_s}{\ell/3}}\sum_{\gamma\in B_s}\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge\mathsf{size}(\mathbf{r})=\gamma\right]\\ &= \frac{1}{\binom{\gamma_s}{\ell/3}}\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge\mathsf{size}(\mathbf{r})\in B_s\right].\end{aligned}$$

The other side of the first inequality follows similarly, using $\ell \ll n$ and thus,

$$\binom{\gamma_s}{\ell/3}\leq 2\cdot\binom{\gamma_{s-1}}{\ell/3}.$$

The second inequality can also be proved similarly. ∎

We will also need to control the bucket B_0 since we chose $s^*\geq 1$.

LEMMA 4.16.

$$\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\mathit{size}(\mathbf{r})\in B_0]=o_n(1)$$

Proof. We'll show that $\mathrm{Var}[\mathsf{size}(\mathbf{r})]\leq O\left(\frac{t^2\sqrt{N}}{n\sqrt{\ell}}\right)$. Using our value of t, $\ell\geq n^{.99}$, and $\mathbb{E}[\mathsf{size}(r)]=2N^{1/4}$, the lemma then follows from Chebychev's inequality.

For a fixed pair x,y consider $\Pr_{\mathbf{r}\sim\mathcal{N}_n}[y\in\mathsf{slab}(\mathbf{r})|\mathbf{r}^Tx=\lambda]$. We will upper bound this probability when $\lambda=t\pm\epsilon$. Let $\alpha n=x^Ty$ and let $y^\perp=y-\alpha x$. Now recall that by the good supporting hyperplane lemma we have that $|\alpha|\leq O(\frac{1}{\sqrt{\ell}})$. So we can assume $|\alpha|\leq .01$. We now compute

$$\begin{aligned}\Pr\left(|r^Ty-t|\leq\epsilon\middle|r^Tx=\lambda\right) &= \int_{t-\alpha\lambda-\epsilon}^{t-\alpha\lambda+\epsilon}\frac{1}{\sqrt{2\pi}\|y^\perp\|}e^{-z^2/2\|y^\perp\|^2}dz\\ &\leq \frac{2\epsilon}{\sqrt{2\pi}\|y^\perp\|}e^{-(t-\alpha t-\epsilon)^2/2\|y^\perp\|^2}\\ &\leq \frac{2\epsilon}{\sqrt{2\pi n(1-\alpha^2)}}e^{-t^2/2n}e^{\alpha t^2/(1+\alpha)n}e^{O(\epsilon t/n)}.\end{aligned}$$

Using $e^x\leq 1+2|x|$ when $x\leq 1$ and that $\epsilon=o(n^{-2})$, we get

$$\leq \frac{2\epsilon}{\sqrt{2\pi n(1-\alpha^2)}}e^{-t^2/2n}(1+4|\alpha|t^2/n)(1+o(n^{-2})).$$

Since $\frac{1}{\sqrt{1-x^2}}\leq 1+x^2$ when $|x|\leq 1/2$, we then have

$$\leq \frac{2\epsilon(1+o(n^{-2}))(1+\alpha^2)}{\sqrt{2\pi n}}e^{-t^2/2n}(1+4|\alpha|t^2/n).$$

This is an increasing function in $|\alpha|$ using our bound from the Good Hyperplane Lemma we get

$$\leq \frac{2\epsilon(1+o(n^{-2}))(1+O\left(\frac{1}{\ell}\right))}{\sqrt{2\pi n}} e^{-t^2/2n}\left(1+O\left(\frac{t^2}{n\sqrt{\ell}}\right)\right).$$

Since $t=\omega(\sqrt{n})$ and $\ell=\omega(1)$ we have that $n^{-2} \ll \frac{1}{\ell} \ll \frac{t^2}{n\sqrt{\ell}}$. So we get that this is at most

$$\Pr_{r\sim\mathcal{N}_n}[y \in \mathsf{slab}(\mathbf{r})]\left(1+O\left(\frac{t^2}{n\sqrt{\ell}}\right)\right).$$

Thus,

$$\mathrm{Var}[\mathsf{size}(r)] = \mathbb{E}\mathsf{size}(r)^2 - (\mathbb{E}\mathsf{size}(r))^2 \leq O\left(\frac{t^2}{n\sqrt{\ell}}\right)(\mathbb{E}\mathsf{size}(r))^2 + \mathbb{E}\mathsf{size}(r) = O\left(\frac{t^2\sqrt{N}}{n\sqrt{\ell}}\right)$$

as claimed. ∎

We'll now need one final technical lemma for the proof to relate samples from $\mathcal{R}$ and samples from a Gaussian random slab.

Lemma 4.17.

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[\Pr_{\mathbf{r}'\sim\mathcal{N}_n}\left[\mathit{size}(\mathbf{r}') \in B^* \mid \mathbf{r}' \in R_{\mathbf{Y}}\right] \leq N^{-1/10}\right] \leq O(N^{-1/20}).$$

Proof. For a $\ell/3$-subset Y, let $G(Y)$ denote the event that $\Pr_{\mathbf{r}'\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}') \in B^* \mid \mathbf{r}' \in R_Y\right] \leq N^{-1/10}$. Now we observe

$$\begin{aligned}
&\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[G(\mathbf{Y})\right] \\
&= \sum_{\gamma\in B^*} \Pr_{\mathbf{r}\sim\mathcal{N}_n^*}\left[\mathsf{size}(\mathbf{r})=\gamma\right]\cdot\frac{1}{\binom{\gamma}{\ell/3}}\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n^*}\left[\mathbf{r}\in R_Y \wedge G(Y) \mid \mathsf{size}(\mathbf{r})=\gamma\right] \\
&\leq \frac{2}{\binom{\gamma^*}{\ell/3}}\cdot\sum_{\gamma\in B^*} \Pr_{\mathbf{r}\sim\mathcal{N}_n^*}\left[\mathsf{size}(\mathbf{r})=\gamma\right]\cdot\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n^*}\left[\mathbf{r}\in R_Y \wedge G(Y) \mid \mathsf{size}(\mathbf{r})=\gamma\right] \\
&= \frac{2}{\binom{\gamma^*}{\ell/3}}\cdot\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n^*}\left[\mathbf{r}\in R_Y \wedge G(Y)\right] \\
&= \frac{2}{\binom{\gamma^*}{\ell/3}}\cdot\frac{1}{\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\mathsf{size}(\mathbf{r})\in B^*]}\cdot\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y \wedge G(Y) \wedge \mathsf{size}(\mathbf{r})\in B^*\right].
\end{aligned}$$

For each $\ell/3$-subset Y of H we have

$$\begin{aligned}
&\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge G(Y)\wedge \mathsf{size}(\mathbf{r})\in B^*\right]\\
&\quad=\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge G(Y)\right]\\
&\qquad\qquad\times\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r})\in B^*\mid \mathbf{r}\in R_Y\wedge G(Y)\right]\\
&\quad\le 2N^{-1/10}\cdot\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge G(Y)\right]\\
&\qquad\qquad\times\sum_{s\neq s^*}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r})\in B_s\mid \mathbf{r}\in R_Y\wedge G(Y)\right]\\
&\quad=2N^{-1/10}\sum_{s\neq s^*}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge G(Y)\wedge \mathsf{size}(\mathbf{r})\in B_s\right]\\
&\quad\le 2N^{-1/10}\sum_{s\neq s^*}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge \mathsf{size}(\mathbf{r})\in B_s\right].
\end{aligned}$$

By Lemma 4.15,

$$\sum_{Y\in\binom{H}{\ell/3}}\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}\in R_Y\wedge G(Y)\wedge \mathsf{size}(\mathbf{r})\in B^*\right]\le 2N^{-1/10}\sum_{s\neq s^*}\binom{\gamma_s}{\ell/3}\cdot\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r})\in B_s\right].$$

We then get

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[G(\mathbf{Y})\right]\le 4N^{-1/10}\cdot\frac{\sum_{s\neq s^*}\binom{\gamma_s}{\ell/3}\cdot\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\mathsf{size}(\mathbf{r})\in B_s]}{\binom{\gamma^*}{\ell/3}\cdot\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\mathsf{size}(\mathbf{r})\in B^*]}\le O\left(n^{1.1}N^{-1/10}\right),$$

using $\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\mathsf{size}(\mathbf{r})\in B_0]=o_n(1)$ and $U\le n^{1.1}$. ∎

With this, we have everything we need to prove the Gaussian Transfer Lemma

Proof. [Proof of the Gaussian Transfer Lemma] Let $G(Y)$ denote the event that $\Pr_{\mathbf{r}'\sim\mathcal{N}_n}\left[\mathsf{size}(\mathbf{r}')\in B^*\mid \mathbf{r}'\in R_{\mathbf{Y}}\right]\le N^{-1/10}$ and $G(Y)^c$ denote the complement of $G(Y)$. By Lemma 4.17,

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[G(\mathbf{Y})\right]\le O(N^{-1/20}).$$

Now observe that for a $(\ell/3)$-set Y in $G(Y)^c$ we have that

$$\begin{aligned}
\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[E(\mathbf{r}',Y)\mid\mathbf{Y}'=Y\right]&=\frac{\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[E(\mathbf{r}',Y),\mathbf{Y}'=Y\mid\mathbf{r}'\in R_Y\right]}{\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[\mathbf{Y}'=Y\mid\mathbf{r}'\in R_Y\right]}\\
&=\frac{\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[E(\mathbf{r}',Y)\mid\mathbf{r}'\in R_Y\right]\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[\mathbf{Y}'=Y\mid\mathbf{r}'\in R_Y,E(\mathbf{r}',Y)\right]}{\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[\mathbf{Y}'=Y\mid\mathbf{r}'\in R_Y\right]}\\
&\le 2\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}\left[E(\mathbf{r}',Y)\mid\mathbf{r}'\in R_Y\right]
\end{aligned}$$

where we used the fact that $\mathsf{size}(\mathbf{r}') \in B^*$.

$$\begin{aligned}&= 2\frac{\Pr_{\mathbf{r}'\sim\mathcal{N}_n}[E(\mathbf{r}', Y) \wedge \mathsf{size}(\mathbf{r}') \in B^* \mid \mathbf{r}' \in R_Y]}{\Pr_{\mathbf{r}'\sim\mathcal{N}_n}[\mathsf{size}(\mathbf{r}') \in B^* \mid \mathbf{r}' \in R_Y]}\\ &\leq 2N^{1/10}\Pr_{\mathbf{r}'\sim\mathcal{N}_n}[E(\mathbf{r}', Y) \in B^* \mid \mathbf{r}' \in R_Y]\end{aligned}$$

So,

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}[E(\mathbf{r}', \mathbf{Y}') \mid \mathbf{Y}' = \mathbf{Y}] \geq 2N^{1/10}\Pr_{\mathbf{r}'\sim\mathcal{N}_n}[E(\mathbf{r}', \mathbf{Y}')|\mathbf{r}' \in R_{\mathbf{Y}}]\right] \leq O(N^{-1/20})$$

as desired. ∎

4.2.2 Biased Hyperplanes Are Hard to Find Let $F_{r,Y}$ denote the failure set of close points $\{x \in H \setminus Y : |r^T x - t| \leq n\epsilon\}$. To prove the hidden slab lemma, we first prove the result for a random Gaussian slab. To do so, we start with a definition: we'll call a $w \in W_Y$ *bad with respect to* Y if it satisfies either

1. $w^T a \geq 0.2t$ for some $a \in H \setminus (Y \cup I_Y)$; or
2. $w^T a = t \pm n\epsilon$ for some $a \in I_Y$.

Otherwise w is good. To simplify notation, we say (Y, w) with $w \in W_Y$ is bad if w is bad with respect to Y and (Y, w) is good otherwise. We write $R_{Y,\mathsf{bad}}$ to denote the set of $r \in R_Y$ such that $\mathrm{proj}(r, S_Y)$ is bad and define $R_{Y,\mathsf{good}}$ similarly.

Intuitively, a $w \in R_{Y,\mathsf{bad}}$ would help the algorithm find a point close to the slab, as either an element in I_Y is already close to the slab or there's an element $a \in H \setminus (Y \cup I_Y)$ where $w^T a$ is potentially close to 1, which is a good candidate to be a point near the slab. We now show it's unlikely for us to get a bad set.

LEMMA 4.18. *For any k-subset Y of H, we have*

$$\Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r} \in R_{Y,\mathsf{bad}} \mid \mathbf{r} \in R_Y\right] \leq N^{-1/4}.$$

Proof. For each $z \in I_Y$ we use $z^\perp$ to denote the orthogonal component of z with respect to $\mathrm{span}(Y)$. We consider sampling $\mathbf{r}$ as sampling three independent multivariate Gaussians and taking the sum:

1. $\mathbf{r}_1$ as a Gaussian over $\mathrm{span}(Y)$ such that $\mathbf{r}_1^T y^i = t \pm \epsilon$ for every $i \in [k]$;
2. $\mathbf{r}_2$ as a Gaussian over $\mathrm{span}(z^\perp : z \in I_Y)$; and
3. $\mathbf{r}_3$ as a Gaussian over the orthogonal subspace of S_Y.

Note that whether $\mathbf{r}$ is bad with respect to Y or not only depends on $\mathbf{r}_1$ and $\mathbf{r}_2$. We consider the two parts and then take a union bound.

First for each $a \in H \setminus (Y \cup I_Y)$ we can write it as $a_1 + a_2 + a_3$ accordingly, and the condition we care about is $\mathbf{r}_1^T a_1 + \mathbf{r}_2^T a_2 \leq .2t$. Letting $a_1 = \sum_{i\in[k]} \alpha_i y^i$, the first term satisfies

$$\mathbf{r}_1^T a_1 = \sum_{i\in[k]} \alpha_i(t \pm \epsilon) \leq t\sum_{i\in[k]} \alpha_i + \epsilon\sum_{i\in[k]} |\alpha_i| \leq 0.1t + O(\sqrt{n}\epsilon).$$

The second term $\mathbf{r}_2^T a_2$ is greater than $0.09t$ with probability at most $1/N^4$ using our choice of t and the promise from the construction of h that $\|a_2\|_2 \le \sqrt{n}/100$. Given that there are no more than N such a's, $\mathbf{r}$ is bad because of the first item with probability at most $1/N^3$.

Next for each $z \in I_Y$, we have from the construction of h that $\|z^\perp\|_2 \ge \sqrt{n}/2$ and thus, after drawing $\mathbf{r}_1$, the $\mathbf{r}_2$ can have $\mathbf{r}_1^T z + \mathbf{r}_2^T z$ land in a window of length $O(n\epsilon)$ with probability at most $O(n\epsilon)$. Given that there are no more than k points in I_Y, $\mathbf{r}$ is bad because of the second item with probability at most $O(n^2\epsilon) \ll N^{-1/4}$. ∎

With this we can now prove the hidden slab lemma for Gaussian slabs.

LEMMA 4.19. *For any (adaptive) randomized algorithm* ALG *that makes at most* $Q \le N^{1/8}$ *queries and any* $\ell/3$ *set* Y *of* H*,*

$$\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\textsf{ALG queries a point in } F_{\mathbf{r},Y} | \mathbf{r} \in R_Y] \le O(N^{-1/8})$$

Proof. It suffices to show that

$$\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\textsf{ALG queries a point in } F_{\mathbf{r},Y} | \mathbf{r} \in R_{Y,\textsf{good}}] \le N^{-1/8}.$$

Note that $\mathbf{r} \in R_{Y,\textsf{good}}$ is only a condition on $\mathbf{w} = \text{proj}(\mathbf{r}, S_Y)$. We denote the corresponding set of good w by $W_{Y,\textsf{good}}$. Now let $w \in W_{y,\textsf{good}}$ and suppose $\text{proj}(\mathbf{r}, S_Y) = w$.

Note that we can assume that ALG is deterministic. In this case, observe that ALG with samples is just a decision tree of depth Q. Let $x^1, \ldots, x^Q$ be the following path of ALG. Set x^1 to be the root and repeat $Q-1$ times:

1. if the current $x^i \in I_Y$, go down to x^{i+1} according to $\text{sgn}(w^T x^i - t)$;
2. otherwise go down to x^{i+1} by pretending the answer to the query of x^i is -1.

We now have that since $w \in W_{Y,\textsf{good}}$,

$$\begin{aligned}
&\Pr_{\mathbf{r}\sim\mathcal{N}_n}[\textsf{ALG queries a point in } F_{\mathbf{r},Y} | \text{proj}(\mathbf{r}, I_Y) = w] \\
&\quad\le \Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\{x^1, ..., x^Q\} \cap F_{\mathbf{r},Y} \ne \emptyset \vee \textsf{ALG doesn't query } x^1, ..., x^Q | \text{proj}(\mathbf{r}, I_Y) = w\right] \\
&\quad\le \Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\exists x \in \{x^1, ..., x^Q\} \setminus (I_Y \cup Y) : \mathbf{r}^T x \ge t - n\epsilon | \text{proj}(\mathbf{r}, I_Y) = w\right] \\
&\quad\le \sum_i \Pr_{\mathbf{r}\sim\mathcal{N}_n}\left[\mathbf{r}^T x^i \ge t - n\epsilon | \text{proj}(\mathbf{r}, I_Y) = w\right] \\
&\quad\le \frac{Q}{N^{1/4}}
\end{aligned}$$

Using $Q \le N^{1/8}$ then proves the lemma. ∎

Using the Gaussian Transfer lemma, we can lift this to a statement for our distribution $\mathcal{R}$.

COROLLARY 4.2. *Let* $(\mathbf{r}, \mathbf{Y}) \sim \mathcal{R}$*, for any (adaptive) randomized algorithm* ALG *that receives* $\mathbf{Y}$ *and makes at most* $Q \le N^{1/8}$ *queries,*

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}[\textsf{ALG queries a point in } F_{\mathbf{r},Y}] \le O(N^{-1/40})$$

Proof. By Lemma 4.19 and the Gaussian Transfer Lemma, we have that

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}[\mathsf{ALG}(\mathbf{Y}) \text{ queries a point in } F_{\mathbf{r}',\mathbf{Y}} \mid \mathbf{Y}'=\mathbf{Y}] \geq \Omega(N^{-1/40})\right] \leq O(N^{-1/20})$$

Now let $E(Y)$ denote the event that $\Pr_{(\mathbf{r}',\mathbf{Y}')\sim\mathcal{R}}[\mathsf{ALG}(Y) \text{ queries a point in } F_{\mathbf{r}',\mathbf{Y}} \mid \mathbf{Y}'=Y] \leq O(N^{-1/40})$ and $E(Y)^c$ denote the complementary event. We then have

$$\begin{aligned}
&\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}\left[\mathsf{ALG}(\mathbf{Y}) \text{ queries a point in } F_{\mathbf{r},\mathbf{Y}}\right] \\
&\qquad = \Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}[\mathsf{ALG}(\mathbf{Y}) \text{ queries a point in } F_{\mathbf{r},\mathbf{Y}} \wedge E(\mathbf{Y})] \\
&\qquad\qquad + \Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}}[\mathsf{ALG}(\mathbf{Y}) \text{ queries a point in } F_{\mathbf{r},\mathbf{Y}} \wedge E(\mathbf{Y})^c] \\
&\qquad \leq O(N^{-1/40}) + O(N^{-1/20})
\end{aligned}$$

as desired. ∎

We can now boost this result to prove the Hidden Slab Lemma.

Proof. [Proof of the Hidden Slab Lemma] Suppose not. Given a set $(\mathbf{r},\mathbf{Y}) \sim \mathcal{R}$ we run ALG on a random d-subset $\mathbf{Y}_d$ of $\mathbf{Y}$. Note that $(\mathbf{r},\mathbf{Y}_d)$ is distributed according to $\mathcal{R}_d$. It then follows from Corollary 4.2 that the first element ALG queries from $F_{\mathbf{r},\mathbf{Y}_d}$ is in $\mathbf{Y}\setminus\mathbf{Y}_d$ with probability $\Omega(N^{-1/40})$. Letting the random seed of ALG be $\boldsymbol{\pi}$ we get

$$\Pr_{\mathbf{r},\mathbf{Y},\mathbf{Y}_d,\boldsymbol{\pi}}[\mathsf{ALG}(\mathbf{Y}_d,\boldsymbol{\pi}) \text{ first queries } F_{\mathbf{r},\mathbf{Y}_d} \text{ from } \mathbf{Y}\setminus\mathbf{Y}_d] = \Omega(N^{-1/40})$$

There is a fixed random seed π that achieves this same bound. Changing perspective, we can get the same distribution by sampling $\mathbf{r} \sim \mathcal{N}_n^*$ choosing a random subset $\mathbf{Y}_d$ and then random extending it to a $\ell/3$-subset $\mathbf{Y}$ of $\mathsf{slab}(r)$. But now for a fixed $(r,Y_d) \in \mathrm{supp}(\mathcal{R}_d)$ consider

$$\Pr_{\mathbf{Y}}[\mathsf{ALG}(Y_d,\pi) \text{ first queries } F_{r,Y_d} \text{ from } \mathbf{Y}\setminus Y_d | \mathbf{Y}_d = Y_d, \mathbf{r}=r]$$

But now the queries that $\mathsf{ALG}(Y_d,\pi)$ makes are fixed. So let x be the first query in F_{r,Y_d}. Since $\ell \leq n$ and $\mathsf{size}(r) \geq N^{1/4}$ we have that with high probability $\mathbf{Y}\setminus Y_d$ doesn't contain x. Namely,

$$\Pr_{\mathbf{Y}}[\mathsf{ALG}(Y_d,\pi) \text{ first queries } F_{r,Y_d} \text{ from } \mathbf{Y}\setminus Y_d | \mathbf{Y}_d = Y_d, \mathbf{r}=r] \leq O(N^{-1/5})$$

a contradiction. ∎

4.3 Proof of the Main Theorem Like in the case of $\mathbb{R}^n$ we will prove the result by showing how one can solve a support size distinction problem using a distribution-free LTF tester. Before we describe this reduction, we will need a perturbation lemma for our slabs:

LEMMA 4.20. *Let $S = \{x : |w^T x - b| \leq \epsilon\}$. Moreover, let $V = \{v^1,...,v^k\} \subseteq \{\pm 1\}^n \cap S$ be vectors such that the matrix M whose ith column is v^i has $\sigma_k(M) \geq c\sqrt{n}$. It then follows that for any function $\psi : \{v^1,...,v^k\} \to \{-1,1\}$ there exists a there exists an LTF $f : \mathcal{P} \to \{-1,1\}$ such that $f(x) = \psi(x)$ for all $x \in \{v^1,...,v^k\}$ and $f(x) = \mathrm{sgn}(w^T x - b)$ if $|w^T x - b| > \frac{2}{c}\epsilon\sqrt{n}$.*

Proof. Let $P : \mathbb{R}^n \to \mathbb{R}^k$ be the orthogonal projection onto span(V). Consider $(PM)^{-1}P$. Let u be a vector such that $u_i = 2\epsilon\psi(v^i)$ and consider the LTF f corresponding to $(w^T + u^T(PM)^{-1}P)x - b$. On the one hand note that

$$f(v^i) = \operatorname{sgn}(w^T v^i + u^T(PM)^{-1}PMe_i - b) = \operatorname{sgn}(w^T v^i - b + 2\epsilon f(v^i)) = f(v^i)$$

Conversely for any $x \in \{\pm 1\}^n$ we have that

$$|u^T(PM)^{-1}Px| \le \sigma_1((PM)^{-1})\|u\|_2\|x\|_2 = \frac{1}{\sigma_k(PM)}\|u\|_2\|x\|_2 \le \frac{2}{c}\sqrt{n}\epsilon.$$

Thus, our LTF f agrees with $\operatorname{sgn}(w^T x - b)$ when $|w^T x - b| > \frac{2}{c}\epsilon\sqrt{n}$. ∎

4.3.1 Simulation Let ALG be a randomized distribution-free tester with success probability 0.9 that uses at most $k \in [n^{.99}, O(n/\log^3(n))]$ samples. Set ℓ to be the smallest number such that $k < \mathsf{SSD}(400n, \frac{\ell}{3200n}, 1/2)$. Towards a contradiction, suppose that ALG makes at most $Q \le N^{1/8}$ (adaptive) queries. We will then show how we can use ALG to determine if an unknown distribution p over $[m]$ with $p(i) = \Omega(1/n)$ for all $i \in \operatorname{supp}(p)$ has small support ($|\operatorname{supp}(p)| \le \ell/8$) or large support ($|\operatorname{supp}(p)| \ge 200n$), which will contradict our choice of ℓ.

Let p be an distribution over $[m]$, *unknown* to the player, that falls into one of the two cases. For notational convenience, we let a be the unknown bit that is set to $a = 1$ if p has small support and $a = 0$ if p has large support. The player starts by drawing the following four objects: (1) $\mathbf{r} \sim \mathcal{N}_n^*$, which defines an $\mathbf{r}$-slab; (2) a bijection $\boldsymbol{\phi} : [400n] \to \mathsf{slab}(r)$ drawn uniformly at random; (3) a map $\boldsymbol{\psi} : \{\pm 1\}^n \to \{\pm 1\}$ drawn uniformly at random and independently from $\boldsymbol{\phi}$ and $\mathbf{r}$; (4) a random string $\boldsymbol{\pi}$ for ALG drawn independently from $\mathbf{r}, \boldsymbol{\phi}, \boldsymbol{\psi}$.

Together with the hidden distribution p, they define the following Boolean function $f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}} : \{\pm 1\}^n \to \{\pm 1\}$ and distribution $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$ over $\{\pm 1\}^n$ as follows (we use f for convenience):

1. The function f is defined as follows when p has large support. For each $i \in \operatorname{supp}(p)$, $f(\boldsymbol{\phi}(i)) = \boldsymbol{\psi}(\boldsymbol{\phi}(i))$; for every point $x \ne \boldsymbol{\phi}(i)$ for any $i \in \operatorname{supp}(p)$, we set $f(x)$ as follows. If $x \notin H$, set $f(x)$ to be the sign of x with respect to H; if $x \in H$, set $f(x) = \operatorname{sgn}(\mathbf{r}^T x - t)$.

2. The function f is defined as follows when p has small support. If $x \notin H$, we set $f(x)$ according to the sign of x with respect to H; if $x \in H$, we set the values according to the LTF from Lemma 4.20 e.g. such that $f(\boldsymbol{\phi}(i)) = \boldsymbol{\psi}(\boldsymbol{\phi}(i))$ for all $i \in \operatorname{supp}(p)$ and f agrees with $\operatorname{sgn}(\mathbf{r}^T x - t)$ for all $x \in H$ with $|\mathbf{r}^T x - t| \ge n\epsilon$.

3. In both cases let $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$ be the distribution supported over $\{\boldsymbol{\phi}(1), \ldots, \boldsymbol{\phi}(m)\}$: the probability of $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$ on $\boldsymbol{\phi}(i)$ is set to be the same as that of p on i for each $i \in [m]$.

Of course the player has no way to construct $f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}$ by herself since she does not know p (she does not even know whether p has large or small support). Before continuing to describe the simulation, we record the following lemma about $(f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}, \mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}})$:

LEMMA 4.21. *If p has small support, then $f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}$ is always a halfspace.*

If p has large support, then $f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}$ is $\Omega(1)$-far from halfspaces with respect to $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$ with probability at least $1 - o_n(1)$ (over the randomness of $\mathbf{r}, \boldsymbol{\phi}$ and $\boldsymbol{\psi}$).

Proof. If p has large support than the claim follows from Lemma 1.2. If p has small support then let the LTF corresponding to h be $\mathrm{sgn}(w^T x - b)$ and the LTF from Lemma 4.20 be $\mathrm{sgn}((\mathbf{r}')^T x - t')$. Then we note that

$$f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}(x) = \mathrm{sgn}\left(\frac{2}{\min_{x\in\{\pm1\}^n : w^T x \neq b} |w^T x - b|}(w^T x - b) + \frac{1}{\max_{x\{\pm1\}^n} |(\mathbf{r}')^T x - t'|}((\mathbf{r}')^T x - t')\right).$$

So f is indeed a halfspace in this case. ∎

After drawing $\mathbf{r}, \boldsymbol{\phi}$ and $\boldsymbol{\psi}$, the player calls ALG to work on $(f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}, \mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}})$ (even though she does not really know the pair). ALG asks for a sequence of k samples from $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$. For this the player turns to the sampling oracle of p and asks for a sequence of k samples $\mathbf{J} = (\mathbf{j}^1, \ldots, \mathbf{j}^k)$ from p. It then sends the following sequence of k points to ALG:

$$\boldsymbol{\phi}(\mathbf{J}) := \big(\boldsymbol{\phi}(\mathbf{j}^1), \ldots, \boldsymbol{\phi}(\mathbf{j}^k)\big)$$

It is clear that the $\boldsymbol{\phi}(\mathbf{J})$ sent to ALG is distributed the same as a sequence of k samples from $\mathcal{D}_{p,\mathbf{r},\boldsymbol{\phi}}$. Now by Lemma 4.21 we observe

$$\Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi},\mathbf{J},\boldsymbol{\pi}}\Big[\mathit{ALG}\big(\boldsymbol{\phi}(\mathbf{J}), \boldsymbol{\pi}; f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}\big) = a\Big] \geq 0.9 - o_n(1).$$

Before continuing the simulation, let's define $\mathbf{Y}$: $\mathbf{Y} = \{\boldsymbol{\phi}(\mathbf{j}^i) : i \in [m]\}$. We observe that $(\mathbf{r}, \mathbf{Y})$ as in the probability space described above is distributed exactly as $\mathcal{R}_d$ for some $d \leq k$.

After receiving $\boldsymbol{\phi}(\mathbf{J})$ and $\boldsymbol{\pi}$, ALG will make Q (adaptive) queries. To answer these queries, we use $\mathbf{r}, \mathbf{Y}$ and $\boldsymbol{\psi}$ (but not p) to define the following Boolean function $g_{\mathbf{r},\mathbf{Y},\boldsymbol{\psi}} : \{\pm1\}^n \to \{\pm1\}$, which the player has in hand:

1. If x is not in H, set $g(x)$ to be the sign of x with respect to H;
2. If $x \in H$ and $\mathbf{r}^T x \notin [t - n\epsilon, t + n\epsilon]$, set $g(x) = \mathrm{sgn}(\mathbf{r}^T x - t)$;
3. If $x \in \mathbf{Y}$ (note by construction $\mathbf{Y} \subseteq \mathsf{slab}(\mathbf{r})$ so $r^T x = t \pm n\epsilon$), set $g(x) = \boldsymbol{\psi}(x)$; and
4. If none above applies ($x \in H$, $\mathbf{r}^T x = t \pm n\epsilon$ and $x \notin \mathbf{Y}$), set $g(x) = -1$. Recall that this is the set $F_{\mathbf{r},\mathbf{Y}}$.

Now by our definition of f, we have that $g_{\mathbf{r},\mathbf{Y},\boldsymbol{\psi}}(x) = f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}(x)$ for all $x \notin F_{\mathbf{r},\mathbf{Y}}$. So the strategy of the player is just to answer all queries of ALG using $g_{\mathbf{r},\mathbf{Y},\boldsymbol{\psi}}$, which she knows, and outputs the same $\{0,1\}$-answer as ALG when the simulation ends. As a result, we have

$$\begin{aligned}&\Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi},\mathbf{J},\boldsymbol{\pi}}\big[\text{player outputs } a\big]\\ &\quad\geq \Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi},\mathbf{J},\boldsymbol{\pi}}\Big[\mathsf{ALG}\big(\boldsymbol{\phi}(\mathbf{J}), \boldsymbol{\pi}; f_{p,\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi}}\big) = a\Big] - \Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi},\mathbf{J},\boldsymbol{\pi}}\Big[\mathsf{ALG}\big(\boldsymbol{\phi}(\mathbf{J}), \boldsymbol{\pi}; g_{\mathbf{r},\mathbf{J},\boldsymbol{\psi}}\big) \text{ queries } F_{\mathbf{r},\mathbf{Y}}\Big].\end{aligned}$$

We now argue that

Lemma 4.22.

$$\Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\psi},\mathbf{J},\boldsymbol{\pi}}\Big[\mathit{ALG}\big(\boldsymbol{\phi}(\mathbf{J}), \boldsymbol{\pi}; g_{\mathbf{r},\mathbf{Y},\boldsymbol{\psi}}\big) \textit{ queries } F_{\mathbf{r},\mathbf{Y}}\Big] = o_n(1)$$

Proof. Indeed, suppose that ALG could query a point in $F_{\mathbf{r},\mathbf{Y}}$ with probability $\Omega(1)$. Since $\mathbf{J}$ and $\boldsymbol{\psi}$ are independent from $\mathbf{r}, \boldsymbol{\phi}, \boldsymbol{\pi}$ it follows that there a fixed set of d distinct samples $J = (j^1, ..., j^k)$ and a fixed ψ such that

$$\Pr_{\mathbf{r},\boldsymbol{\phi},\boldsymbol{\pi}}\left[\mathsf{ALG}\big(\boldsymbol{\phi}(J), \boldsymbol{\pi}; g_{\mathbf{r},\mathbf{Y},\psi}\big) \text{ queries } F_{\mathbf{r},\mathbf{Y}}\right] = \Omega(1)$$

Now we consider a new algorithm ALG' which given $\mathbf{Y}$, $\boldsymbol{\pi}$, and a random bijection $\boldsymbol{\phi}'$ from $\{j^1, ..., j^k\}$ to $\mathbf{Y}$ and simulates $\mathsf{ALG}(\boldsymbol{\phi}'(J), \boldsymbol{\pi})$. If ALG ever tries to query a point in $\mathbf{Y}$, ALG' simply answers using ψ rather than actually making the query. Similarly, for queries not in H, ALG' avoids the query and simply gives ALG the sign of x with respect to H. Since $\mathbf{r}, \mathbf{Y}, \boldsymbol{\phi}'(J)$ is distributed identically to $\mathbf{r}, \mathbf{Y}, \boldsymbol{\phi}(J)$ this is a faithful simulation of $\mathsf{ALG}\big(\boldsymbol{\phi}(J), \boldsymbol{\pi}; g_{\mathbf{r},\mathbf{Y},\psi}\big)$ so long as no point in $F_{\mathbf{r},\mathbf{Y}}$ is queried. Thus,

$$\Pr_{(\mathbf{r},\mathbf{Y})\sim\mathcal{R}_d,\boldsymbol{\phi}',\boldsymbol{\pi}}\left[\mathsf{ALG}'\big(\mathbf{Y}, \boldsymbol{\pi}, \boldsymbol{\phi}'\big) \text{ queries } F_{\mathbf{r},\mathbf{Y}}\right] = \Omega(1)$$

but this contradicts the Hidden Slab Lemma. ∎

Thus, the player succeeds with probability $0.9 - o_n(1)$, a contradiction with how we defined ℓ. So any such algorithm must make at least $N^{1/8}$ queries. To finish the proof, note that by Theorem 1.3 we have that $\mathsf{SSD}(400n, \frac{\ell}{3200n}, 1/2) = \Omega\left(\frac{\ell^2}{n\log(n)}\right)$ and thus $\ell = \tilde{O}(\sqrt{nk})$. The good hyperplane lemma then implies that $N^{1/8} = \exp(\tilde{\Omega}(\sqrt{n/k}))$.

References

[1] N. Ailon and B. Chazelle. Information theory in property testing and monotonicity testing in higher dimension. *Information and Computation*, 204(11):1704–1717, 2006.

[2] Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active property testing. In *IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 21–30, 2012.

[3] G. Bennett, L.E. Dor, W.B. Johnson, V. Goodman, and C.M. Newman. On uncomplemented subspaces. *Israel Journal of Mathematics*, 26(2), 1977.

[4] Eric Blais, Renato Ferreira Pinto Jr., and Nathaniel Harms. VC dimension and distribution-free sample-based testing. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 504–517, 2021

[5] Xi Chen and Jinyu Xie. Tight bounds for the distribution-free testing of monotone conjunctions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 54–71. SIAM, 2016.

[6] E. Dolev and D. Ron. Distribution-free testing for monomials with a sublinear number of queries. *Theory of Computing*, 7(1):155–176, 2011.

[7] Xi Chen and Jinyu Xie. Tight bounds for the distribution-free testing of monotone conjunctions. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 54–71. SIAM, 2016.

[8] E. Dolev and D. Ron. Distribution-free testing for monomials with a sublinear number of queries. *Theory of Computing*, 7(1):155–176, 2011.

[9] Rogers Epstein and Sandeep Silwal. Property testing of LP-type problems. In *International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, page 98:1–98:18, 2020.

[10] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, July 1998.

[11] D. Glasner and R. Servedio. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing*, 5(10):191–216, 2009.
[12] Dana Glasner and Rocco A Servedio. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing*, 5(1):191–216, 2009.
[13] Nathaniel Harms. Testing halfspaces over rotation-invariant distributions. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 694–713, 2019.
[14] S. Halevy and E. Kushilevitz. Distribution-free property-testing. *SIAM Journal on Computing*, 37(4):1107–1138, 2007.
[15] S. Halevy and E. Kushilevitz. Distribution-free connectivity testing for sparse graphs. *Algorithmica*, 51(1):24–48, 2008.
[16] S. Halevy and E. Kushilevitz. Testing monotonicity over graph products. *Random Structures & Algorithms*, 33(1):44–67, 2008.
[17] Jeff Kahn, János Komlós, and Endre Szemerédi. On the probability that a random ±1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240, 1995.
[18] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.
[19] Kevin Matulef, Ryan O'Donnell, Ronitt Rubinfeld, and Rocco A Servedio. Testing halfspaces. *SIAM Journal on Computing*, 39(5):2004–2047, 2010.
[20] Andrew M Odlyzko. On subspaces spanned by random selections of ±1 vectors. *Journal of combinatorial theory, Series A*, 47(1):124–133, 1988.
[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
[22] Terence Tao and Van Vu. On random ±1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006.
[23] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
[24] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM Symposium on Theory of Computing*, pages 685–694, 2011.
[25] Gregory Valiant and Paul Valiant. The power of linear estimators. In *IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412, 2011.
[26] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Annals of Statistics*, 47(2):857–883, 2019.

A Missing Proofs

Proof. [Proof of Lemma 1.1] It suffices to prove the claim when $k = n + 1$. Applying an affine transformation we can assume that $v_{n+1} = 0$ and $v_i = e_i$ for $i \leq n$. Fix an arbitrary function f. We then let $w \in \mathbb{R}^n$ with $w_i = f(v_i)$ and observe $f(x) = \mathrm{sgn}(\langle w, x\rangle + f(v_{n+1})/2)$. ∎

Proof. [Proof of Lemma 1.2] Fix an LTF g. For any point v_i, the probability that $g(v_i)$ agrees with $f(v_i)$ is $1/2$. So by Chernoff bounds the probability that f and g disagree on fewer than $k/4$ points is at most $e^{-k/16}$. On the other hand, using VC dimension and the Sauer-Shelah Lemma, there are $2^{kH\left(\frac{n+1}{k}\right)}$ LTFs. A union bound shows that the probability that the coloring is close to an LTF is

$$2^{kH\left(\frac{n+1}{k}\right)-k/(16\ln(2))} = o_n(1)$$

when $k \geq 100(n+1)$. Thus with high probability, every LTF disagrees with f on $k/4$ points. Since each point appears with probability $\Omega(1/k)$ it follows they are $\Omega(1)$-far under $\mathcal{D}$. ∎

Proof. [Proof of Lemma 4.4] For completeness, we include a proof. We first note that we can write S as $\{v : Wv = \theta\}$ for some $W \in \mathbb{R}^{(n-k)\times n}$ with rank $n-k$ and $\theta \in \mathbb{R}^{n-k}$. Moreover, we can assume

without loss of generality that $W = (I_{n-k} \mid R)$ for some $R \in \mathbb{R}^{(n-k)\times k}$ as applying row operations and permuting columns does not change the number of points in the affine space. Now for a subset $\{i_1, ..., i_\ell\} = I \subseteq [n-k]$ we define $f_I : \{\pm 1\}^n \to \{\pm 1\}^n$ as the function that flips the $i_1, i_2, ..., i_\ell$ bits. We observe that if $x \in \{\pm 1\}^n \cap S$ then $f_I(x)$ satisfies $(Wf_I(x))_i \neq \theta_i$ for $i \in I$ and $(Wf_I)_j = \theta_j$ for $j \notin I$. So it follows that for $I, J \subseteq [n-k]$ and $I \neq J$ and $x, y \in \{\pm 1\}^n \cap S$, $f_I(x) \neq f_J(y)$. So, $\bigsqcup_{I\subseteq[n-k]} f_I(\{\pm 1\}^n \cap S) \subseteq \{\pm 1\}^n$. Since each f_I is injective, taking the cardinality of both sides gives $2^{n-k}|S| \leq 2^n$ as desired. ∎