# What is the Value of Data? on Mathematical Methods for Data Quality Estimation

**Netanel Raviv**[⋆], **Siddharth Jain**[†], and **Jehoshua Bruck**[†]
[⋆]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis 63130, MO, USA
[†]Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA

### Abstract

Data is one of the most important assets of the information age, and its societal impact is undisputed. Yet, rigorous methods of assessing the quality of data are lacking. In this paper, we propose a formal definition for the quality of a given dataset. We assess a dataset's quality by a quantity we call the *expected diameter*, which measures the expected disagreement between two randomly chosen hypotheses that explain it, and has recently found applications in active learning. We focus on Boolean hyperplanes, and utilize a collection of Fourier analytic, algebraic, and probabilistic methods to come up with theoretical guarantees and practical solutions for the computation of the expected diameter. We also study the behaviour of the expected diameter on algebraically structured datasets, conduct experiments that validate this notion of quality, and demonstrate the feasibility of our techniques.

## I. Introduction

Recent advances in machine learning (ML) have revolutionized our society in more ways than one. Yet, ML techniques are highly prone to *garbage-in-garbage-out* issues, where processing uninformative, repetitive, or noisy data leads to nonsensical conclusions. However, even in the noiseless setting, by merely observing a large dataset it is hard to evaluate how informative it is, and what would be the accuracy of an arbitrary model that explains it over unseen data points.

Since the ML paradigm is inherently heuristic, it is essential to develop methods to rigorously determine the value of datasets; such methods can be used to explain the success or failure of one learning method with respect to another, and to determine the intrinsic value of a given dataset. In particular, it is natural to aspire to a *universal* notion of value, one that is devoid of the contextual use of the data, and does not pertain to any particular learning algorithm.

A few approaches exist in the literature, that aim towards assessment of a *specific* learning method with respect to the dataset it operates on. For example, many learning algorithms are analyzed with respect to the *size* of a randomly chosen dataset on which they operate [9], a measure called *sample complexity*, that prioritizes quantity over quality. However, real-world datasets are rarely purely random and are often laboriously collected (e.g., in medical research). Moreover, quantity does not necessarily correlate with quality, as one can easily come up with two datasets of equal size, whose respective sets of consistent hypotheses (i.e., that explain the data well) are substantially different in terms of their variance[1]. Hence, the size of a given dataset does not always reflect its value.

An additional commonly used notion of data quality is its *margin*, i.e., the minimum Euclidean distance between the convex hulls of the positive and the negative points, (e.g., for the well-known SVM method [9, Sec. 15]). However, one can similarly construct two datasets with identical margins, and substantially different sets of consistent hypotheses, and even such that the SVM method produces the same output (see Figure 1).

In this paper we propose a method for assessing the intrinsic quality of a given dataset $\mathcal{D}$. For the reasons discussed above, our aim is to provide methods that are *performance-independent*, i.e., that do not rely on finding a consistent hypothesis and validating its performance over unseen data points. Instead, we provide a measure that explains the performance of *any* hypothesis from a given hypotheses class, regardless of the learning algorithm that is used to obtain it.

[1] More generally, the *No-Free-Lunch* theorem [9, Thm. 5.1] roughly states that for every learning method there exists a dataset on which it fails.

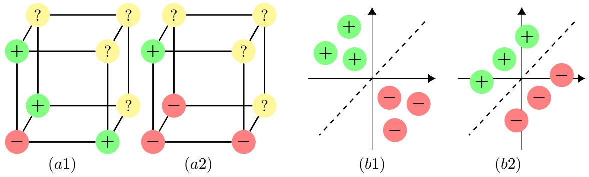


Fig. 1: Datasets $(a1)$ and $(a2)$, that reside in $\{\pm 1\}^3$, are both of size $4$. However, every two affine hyperplanes that classify $(a1)$ correctly (i.e., agree on all green and red points) agree on all the remaining unknown (yellow) points, whereas some affine hyperplanes that classify $(a2)$ correctly do not. Hence, $(a1)$ is intuitively more valuable than $(a2)$. An SVM algorithm on datasets $(b1)$ and $(b2)$ in $\mathbb{R}^2$ yields identical separators, given as dashed lines. However, $(b2)$ is clearly more informative than $(b1)$ due to smaller variability of the consistent hypotheses, even though their margins are identical.

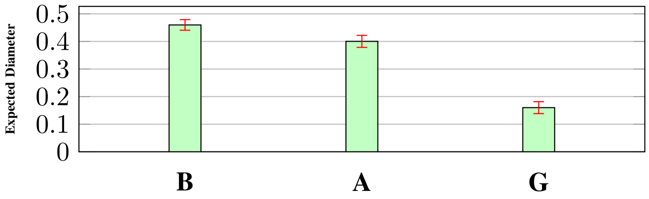


Fig. 2: The mean and standard deviation of the expected diameter of randomly generated *bad* (**B**), *arbitrary* (**A**), and *good* (**G**) datasets, see Section VIII.

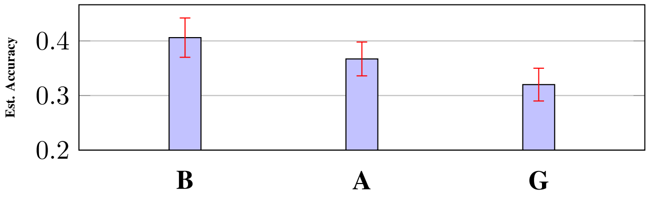


Fig. 3: The mean and standard deviation of the estimated accuracy of perceptron with respect to a uniform prior, on the same datasets as in Figure 2.

Specifically, with respect to a set of hypotheses $\mathcal{H}$ that agree on a dataset $\mathcal{D}$, we define the quality of $\mathcal{D}$ as the expected disagreement between two random members of $\mathcal{H}$, a property that we call *expected diameter*. Focusing on expected disagreement between randomly chosen hypotheses (rather than, say, on the maximum disagreement), encapsulates the following meaningful aspects of our goal.

First, since all hypotheses in $\mathcal{H}$ explain the dataset equally well, we naturally associate a probability distribution on $\mathcal{H}$, often called a *prior*, which reflects the user's belief regarding their likelihood. Second, as most classic and contemporary ML techniques employ randomness in one way or another, the output of a random ML algorithm can also be viewed as a probability distribution on $\mathcal{H}$. The expected diameter captures the tangent point of these two concepts; it measures the expected disagreement between a hypothesis chosen according to the prior on $\mathcal{H}$, and one that is chosen according to learning algorithm. To keep the expected diameter oblivious to any subjective prior and to any particular learning algorithm, we consider both distributions *uniform* on $\mathcal{H}$. The precise nature of this uniformity, alongside a formal description of the above intuition, will be given shortly in Section II.

Before summarizing our contributions, we demonstrate experimentally that the expected diameter indeed predicts the success of learning. Figure 2 presents the mean and standard deviation of the expected diameter on 300 randomly generated datasets of *identical* size and dimension, out of which 100 are *bad* (**B**), i.e., contain redundant information, 100 are *arbitrary* (**A**), i.e., chosen entirely at random, and 100 are *good* (**G**), i.e., contain many informative pairs of data points. In Figure 3 we used *the same* datasets as in Figure 2 and estimated the distance between a hypothesis chosen according to a uniform prior (representing the "true" function), and a hypothesis produced by a randomized perceptron algorithm; it is evident from these experiments that lower expected diameter correlates with better accuracy. Formal description and technical details are given in Section VIII.

*Our Contribution:* We focus on Boolean datasets and the hypotheses class of homogeneous linear separators; a class that is also known as *halfspaces* or *linear threshold functions*, and encapsulates many other classes by a set of known reductions [3, Table I]. We begin by presenting an intriguing connection to Fourier analysis of Boolean functions in the form of a polynomial algebraic algorithm for approximating the expected diameter (Section IV). This algorithm applies to any distribution on $\mathcal{H}$, but is most useful for ones that are in some sense "short", which includes the uniform ones. A surprising corollary of this part is that the expected diameter can be approximated efficiently *without* the ability to randomly sample a hypothesis according to the underlying probability distribution on $\mathcal{H}$; an appealing feature since sampling is often hard or unknown.

Albeit being polynomial, the complexity of this algorithm is rather prohibitive, and hence in Section V we focus on a particular important case of a samplable distribution on $\mathcal{H}$. For this distribution we present two different probabilistic algorithms, and analyze their theoretical complexity and probabilistic guarantees. We continue in Section VI with a structural theorem, which shows that datasets with a certain algebraic structure possess a convenient uniformity of the expected diameter. This uniformity is formulated by using tools from Boolean algebra, group theory, and graph theory, and is independent of any particular way of computing the expected diameter. The case of data over the real-number field, which is somewhat easier to handle, is discussed in Section VII. We conclude the paper in Section VIII by demonstrating some of our methods experimentally. Formal definitions and mathematical background are given shortly in Section II.

## II. Preliminaries

For a given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \{\pm 1\}^n, y_i \in \{\pm 1\}, i \in [k]\}$ let $\mathcal{X} \triangleq \{\mathbf{x}_i\}_{i=1}^k$, and let $\mathcal{H} = \mathcal{H}(\mathcal{D})$ be the set of all homogeneous halfspaces $h : \{\pm 1\}^n \to \{\pm 1\}$, $h(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x})$ for some $\mathbf{w} \in \mathbb{R}^n$, such that $h(\mathbf{x}_i) = y_i$ for every $i \in [k]$. We call $\mathcal{H}$ the set of *consistent hypotheses*, and occasionally abuse the notation by using $\mathcal{H}$ to denote an unspecified probability distribution over the set of consistent hypotheses. For every pair of halfspaces $h_1$ and $h_2$ define their respective distance as $d(h_1, h_2) = \frac{1}{2^n} \sum_{\mathbf{x} \in \{\pm 1\}^n} \frac{1 - h_1(\mathbf{x}) h_2(\mathbf{x})}{2}$, which amounts to the fraction of $\mathbf{x}$'s on which $h_1$ and $h_2$ disagree.

We measure the quality of $\mathcal{D}$ according to its *expected diameter*, defined as follows.

**Definition 1.** *For a given dataset $\mathcal{D}$ and a given probability distribution $\mathcal{H}$ over its set of consistent hypotheses, the* expected diameter *of $\mathcal{D}$ is $\mathbb{E}_{h_1, h_2 \sim \mathcal{H}}\, d(h_1, h_2)$. The dependence on $\mathcal{H}$ is omitted if unspecified or clear from the context.*

The aim of this paper is to devise techniques for computing the expected diameter of a given dataset $\mathcal{D}$, which is a real number between 0 and $\frac{1}{2}$ (see Appendix A). We argue that the most suitable probability distribution for data quality estimation is the *uniform distribution* $\mathcal{H}_{uni}$, defined as $\Pr(h) = 1/|\mathcal{H}|$ for every $h \in \mathcal{H}$ (see Subsection II-A). Results for $\mathcal{H}_{uni}$ (and more broadly, any distribution $\mathcal{H}$ such that $c(\mathcal{H}) \triangleq |\mathcal{H}| \sum_{h \in \mathcal{H}} \Pr(h)^2$ is small) are given in Section IV by using Fourier analysis. Due to prohibitive (albeit polynomial) complexity in Section IV, we study a surrogate distribution $\mathcal{H}_{vol}$, that we call the *volume distribution*, in Section V. To define $\mathcal{H}_{vol}$, notice that the discrete set $\mathcal{H}$ naturally admits a continuous one (often called *the version space*)

$$\mathcal{V} \triangleq \{\mathbf{w} \in \mathbb{R}^n | y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 0 \ \forall i \in [k] \text{ and } \|\mathbf{w}\|_2 \leq 1\},$$

which is partitioned to $|\mathcal{H}|$ parts $\mathcal{V}_h = \{\mathbf{w} \in \mathcal{V} | \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}) = h(\mathbf{x}) \text{ for all } \mathbf{x} \in \{\pm 1\}^n\}$ for $h \in \mathcal{H}$. Hence, in $\mathcal{H}_{vol}$ we define $\Pr(h) = \operatorname{Vol}(\mathcal{V}_h) / \operatorname{Vol}(\mathcal{V})$ for every $h \in \mathcal{H}$. The volume distribution is (approximately) samplable by using algorithms for sampling from convex bodies (see below). Namely, one can sample $\mathbf{w} \in \mathcal{V}$ (approximately) uniformly at random, and output $\operatorname{sign}(\mathbf{w} \cdot \mathbf{x})$. The authors are not aware of any efficient algorithm[2] to sample $h \in \mathcal{H}_{uni}$, but nevertheless, we are able to estimate the expected diameter under $\mathcal{H}_{uni}$ without sampling.

We focus on probabilistic algorithms, that for some $\epsilon, \eta > 0$, guarantee at most $\epsilon$ additive deviation from the expected diameter with probability at least $1 - \eta$. In what follows we use the standard notation $[n] \triangleq \{1, 2, \ldots, n\}$, we use lowercase bold letters to denote vectors and regular lowercase letters to denote scalars or functions (e.g., $\mathbf{x} = (x_1, \ldots, x_n)$).

### A. Why Expected Diameter?

Clearly, a natural measure for the success of a learning algorithm is $d(f, g)$, where $f$ is the "true" function, and $g$ is the output of the algorithm. However, in reality the existence of a "true" function is merely an assumption (known as the *realizability assumption* [9, Def. 2.1]), and hence one normally seeks a "most probable" $f$, a notion which requires probabilistic assumptions on the data gathering process. For datasets that might contain significant bias, one can only assume that *all* $f$'s that classify the dataset correctly are equally likely.

On the other hand, choosing a learning method, even for a given hypothesis class, is a formidable task for many data scientists. For example, one may choose different types of gradient descent, loss functions, and regularization parameters, or randomize the choice of hyperparameters, and end up with a different function $g$. Further, algorithms which process the dataset sequentially, such as the well-known perceptron, are susceptible to the order by which the datapoints are processed. Since we aim for the most uniform notion of data quality, we coalesce all these aspects into one by viewing $g$ as chosen uniformly at random.

Specifically, the accuracy of (a given run of) any probabilistic learning algorithm $A$ on $\mathcal{D}$ is naturally measured by $d(f, g)$, where $f$ is the "true" function by which $\mathcal{D}$ is labeled and $A(\mathcal{D}) = g$. Therefore, letting $\mathcal{H}_{prior}$ be the prior at hand, and $\mathcal{H}_A$ be the probability distribution on $\mathcal{H}$ that is induced by $A$, the expected accuracy of $A$ equals $\mathbb{E}_{h_1 \sim \mathcal{H}_{prior}, h_2 \sim \mathcal{H}_A}\, d(h_1, h_2)$. Since our aim is to obtain a *universal* notion of data quality, we consider both $\mathcal{H}_{prior}$ and $\mathcal{H}_A$ as some general distribution $\mathcal{H}$, and measure the quality of $\mathcal{D}$ by using the expected diameter according to that $\mathcal{H}$.

As explained above, for technical reasons we study two different interpretations of a "uniform" distribution over $\mathcal{H}$. In $\mathcal{H}_{vol}$, the weight vector $\mathbf{w}$ is chosen according to a *continuous* uniform distribution on the version space. On the contrary, $\mathcal{H}_{uni}$ is a *discrete* uniform distribution on the (finite) set $\mathcal{H}$, i.e., where every hypothesis is chosen with probability $1/|\mathcal{H}|$. Specializing/generalizing this question to particular priors, particular learning algorithms, non-separable datasets, different hypotheses classes, or hypotheses that do not classify $\mathcal{D}$ perfectly, are left for future research.

[2] Of course, one can get $\mathcal{H}_{uni}$ by rejection sampling, but the resulting complexity is super-exponential.

### B. Previous Work

We first note that independently of this work, a similar quantity appeared in [12] for applications in active learning, but was not studied in depth. Extremal questions of similar flavor appeared in [7, Sec. 8], which studies the notion of *specifying sets*. For a given class $H$ of Boolean functions and a function $f \in H$, a specifying set for $f$ in $H$ is a dataset such that $f$ is the unique function in $H$ which classifies it correctly. It is readily verified that a dataset is a specifying set if and only if its expected diameter is zero.

Our notion for the value of data is not to be confused with similar terms in the data acquisition literature (e.g., [1], [2]). In this line of works, data is acquired from individuals that fix its price arbitrarily (normally as a function of their personal perception of privacy infringement), and no rigorous notion of data quality is discussed. Finally, [13] presents a novel learning framework that captures inter-dependence between data points; this idea is substantially different from ours, but it can also be viewed as relating to data quality.

### C. Mathematical Background

*Fourier Analysis of Boolean Functions [5] (Section IV):* Every Boolean function $f : \{\pm 1\}^n \to \mathbb{R}$ can be represented as a linear combination over $\mathbb{R}$ of the functions $\{\chi_S(\mathbf{x})\}_{S\subseteq[n]}$, where $\chi_S(\mathbf{x}) = \prod_{j\in S} x_j$ for every $S \subseteq [n]$. The coefficient of $\chi_S(\mathbf{x})$ in this linear combination is called the *Fourier coefficient* of $f$ at $S$, and it is denoted by $\hat{f}(S)$. The collection of all Fourier coefficients of $f$ is called the *Fourier spectrum* of $f$. Each Fourier coefficient $\hat{f}(S)$ equals the inner product between $f$ and $\chi_S$, defined as $\langle f, \chi_S\rangle \triangleq \mathbb{E}_{\mathbf{x}} f(\mathbf{x})\chi_S(\mathbf{x})$, where $\mathbf{x}$ is chosen uniformly at random. For any two Boolean functions $f$ and $g$, their inner product can be computed by the inner product (in the usual sense) of their respective Fourier spectra, a result known as Plancheral's identity (or Parseval's identity if $f = g$): $\langle f, g\rangle = \sum_{s\subseteq[n]} \hat{f}(S)\hat{g}(S)$. Finally, an attractive feature of Fourier analytic methods on halfspaces is that their largest Fourier coefficients appear on lower degree terms, a property known as *Fourier concentration*, and given in the following lemma.

**Lemma 1.** *[8] For an integer $a \geq 0$ and a function $f : \{\pm 1\}^n \to \mathbb{R}$, let $W^{\geq a}[f] \triangleq \sum_{|S|\geq a} \hat{f}(S)^2$. For every $0 < b < 1$, every halfspace $f$ satisfies that $W^{\geq a}[f] \leq b$, where $a = O(1/b^2)$.*

*Random Sampling from Convex Bodies (Section V and Section VIII):* In the sequel we require an algorithm that is given a set of constraints that define a convex body $\mathcal{B} \subseteq \mathbb{R}^n$, and returns a point which is chosen uniformly at random from it. In particular, we focus on the *Hit-and-Run* (H&R) algorithm [11], which works well in theory [4] as well as in our experimental results (Section VIII). This algorithm begins with a "sufficiently random" starting point $\mathbf{v}_0$, chooses a random direction $\mathbf{l} \in \mathbb{R}^n$, chooses a uniformly random point $\mathbf{v}_1$ from the chord $\{\mathbf{v} + t\mathbf{l} | t \in \mathbb{R}\} \cap \mathcal{V}$, and repeats the process. After $O^*(n^3 \frac{1}{\epsilon^2} \ln(\frac{2}{\epsilon}))$ of these steps, it is known that the resulting distribution is $\epsilon$-close to uniform, but in practice convergence is apparent much faster. Thanks to Lemma 1 of [10], to generate multiple random points in $\mathcal{V}$ one does not need to run the algorithm anew for each point, and consecutive points are sufficient. To simplify our analysis, and since H&R performs very well in practice, we neglect the error that is introduced by H&R.

*Hypercube Symmetries, Boolean Arithmetic, and Group Actions (Section VI):* An *automorphism* of a graph $G = (V, E)$ is an injective function $\sigma : V \to V$ which preserves edge-vertex connectivity, and the set of all automorphisms of a graph form a group $\text{Aut}(G)$ under composition. The Boolean field $\mathbb{F}_2$ is the set $\{\pm 1\}$ with the actions $\oplus$ and $\odot$, where $x \odot y = -1$ if and only if $x = y = -1$ and $x \oplus y = -1$ if and only if $x \neq y$. The set $\mathbb{F}_2^n$ is a vector space, and for vectors $\{\mathbf{v}_i\}$ in it we denote their linear span over $\mathbb{F}_2$ by $\text{span}_{\mathbb{F}_2}\{\mathbf{v}_i\}$.

We shall make use of the automorphism group $\text{Aut}(G)$ of the Boolean hypercube graph, whose vertices are $\mathbb{F}_2^n$, and two vertices are connected if their respective Hamming distance equals one (i.e., they are distinct in precisely one entry). It is widely known ([6, Prob. 3.11]) that $\text{Aut}(G) = S_n \times \mathbb{F}_2^n$, where $S_n$ is the permutation group on $[n]$. That is, every $\sigma \in \text{Aut}(G)$ corresponds to a permutation $\pi \in S_n$ and a vector $\mathbf{v} \in \mathbb{F}_2^n$ such that $\sigma(\mathbf{x}) = (x_{\pi(1)}, \ldots, x_{\pi(n)}) \oplus \mathbf{v} \triangleq \pi(\mathbf{x}) \oplus \mathbf{v}$, and hence we denote $\sigma = (\pi, \mathbf{v})$. It is an easy exercise to verify that if $\sigma = (\pi, \mathbf{v})$ then $\sigma^{-1} = (\pi^{-1}, \pi^{-1}(\mathbf{v}))$. Finally, for $\mathbf{w} \in \mathbb{R}^n$ and $\sigma = (\pi, \mathbf{v}) \in \text{Aut}(G)$ we let $\sigma(\mathbf{w}) \triangleq \pi(\mathbf{w}) \star \mathbf{v}$, where $\star$ is the point-wise product over $\mathbb{R}$, and notice that $\sigma$ is an invertible linear operator over $\mathbb{R}$, whose determinant is either $1$ or $-1$.

For a set $\mathcal{X} \subseteq \mathbb{F}_2^n$ let $\text{Stab}(\mathcal{X}) \subseteq \text{Aut}(G)$ be the set of all $\sigma \in \text{Aut}(G)$ such that $\sigma(\mathbf{x}) = \mathbf{x}$ for every $\mathbf{x} \in \mathcal{X}$, and notice that $\text{Stab}(\mathcal{X})$ is a subgroup of $\text{Aut}(G)$. Let $\mathbb{F}_2^n/\mathcal{X}$ be the set of all cosets of $\mathcal{X}$, i.e., all sets of the form $\mathcal{X} \oplus \mathbf{v} \triangleq \{\mathbf{x} \oplus \mathbf{v} | \mathbf{x} \in \mathcal{X}\}$ for some $v \in \mathbb{F}_2^n$. For $\mathcal{C}_1, \mathcal{C}_2 \in \mathbb{F}_2^n/\mathcal{X}$ we say that $\mathcal{C}_1 \sim \mathcal{C}_2$ if there exists $\sigma \in \text{Stab}(\mathcal{X})$ such that $\sigma(\mathcal{C}_1) = \mathcal{C}_2$. Since $\text{Stab}(\mathcal{X})$ is a group, we have that $\sim$ is an equivalence relation, and as such, partitions $\mathbb{F}_2^n/\mathcal{X}$ into $t$ disjoint equivalence classes $\mathcal{O}_1, \ldots, \mathcal{O}_t$ for some $t$, each of which is called an *orbit*.

## III. Basic Relations

We begin by making the following observation.

$$\begin{aligned}\mathop{\mathbb{E}}_{h_1,h_2} d(h_1,h_2) &= \mathop{\mathbb{E}}_{h_1,h_2}\left[\frac{1-\mathbb{E}_{\mathbf{x}}[h_1(\mathbf{x})h_2(\mathbf{x})]}{2}\right]\\ &= \mathop{\mathbb{E}}_{h_1,h_2}\left[\frac{1-\langle h_1,h_2\rangle}{2}\right]\\ &= \frac{1-\mathbb{E}_{h_1,h_2}[\langle h_1,h_2\rangle]}{2}.\end{aligned}$$

Therefore, computing $\mathbb{E}_{h_1,h_2} d(h_1,h_2)$ is equivalent to computing $\mathbb{E}_{h_1,h_2}[\langle h_1,h_2\rangle]$. We shall focus on the latter, for which we have

$$\begin{aligned}\mathop{\mathbb{E}}_{h_1,h_2}[\langle h_1,h_2\rangle] &= \mathop{\mathbb{E}}_{h_1,h_2}[\mathop{\mathbb{E}}_{\mathbf{x}}[h_1(\mathbf{x})h_2(\mathbf{x})]]\\ &\overset{(a)}{=} \mathop{\mathbb{E}}_{\mathbf{x}}[\mathop{\mathbb{E}}_{h_1,h_2}[h_1(\mathbf{x})h_2(\mathbf{x})]] \overset{(b)}{=} \mathop{\mathbb{E}}_{\mathbf{x}}\left(\mathop{\mathbb{E}}_{h} h(\mathbf{x})\right)^2\\ &\overset{(c)}{=} \mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2 \overset{(d)}{=} \sum_{S\subseteq[n]} \hat{H}(S)^2, \qquad (1)\end{aligned}$$

where $(a)$ holds since the probability spaces are finite, $(b)$ holds since $h_1$ and $h_2$ are chosen independently, in $(c)$ we denote $H(\mathbf{x}) \triangleq \mathbb{E}_h h(\mathbf{x})$, and $(d)$ follows from Parseval's identity. Notice that the function $H(\mathbf{x})$ satisfies

$$\begin{aligned}H(\mathbf{x}) &= \sum_{h\in\mathcal{H}} \Pr(h)\cdot h(\mathbf{x})\\ &= \sum_{h\in\mathcal{H}|h(\mathbf{x})=1} \Pr(h) - \sum_{h\in\mathcal{H}|h(\mathbf{x})=-1} \Pr(h)\\ &= \Pr_{\mathbf{w}\in\mathcal{V}}(\mathbf{w}\cdot\mathbf{x}\ge 0) - \Pr_{\mathbf{w}\in\mathcal{V}}(\mathbf{w}\cdot\mathbf{x}< 0)\\ &= 2\Pr_{\mathbf{w}\in\mathcal{V}}(\mathbf{w}\cdot\mathbf{x}\ge 0) - 1, \qquad (2)\end{aligned}$$

where $\mathbf{w}\in\mathcal{V}$ is chosen according to the distribution on $\mathcal{V}$ that is induced by $\mathcal{H}$.

## IV. Fourier Analytic Approximation of the Expected Diameter

In this section we use the fact that $\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle = \sum_{S\subseteq[n]}\hat{H}(S)^2$ (1). To this end, we first observe that for every $S\subseteq[n]$,

$$\begin{aligned}\hat{H}(S) &= \mathop{\mathbb{E}}_{\mathbf{x}}\chi_S(\mathbf{x})H(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{x}}\chi_S(\mathbf{x})\mathop{\mathbb{E}}_{h} h(\mathbf{x})\\ &= \mathop{\mathbb{E}}_{h}\mathop{\mathbb{E}}_{\mathbf{x}}\chi_S(\mathbf{x})h(\mathbf{x}) = \mathop{\mathbb{E}}_{h}\hat{h}(S). \qquad (3)\end{aligned}$$

Namely, the Fourier spectrum of $H$ is the expectation of the Fourier spectra of $h\in\mathcal{H}$. Moreover, every coefficient $\hat{H}(S)$ can be approximated by observing the dataset. That is, for $\ell(S) \triangleq \frac{1}{k}\sum_{i=1}^k \chi_S(\mathbf{x}_i)y_i = \frac{1}{k}\sum_{i=1}^k \chi_S(\mathbf{x}_i)H(\mathbf{x}_i)$, we have that

$$\Pr\left(|\ell(S)-\hat{H}(S)|\le\epsilon\right) \ge 1-2e^{-\frac{k\epsilon^2}{2}} \qquad (4)$$

for every $\epsilon > 0$ by Hoeffding's inequality, where the probability is over the choice of the dataset. Hence, it follows that $\sum_{S\subseteq[n]}\ell(S)^2 \approx \sum_{S\subseteq[n]}\hat{H}(S)^2$. However, we wish to avoid computing $\ell(S)$ over all $2^n$ subsets $S$ of $[n]$. To this end, we prove a Fourier concentration bound for $H$, that follows from Lemma 1 by the Cauchy-Schwartz inequality, and depends on the parameter $c(\mathcal{H}) = |\mathcal{H}|\sum_{h\in\mathcal{H}}\Pr(h)^2$.

**Lemma 2.** *For $a\in\mathbb{N}$ and $b\in\mathbb{R}$, if $W^{\ge a}[h]\le b$ for every $h\in\mathcal{H}$, then $W^{\ge a}[H]\le b\cdot c(\mathcal{H})$, and therefore, $\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle - b\cdot c(\mathcal{H}) \le \sum_{|S|<a}\hat{H}(S)^2 \le \mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle$.*

*Proof.* We have:

$$\begin{aligned}W^{\ge a}[H] &= \sum_{|S|\ge a}\hat{H}(S)^2 \overset{(3)}{=} \sum_{|S|\ge a}\left(\mathop{\mathbb{E}}_{h}\hat{h}(S)\right)^2\\ &= \sum_{|S|\ge a}\left(\sum_{h\in\mathcal{H}}\Pr(h)\cdot\hat{h}(S)\right)^2\end{aligned}$$

$$
\begin{aligned}
&\overset{(\dagger)}{\leq} \sum_{|S|\geq a}\left(\sum_{h\in\mathcal{H}}\Pr(h)^2\right)\left(\sum_{h\in\mathcal{H}}\hat{h}(S)^2\right)\\
&= \sum_{h_1\in\mathcal{H}}\sum_{h_2\in\mathcal{H}}\Pr(h_1)^2\sum_{|S|\geq a}\hat{h}_2(S)^2\\
&\overset{(\ddagger)}{\leq} b\sum_{h_1\in\mathcal{H}}\sum_{h_2\in\mathcal{H}}\Pr(h_1)^2\\
&= b|\mathcal{H}|\cdot\sum_{h}\Pr(h)^2 = b\cdot c(\mathcal{H}),
\end{aligned}
$$

where (†) follows from the Cauchy-Schwartz inequality, and (‡) from $W^{\geq a}[h]\leq b$. The second part of the lemma follows directly from (1). □

Now, we approximate $\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle$ by $\sum_{|S|<a}\ell(S)^2$. Since $\chi_S(\mathbf{x})=\prod_{j\in S}x_j$ for every $S\subseteq[n]$, this approximation can be computed in $\binom{n}{a}\cdot O(ka)$ time. To estimate the error, let $\epsilon_S\triangleq\ell(S)-\hat{H}(S)$, and notice that $\hat{H}(S)^2=\ell(S)^2-2\epsilon_S\cdot\ell(S)+\epsilon_S^2$ for every $S\subseteq[n]$. Therefore, since $|\ell(S)|\leq 1$, it follows that

$$\hat{H}(S)^2-2|\epsilon_S|-\epsilon_S^2\leq\ell(S)^2\leq\hat{H}(S)^2+2|\epsilon_S|-\epsilon_S^2$$

for every $S\subseteq[n]$. Thus, by applying Lemma 1 and Lemma 2 it follows that

$$
\begin{aligned}
\mathop{\mathbb{E}}_{h_1,h_2}\langle h_1,h_2\rangle-b\cdot c(\mathcal{H})-2\binom{n}{<a}\epsilon_{a,\max}-\binom{n}{<a}\epsilon_{a,\max}^2 &\leq \sum_{|S|<a}\ell(S)^2\\
&\leq \mathop{\mathbb{E}}_{h_1,h_2}\langle h_1,h_2\rangle+2\binom{n}{<a}\epsilon_{a,\max},
\end{aligned}
\tag{5}
$$

where $\binom{n}{<a}\triangleq\sum_{j=0}^{a-1}\binom{n}{j}$, $\epsilon_{a,\max}\triangleq\max\{|\epsilon_S||S\in\binom{[n]}{<a}\}$, and $b=O(1/\sqrt{a})$.

Clearly, to have any meaningful asymptotic conclusions from (5), we must have that $\epsilon_{a,\max}=o(\frac{1}{n^a})$, where $a$ is seen as constant. To this end, a sufficient condition is that $|\epsilon_S|\leq\epsilon$ for every $S\in\binom{[n]}{<a}$ and some $\epsilon=o(\frac{1}{n^a})$. Since for each individual $S\in\binom{[n]}{<a}$ the event $|\epsilon_S|\leq\epsilon$ occurs with probability at least $1-2e^{-\frac{k\epsilon^2}{2}}$ (4), it follows that a fraction of at least

$$1-2\binom{n}{<a}e^{-\frac{k\epsilon^2}{2}}\triangleq 1-\eta$$

of all datasets of size $k$ satisfy the condition $|\epsilon_S|\leq\epsilon$ for every $S\in\binom{[n]}{<a}$. Hence, we must have that

$$\epsilon=\sqrt{\frac{2}{k}\cdot\ln\left(\frac{2}{\eta}\cdot\binom{n}{<a}\right)}=o\left(\frac{1}{n^a}\right). \tag{6}$$

Therefore, it follows that one can for example fix $\eta=e^{-n}$, and then $k=\Omega(n^{2a+2})$ suffices to satisfy (6).

To summarize, the scheme in this section applies for any probability distribution on $\mathcal{H}$, but provides better approximation results for distributions with smaller value for $c(\mathcal{H})$; that includes $\mathcal{H}_{uni}$, for which $c(\mathcal{H}_{uni})=1$. An accuracy-complexity tradeoff is readily seen—a larger value for $a$ increases the complexity of the algorithm and the required number of points in the dataset, but results in a better approximation. Notice also that the algorithm in this section is deterministic; the probability analysis guarantees a successful approximation for all but an exponentially small fraction of the datasets. We summarize this section in the following theorem, and continue to study the special case of the volume distribution in the next section.

**Theorem 1.** *Whenever $k=\Omega(n^{2a+2})$ we have that*

$$\mathop{\mathbb{E}}_{h_1,h_2}\langle h_1,h_2\rangle-\frac{c(\mathcal{H})}{\Omega(\sqrt{a})}+o(1)\leq\sum_{|S|<a}\ell(S)^2\leq\mathop{\mathbb{E}}_{h_1,h_2}\langle h_1,h_2\rangle+o(1),$$

*for all but exponentially small fraction of possible datasets. Namely, for probability distributions on $\mathcal{H}$ whose respective $c(\mathcal{H})$ is constant, one can approximate $\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle$ with high probability up to to arbitrary (constant) precision in polynomial time, while operating on polynomially many points.*

## V. Approximations for the Volume Distribution

The algorithms below require random sampling from $\mathcal{V}$, for which the H&R algorithm is used. We emphasize that every use of the H&R algorithm requires a "warm-up", after which the points are sufficiently random. Moreover, choosing a point uniformly at random from the chord at each step can be done in $O(nk)$ time (Lemma 6 in Appendix A). For the sake of brevity, we omit the warm-up phase from the complexity analysis.

*The Direct Algorithm (**DIR**):* Let $m = m(\epsilon, \eta)$, $\ell = \ell(\epsilon, \eta)$ be integers that will be computed in the sequel. This algorithm chooses $m$ pairs $(\mathbf{w}_{i_t}, \mathbf{w}_{j_t})_{t=1}^{m}$ and $\ell$ binary vectors $\mathbf{z}_{t,j}$ for every $t \in [m]$, and returns

$$\mathbf{est}_D \triangleq \frac{1}{m\ell} \sum_{t=1}^{m} \sum_{j=1}^{\ell} \operatorname{sign}(\mathbf{w}_{i_t} \mathbf{z}_{t,j}) \operatorname{sign}(\mathbf{w}_{j_t} \mathbf{z}_{t,j}).$$

It is readily verified that the complexity of this approximation is $O(mn(k + \ell))$. By repeated applications of Hoeffding's inequality, that are detailed in Appendix B, it follows that

$$\Pr\left(\left|\mathbf{est}_D - \mathop{\mathbb{E}}_{h_1,h_2} \langle h_1, h_2 \rangle\right| \leq \epsilon\right) \geq 2\left(1 - e^{-\frac{m\delta^2}{2}}\right) \cdot \left(1 - e^{-\frac{\ell(\epsilon-\delta)^2}{2}}\right)^{m} - 1,$$

where $\epsilon = \delta + \mu$. Hence, for example, one can choose $\delta = \frac{\epsilon}{2}$ and $m = \frac{c}{\epsilon^2}$ for some constant $c$, and then

$$\ell = -\frac{8}{\epsilon^2} \ln\left(1 - \left(1 - \frac{1 - \frac{\eta}{2}}{1 - e^{-c/8}}\right)^{\epsilon^2/c}\right),$$

and the overall complexity is

$$O\left(\frac{nk}{\epsilon^2} + \frac{n}{\epsilon^4} \ln\left(1 - \left(1 - \frac{1 - \frac{\eta}{2}}{1 - e^{-c/8}}\right)^{\epsilon^2/c}\right)^{-1}\right).$$

*The Alternative Algorithm (**ALT**):* Let $s = s(\epsilon, \eta)$ and $r = r(\epsilon, \eta)$ be integers that will be computed in the sequel. This algorithm estimates $\mathbb{E}_{h_1,h_2}\langle h_1, h_2 \rangle$ by using its equality to $\mathbb{E}_{\mathbf{x}} H(\mathbf{x})^2$, which in turn equals $\mathbb{E}_{\mathbf{x}}(2 \Pr_{\mathbf{w} \in \mathcal{V}}(\mathbf{w} \cdot \mathbf{x} \geq 0) - 1)^2$ (see Section II). Naïvely, one can estimate this quantity as $\frac{1}{r}\sum_{i=1}^{r}\left(2 \cdot \frac{1}{s}\sum_{j=1}^{s} \mathbb{1}(\mathbf{w}_j \mathbf{z}_i \geq 0) - 1\right)^2$ where $\mathbb{1}$ is a Boolean indicator, and where the $\mathbf{z}_i$'s and $\mathbf{w}_j$'s are chosen uniformly at random from $\{\pm 1\}^n$ and from $\mathcal{V}$, respectively. However, in Appendix C it is shown that the following approximation is usually better.

$$\mathbf{est}_A \triangleq \frac{1}{r} \sum_{i=1}^{r} \left(-4 \left(\frac{1}{s/2} \sum_{j=1}^{s/2} (1 - \mathbb{1}_{i,j,i}) \mathbb{1}_{i,j+\frac{s}{2},i}\right) + 1\right), \tag{7}$$

where $\mathbb{1}_{a,b,c}$ stands for $\mathbb{1}(\mathbf{w}_{a,b}\mathbf{z}_c \geq 0)$. The complexity of this algorithm is $O(sn(k+r))$. According to a probabilistic analysis that is given in Appendix D, we have that

$$\Pr\left(\left|\mathbf{est}_A - \mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2\right| \leq \delta + 4\mu\right) \geq 2\left(1 - e^{-\frac{r\delta^2}{2}}\right) \cdot (1 - r \cdot e^{-s\mu^2}) - 1,$$

where $\epsilon = \delta + 4\mu$. Once again, we choose, say, $\delta = \frac{\epsilon}{2}$ and $r = \frac{c'}{\epsilon^2}$ for some constant $c'$, and get

$$s = -\frac{64}{\epsilon^2} \log\left(\frac{\epsilon^2}{c'}\left(1 - \frac{1 - \frac{\eta}{2}}{1 - e^{-c'/8}}\right)\right),$$

and the overall complexity is

$$O\left(\frac{n}{\epsilon^2} \log\left(\frac{\epsilon^2}{c'}\left(1 - \frac{1 - \frac{\eta}{2}}{1 - e^{-c'/8}}\right)\right)^{-1}\left(k + \frac{c'}{\epsilon^2}\right)\right).$$

Practically, in Section VIII we run **DIR** and **ALT** on randomly generated datasets until convergence is apparent. While the resulting approximations are comparable, **ALT** demonstrates faster convergence times as the number of sampled **z**'s ($\ell$ in **DIR** and $r$ in **ALT**) grows. This phenomenon is yet to be explained.

## VI. Expected Diameter of Structured Data

In this section an additional appealing property of the expected diameter is revealed. It is shown that algebraic features of the set $\mathcal{X}$ can be exploited to perform significantly less computations. This result will be particularly useful whenever $\mathcal{X}$ is a *subcube* of $\{\pm 1\}^n$, and applies for both $\mathcal{H}_{uni}$ and $\mathcal{H}_{vol}$.

The main result of this section is that the expected distance is uniform on *cosets* of $\mathcal{X}$. In what follows, for any subset $\mathcal{C} \subseteq \{\pm 1\}^n$ we define the $\mathcal{C}$*-restricted distance* (restricted distance, in short)

$$d_{\mathcal{C}}(h_1, h_2) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c} \in \mathcal{C}} \frac{1 - h_1(\mathbf{c}) h_2(\mathbf{c})}{2}. \tag{8}$$

**Lemma 3.** *(The Coset Lemma) Let $\sigma \in \mathrm{Aut}(G)$ such that $\sigma(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, for cosets $\mathcal{C}_1$ and $\mathcal{C}_2$ of $\mathcal{X}$ such that $\sigma(\mathcal{C}_1) = \mathcal{C}_2$, we have that*

$$\mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_1}(h_1, h_2) = \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_2}(h_1, h_2).$$

A proof is given in Appendix E. The uniformity of the expected distance on cosets in the same orbit allows us to develop the following formula.

**Corollary 1.** *Assume that $\mathbb{F}_2^n / \mathcal{X}$ is partitioned to the orbits $\mathcal{O}_1, \ldots, \mathcal{O}_t$, and pick $\mathcal{C}_i \in \mathcal{O}_i$ arbitrarily for every $i \in [t]$. Then, we have*

$$\begin{aligned} \mathop{\mathbb{E}}_{h_1,h_2} d(h_1, h_2) &= \mathop{\mathbb{E}}_{h_1,h_2} \frac{|\mathcal{X}|}{2^n} \sum_{i=1}^{t} |\mathcal{O}_i| d_{\mathcal{C}_i}(h_1, h_2) \\ &= \frac{|\mathcal{X}|}{2^n} \sum_{i=1}^{t} |\mathcal{O}_i| \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_i}(h_1, h_2). \end{aligned}$$

*Namely, in order to compute $\mathbb{E}_{h_1,h_2} d(h_1, h_2)$, it suffices to compute the expected distance when restricted to* orbit representatives *from the orbits of $\mathcal{X}$.*

Of course, utilizing Corollary 1 for efficient computation of $\mathbb{E}_{h_1,h_2} d(h_1, h_2)$ strongly depends on the structure of $\mathcal{X}$, and the size of the respective orbits. In what follows we provide an example for a structure for which Corollary 1 is particularly powerful.

For $\mathbf{v} \in \mathbb{F}_2^n$ and $I \subseteq [n]$, the set $\mathcal{X}$ is called a $(\mathbf{v}, I)$-subcube (subcube, in short), if $\mathcal{X} = \mathrm{span}_{\mathbb{F}_2}\{\mathbf{e}_i\}_{i \in I} \oplus \mathbf{v}$, where $\mathbf{e}_i$ is the $i$'th unit vector (i.e., $e_{i,j} = -1$ if $i = j$, and $1$ otherwise). It is readily verified that $\mathcal{X}$ is an affine subspace of $\mathbb{F}_2^n$ of dimension $|I|$. The following results are proved in Appendix F.

**Lemma 4.** *If $\mathcal{X}$ is a $(\boldsymbol{v}, I)$-subcube for some $I = \{i_j\}_{j=1}^{\ell}$ and $\boldsymbol{v} \in \mathbb{F}_2^n$, then $\mathcal{X}$ has $n - |I|$ orbits, and a set of representatives is given by $\mathcal{C}_i = \mathcal{X} \oplus \boldsymbol{u}_i$, where $\boldsymbol{u}_i$ is any vector whose Hamming weight*[3] *on $[n] \setminus I$ is $i$.*

**Corollary 2.** *If $\mathcal{X}$ is a $(\boldsymbol{v}, I)$-subcube for some $I \subseteq [n]$ and $\boldsymbol{v} \in \mathbb{F}_2^n$, then*

$$\mathop{\mathbb{E}}_{h_1,h_2} d(h_1, h_2) = 2^{|I| - n} \sum_{i=1}^{n - |I|} \binom{n - |I|}{i} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_i}(h_1, h_2).$$

A particularly attractive property of Corollary 2 is that the significant contribution to $\mathbb{E}_{h_1,h_2} d(h_1, h_2)$ comes from $O(\sqrt{n - |I|})$ of indices $i \in [n - |I|]$ (See Appendix G). Hence, for example, the contribution of every randomly chosen pair $h_1, h_2$ to the expected diameter can be computed *exactly* in $O(nk(n - |I|))$ time, or approximated closely in $O(nk\sqrt{n - |I|})$ time.

## VII. The Case of Data Over $\mathbb{R}$

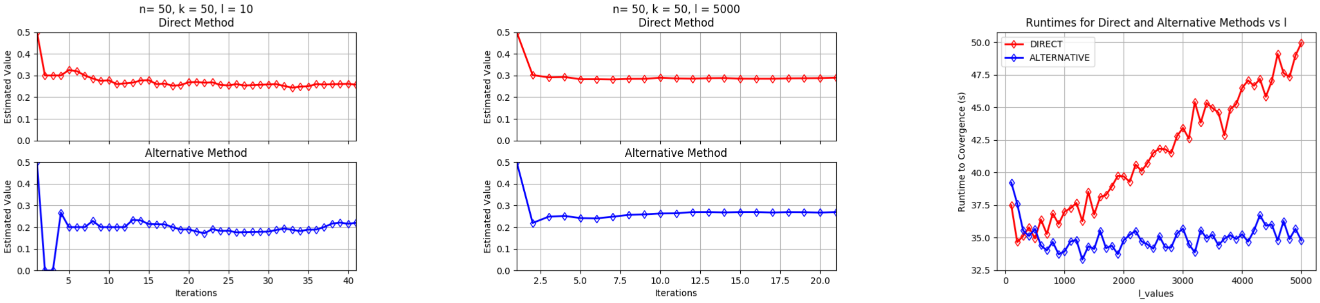
Consider the case where $\mathbf{x}_i \in \mathbb{R}^n$ rather than $\mathbf{x}_i \in \{\pm 1\}^n$. While the definitions of $\mathcal{H}$ and $\mathcal{V}$ extend verbatim to this case, one must revise the definition of distance. Aiming to reflect the fraction of disagreement, we define

$$d(h_1, h_2) = \frac{1}{\mathrm{Vol}(\mathbb{B}_n)} \int_{\mathbb{B}_n} \frac{1 - h_1(x) h_2(x)}{2} dx,$$

where $\mathbb{B}_n$ is the $n$-dimensional unit ball. However, one can easily notice that this definition is equivalent to the definition of *angle*. Therefore, one can settle for

$$d(h_1, h_2) = \frac{1}{\pi} \cdot \arccos\left( \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{\|\mathbf{w}_1\|_2 \cdot \|\mathbf{w}_2\|_2} \right),$$

[3] The Hamming weight of $u$ on $[n] \setminus I$ is the size of the set $\{j \in [n] \setminus I \mid u_j = -1\}$.



(a) (b) (c)

Fig. 4: Runtime comparison and convergence plots for **DIR** and **ALT** (Section V).

where $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n$ are vectors that define $h_1, h_2$. Hence, assuming the distribution $\mathcal{H}_{vol}$ on $\mathcal{H}$ ($\mathcal{H}_{uni}$ is not well-defined in this case), one can estimate the average distance by the simple algorithm the averages the above expression over $t$ random pairs from $\mathcal{V}$, i.e.,

$$\frac{1}{t}\sum_{\ell=1}^{t} \arccos\left(\frac{\mathbf{w}_{i_\ell}\cdot\mathbf{w}_{j_\ell}}{\|\mathbf{w}_{i_\ell}\|_2\cdot\|\mathbf{w}_{j_\ell}\|_2}\right),$$

where $\{\mathbf{w}_{i_\ell}\}_{\ell=1}^{t}$ and $\{\mathbf{w}_{j_\ell}\}_{\ell=1}^{t}$ are chosen uniformly at random by H&R. Notice that this algorithm can be used in the case of Boolean halfspaces as well (i.e., where $\mathbf{x}_i \in \{\pm 1\}^n$), but the above distance measure does not reflect the *Boolean* disagreement between the halfspaces, since it is not clear how many hypercube points lie in the intersection of two halfspaces.

## VIII. Experimental Results

We ran our experiments on an Intel Core $i5$-$4570$, $3.20$GHz $4 \times 4$ with $3.8$GiB RAM memory and ubuntu: $16.04$ LTS operating system. We used $10^5$ iterations of H&R as a warm-up. Afterwards, $500$ intermediate steps were made to generate consecutive samples. Both **DIR** and **ALT** were run until no more than $5 \cdot 10^{-2}$ additive difference in the estimation was observed during $10$ iterations. Our experiments demonstrate the feasibility of some of our techniques, but are inconclusive as of which one among **DIR** and **ALT** is preferable.

*Expected Diameter vs. Accuracy:* In the experiment of Figure 2, 300 datasets of size $k = 20$ and dimension $n = 50$ were generated at random and labeled by a halfspace $h$ with a standard Gaussian weight vector $\mathbf{w}$. All points in the *arbitrary* (**A**) datasets were generated at random from $\mathrm{Bern}(0.5)$. In the *bad* (**B**) datasets, $k/2$ points $\mathbf{x}_i$ were chosen by $\mathrm{Bern}(0.5)$, and then their negation $-\mathbf{x}_i$ was added to the dataset (notice that $\mathrm{sign}(\mathbf{w}\cdot\mathbf{x}) = -\,\mathrm{sign}(\mathbf{w}\cdot(-\mathbf{x}))$ for every $\mathbf{x}$, and hence having both $\mathbf{x}$ and $-\mathbf{x}$ does not contribute to the learner more than just having either). In the *good* (**G**) datasets, we applied a simple iterative algorithm to find $k/2$ "boundary" pairs $\mathbf{x}, \mathbf{y} \in \{\pm 1\}^n$, i.e., such that $h(\mathbf{x}) \neq h(\mathbf{y})$, and the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ is 1. After generating these datasets, the algorithm **DIR** was applied until convergence.

In Figure 3, for each one of the **A**, **B**, and **G** datasets, we conducted the following experiment—First, the perceptron algorithm was applied, where the starting point and the order of the points is randomized. Then, a random consistent hypothesis is chosen with H&R (the "true" function), and the distance between these two functions is estimated. It is evident that on average, the performance of perceptron is superior in datasets with lower expected diameter.

*Performance Comparison:* Let $l$ be the number of samples from $\{\pm 1\}^n$ in each iteration of either **DIR** or **ALT**. We observed greater stability when increasing $l$ in both algorithms (e.g., Figure 4a vs. Figure 4b), but in **DIR** one has to pay a much greater penalty in terms of running time for increasing $l$. This is apparent in Figure 4c, where the run-times are averaged over 20 independent arbitrary datasets (see above).

## References

[1] J. Abernethy, Y. Chen, C.-J. Ho, and B. Waggoner, "Low-cost learning via active data procurement," *ACM Conference on Economics and Computation* (EC), pp. 619–636, 2015.

[2] Y. Chen, N. Immorlica, B. Lucier, V. Syrgkanis, and J. Ziani, "Optimal data acquisition for statistical estimation," *ACM Conference on Economics and Computation* (EC), pp. 27–44, 2018.

[3] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, vol. 3, pp. 326–334, 1965.

[4] L. Lovász, "Hit-and-run mixes fast," *Mathematical Programming*, vol. 86, no. 3, pp. 443–461, 1999.

[5] R. O'Donnell. Analysis of Boolean functions. Cambridge University Press, 2014.

[6] F. T. Leighton, Introduction to parallel algorithms and architectures: Arrays, trees, hypercubes. Elsevier, 2014.

[7] M. Anthony. Discrete mathematics of neural networks: selected topics. Vol. 8. Siam, 2001.

[8] Y. Peres, "Noise stability of weighted majority," *arXiv preprint math/0412377*, 2004.

[9] S. Shalev-Shwartz and S. Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

[10] R. L. Smith, "Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions," *Operations Research*, vol. 32, no. 6, 1296–1308, 1984.

[11] R. L. Smith, "The hit-and-run sampler: a globally reaching Markov chain sampler for generating arbitrary multivariate distributions," *The IEEE Computer Society 28th conference on Winter simulation*, pp. 260–264, 1996.

[12] C. Tosh and S. Dasgupta, "Diameter-Based Active Learning," *International Conference on Machine Learning* (ICML), pp. 3444–3452, 2017.

[13] V. Vapnik, Vladimir and R. Izmailov, "Rethinking statistical learning theory: learning using statistical invariants," *Machine Learning*, vol. 108, no. 3, pp. 381–423, 2019.

[14] E. W. Weisstein, "Central Limit Theorem," *MathWorld–A Wolfram Web Resource*, http://mathworld.wolfram.com/CentralLimitTheorem.html.

## Appendix A
## Omitted proofs

**Lemma 5.** *(Range of the expected diameter) For every dataset $\mathcal{D}$ and every probability distribution $\mathcal{H}$,*

$$\mathop{\mathbb{E}}_{h_1,h_2\in\mathcal{H}} d(h_1,h_2) \le 0.5.$$

*Proof.* According to Section III we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{h_1,h_2\in\mathcal{H}} d(h_1,h_2) &= \frac{1-\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle}{2}\\ &= \frac{1-\mathbb{E}_{\mathbf{x}}H(\mathbf{x})^2}{2},\end{aligned}$$

and since $\mathbb{E}_{\mathbf{x}}H(\mathbf{x})^2$ is nonnegative, the claim follows. □

**Lemma 6.** *(The complexity of the chord function) Given $\mathbf{v}\in\mathbb{R}^n$ and $\mathbf{l}\in\mathbb{R}^n$, one can choose a random elements from $\{\mathbf{v}+t\mathbf{l}|t\in\mathbb{R}\}\cap\mathcal{V}$ in $O(nk)$ time.*

*Proof.* Given $\mathbf{v}$ and $\mathbf{l}$ in $\mathbb{R}^n$, we ought to find the values $t_1$ and $t_2$ that define the body

$$\begin{aligned}&\mathcal{V}\cap\{\mathbf{v}+t\cdot\mathbf{l}\mid t\in\mathbb{R}\}\\ &=\{\mathbf{v}+t\cdot\mathbf{l}\mid t\in\mathbb{R},\ y_i((\mathbf{v}+t\cdot\mathbf{l}))\cdot\mathbf{x}_i\ge 0\\ &\qquad\qquad \text{for all } i\in[k],\ \text{and } \|\mathbf{v}+t\cdot\mathbf{l}\|_2\le 1\}\\ &=\{\mathbf{v}+t\cdot\mathbf{l}\mid t_1\le t\le t_2\}.\end{aligned}\tag{9}$$

In $O(nk)$ time we can turn each of the $k$ linear constrains in (9) into either an upper or a lower bound on $t$ (depending on whether $y_i=1$ or $y_i=-1$), and intersect them to obtain a bound of the form $m_1\le t\le m_2$. Further, the $\ell_2$-norm constraint in (9) can be turned to a quadratic inequality of the form $a_2t^2+a_1t+a_0\ge 0$ in $O(n)$ time, and then turned to to a bound of the form $c_1\le t\le c_2$ in $O(1)$ time by solving it (notice that it will not be of the form "$t\le c_1$ or $c_2\le t$" since $\mathcal{V}$ is convex). Then, we intersect the segments $[m_1,m_2]$ and $[c_1,c_2]$ to find $t_1$ and $t_2$. □

## Appendix B
## Probabilistic analysis of DIR

We analyze the relation between $m$ and $\ell$ and the guaranteed approximation. First, for every $t\in[m]$, by the Hoeffding inequality we have that

$$\Pr\left(\mathop{\mathbb{E}}_{\mathbf{x}}(h_{i_t}(\mathbf{x})h_{j_t}(\mathbf{x}))\le\frac{1}{\ell}\sum_{j=1}^{\ell}h_{i_t}(\mathbf{z}_{t,j})h_{j_t}(\mathbf{z}_{t,j})+\mu\right)\ge 1-e^{-\frac{\ell\mu^2}{2}}\tag{10}$$

for every $\mu>0$, where the probability is over the random choice of $\mathbf{z}_{t,1},\ldots,\mathbf{z}_{t,\ell}$, and where $h_{i_t}(\mathbf{x})\triangleq\operatorname{sign}(\mathbf{w}_{i_t}\cdot\mathbf{x})$ (resp. $h_{j_t}$). Also by the Hoeffding inequality, we have that

$$\Pr\left(\mathbb{E}_{h_1,h_2}\langle h_1,h_2\rangle\le\frac{1}{m}\sum_{t=1}^{m}\mathop{\mathbb{E}}_{\mathbf{x}}(h_{i_t}(\mathbf{x})h_{j_t}(\mathbf{x}))+\delta\right)\ge 1-e^{-\frac{m\delta^2}{2}}\tag{11}$$

for every $\delta>0$. It is straightforward to show that if (10) holds for every $t\in[m]$ and (11) holds, then

$$\mathop{\mathbb{E}}_{h_1,h_2}\langle h_1,h_2\rangle-\frac{1}{m\ell}\sum_{t=1}^{m}\sum_{j=1}^{\ell}h_{i_t}(\mathbf{z}_j)h_{j_t}(\mathbf{z}_j)\le\delta+\mu.$$

Therefore, since (10) is true for any pair in $\mathcal{H} \times \mathcal{H}$, by applying symmetric arguments to (10) and (11), we have that

$$\Pr\left(\left|\frac{1}{m\ell}\sum_{j=1}^{\ell}\sum_{t=1}^{m} h_{i_t}(\mathbf{z}_j)h_{j_t}(\mathbf{z}_j) - \mathop{\mathbb{E}}_{h_1,h_2}\langle h_1, h_2\rangle\right| \le \epsilon\right) = \Pr\left(\left|\mathbf{est} - \mathop{\mathbb{E}}_{h_1,h_2}\langle h_1, h_2\rangle\right| \le \epsilon\right)$$
$$\ge 2\left(1 - e^{-\frac{m\delta^2}{2}}\right)\left(1 - e^{-\frac{\ell(\epsilon-\delta)^2}{2}}\right)^m - 1,$$

where $\epsilon = \delta + \mu$.

## Appendix C
## Learning a function of a Bernoulli variable

In what follows, the samples $x_1, \ldots, x_n$ correspond to the Bernoulli variables $\mathbb{1}(\mathbf{w} \cdot \mathbf{x} \ge 0)$ that are mentioned in the description of **ALT**, and the parameter $p$ equals $\mathbb{E}_{\mathbf{w}\in\mathcal{V}}\,\mathbb{1}(\mathbf{w} \cdot \mathbf{x} \ge 0) = \Pr_{\mathbf{w}\in\mathcal{V}}(\mathbf{w} \cdot \mathbf{x} \ge 0)$, where $\mathbf{x}$ in any element of $\{\pm 1\}^n$.

*Problem:* Given $n$ i.i.d samples $x_1, \ldots, x_n$ from $Bern(p)$, find the best possible approximation to $(2p-1)^2$. That is, for a given probability $\mu$, find as small as possible $\eta$ and a function $f$ for which

$$\Pr\left(\left|f(x_1, \ldots, x_n) - (2p-1)^2\right| \le \eta\right) \ge \mu.$$

*Solution 1:* $f(x_1, \ldots, x_n) = (\frac{2}{n}\sum_{i=1}^n x_i - 1)^2$. Let $\delta$ be such that $\mu = 1 - 2e^{-2n\delta^2}$. By Hoeffding's inequality we have that

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n x_i - p\right| \le \delta\right) \ge \mu.$$

Therefore, with probability $\mu$ we have that

$$\left(\frac{2}{n}\sum_{i=1}^n x_i - 1\right)^2 = 4\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 - 4\left(\frac{1}{n}\sum_{i=1}^n x_i\right) + 1$$
$$\le 4(p+\delta)^2 - 4(p-\delta) + 1$$
$$= (2p-1)^2 + (8p\delta + 4\delta) + 4\delta^2.$$

Similarly, we have a lower bound of $(2p-1)^2 - (8p\delta + 4\delta) + 4\delta^2$, and thus, neglecting $4\delta^2$, we have $\eta = 8p\delta + 4\delta$.

*Solution 2:* $f(x_1, \ldots, x_n) = -4\left(\frac{2}{n}\sum_{j=1}^{n/2}(1 - x_{1,j})x_{2,j}\right) + 1$, where $x_1, \ldots, x_n$ are indexed as $x_{1,1}, \ldots, x_{1,n/2}, x_{2,1}, \ldots, x_{2,n/2}$. It is readily seen that if $x$ and $y$ are chosen i.i.d from $Bern(p)$ then $\mathbb{E}[y(1-x)] = p(1-p)$. Hence, by fixing $\delta'$ such that $\mu = 1 - 2e^{-n\delta'^2}$, by Hoeffding's inequality we have that

$$\Pr\left(\left|\frac{1}{n/2}\sum_{j=1}^{n/2}(1 - x_{1,j})x_{2,j} - p(1-p)\right| \le \delta'\right) \ge \mu.$$

Therefore, with probability $\mu$ we have that

$$f(x_1, \ldots, x_n) = -4\left(\frac{1}{n/2}\sum_{j=1}^{n/2}(1 - x_{1,j})x_{2,j}\right) + 1$$
$$\le -4(p(1-p) - \delta') + 1$$
$$= (2p-1)^2 + 4\delta'.$$

Similarly, we can guarantee a lower bound of $(2p-1)^2 - 4\delta'$, and thus $\eta = 4\delta'$.

We are left to compare the confidence intervals. Since $\mu = 1 - 2e^{-2n\delta^2} = 1 - 2e^{-n\delta'^2}$, it follows that $\delta' = \sqrt{2} \cdot \delta$. Therefore, in Solution 2 we have $\eta = 4\delta' = 4\sqrt{2} \cdot \delta \approx 5.65\delta$. It readily follows that $8p\delta + 4\delta < 5.65\delta$ for $p < \frac{1.65}{8}$. Hence, Solution 2 is a better estimation whenever $p > \frac{1.65}{8}$. Since Solution 2 covers a broader range of $p$ values we prefer it over Solution 1 in **ALT**.

## Appendix D
## Probabilistic analysis of ALT

In this analysis, we employ the abbreviated notations $P_{\mathbf{x}} \triangleq \Pr_{\mathbf{w}\in\mathcal{V}}(\mathbf{w}\cdot\mathbf{x} \geq 0)$ and $\mathbb{1}_{j,i} \triangleq \mathbb{1}(\mathbf{w}_j\mathbf{z}_i \geq 0)$. First observe that for every $\mathbf{x}\in\{\pm1\}^n$ we have that $\mathbb{E}_{\mathbf{w}\in\mathcal{V}}\,\mathbb{1}(\mathbf{w}\cdot\mathbf{x}\geq 0) = P_{\mathbf{x}}$. Hence, it readily follows that

$$\mathop{\mathbb{E}}_{\mathbf{w}_1,\mathbf{w}_2\in\mathcal{V}}[(1-\mathbb{1}(\mathbf{w}_1\cdot\mathbf{x}\geq 0))\mathbb{1}(\mathbf{w}_2\cdot\mathbf{x}\geq 0)] = (1-P_{\mathbf{x}})P_{\mathbf{x}} \text{ for every } \mathbf{x}\in\{\pm1\}^n,$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are chosen independently and uniformly from $\mathcal{V}$. Therefore, by the Hoeffding inequality, for every $\mathbf{x}\in\{\pm1\}^n$ we have that

$$\Pr\left((1-P_{\mathbf{x}})P_{\mathbf{x}} \leq \frac{1}{s/2}\sum_{j=1}^{s/2}(1-\mathbb{1}(\mathbf{w}_j\mathbf{x}\geq 0))\mathbb{1}(\mathbf{w}_{j+s/2}\mathbf{x}\geq 0)+\mu\right) \geq 1-e^{-s\mu^2} \tag{12}$$

for every $\mu > 0$. That is, at most an $e^{-s\mu^2}$ fraction of the $s$-tuples in $\mathcal{V}^s$ are "bad for $\mathbf{x}$", i.e., tuples for which the event in (12) *does not* occur. Therefore, since this claim is true for any $\mathbf{x}\in\{\pm1\}^n$, it follows that given any $\mathbf{z}_1,\ldots,\mathbf{z}_r$ in $\{\pm1\}^n$, at most an $r\cdot e^{-s\mu^2}$ fraction of $\mathcal{V}^s$ are bad for at least one $\mathbf{z}_j$, and the rest of $\mathcal{V}^s$ are "good" for all $\mathbf{z}_j$'s. Also by the Hoeffding inequality, we have

$$\Pr\left(\mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2 \geq \frac{1}{r}\sum_{i=1}^{r}(2P_{\mathbf{z}_i}-1)^2-\delta\right) \geq 1-e^{-\frac{r\delta^2}{2}}.$$

for every $\delta > 0$. Now, notice that if:

1) $(1-P_{\mathbf{z}_i})P_{\mathbf{z}_i} \leq \frac{1}{s/2}\sum_{j=1}^{s/2}(1-\mathbb{1}_{j,i})\mathbb{1}_{j+s/2,i}+\mu$ for some $\mu>0$ and every $i\in[r]$; and
2) $\mathbb{E}_{\mathbf{x}} H(\mathbf{x})^2 \geq \frac{1}{r}\sum_{i=1}^{r}(2P_{\mathbf{z}_i}-1)^2-\delta$ for some $\delta>0$, then (7) satisfies:

$$\begin{aligned}\frac{1}{r}\sum_{i=1}^{r}\left(-4\left(\frac{1}{s/2}\sum_{j=1}^{s/2}(1-\mathbb{1}_{j,i})\mathbb{1}_{j+s/2,i}\right)+1\right) &\leq \frac{1}{r}\sum_{i=1}^{r}\left(-4\left((1-P_{\mathbf{z}_i})P_{\mathbf{z}_i}-\mu\right)+1\right)\\ &= \frac{1}{r}\sum_{i=1}^{r}\left((2P_{\mathbf{z}_i}-1)^2+4\mu\right) = \frac{1}{r}\sum_{i=1}^{r}(2P_{\mathbf{z}_i}-1)^2+4\mu\\ &\leq \mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2+\delta+4\mu.\end{aligned}$$

Hence, it follows that

$$\Pr(\textbf{est} \leq \mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2+\delta+4\mu) \geq \left(1-e^{-\frac{r\delta^2}{2}}\right)\left(1-r\cdot e^{-s\mu^2}\right),$$

which by symmetry implies that

$$\Pr(|\textbf{est}-\mathop{\mathbb{E}}_{\mathbf{x}} H(\mathbf{x})^2| \leq \delta+4\mu) \geq 2\left(1-e^{-\frac{r\delta^2}{2}}\right)\left(1-r\cdot e^{-s\mu^2}\right)-1.$$

## Appendix E
## Proof of the Coset Lemma

We begin with a quick sanity check.

**Lemma 7.** *If $h$ is a halfspace and $\sigma\in\operatorname{Aut}(G)$ then $h^\sigma$ is a halfspace as well.*

*Proof.* Let $\mathbf{w}\in\mathbb{R}^n$ be any vector that defines $h$, and denote $\sigma=(\pi,\mathbf{v})$. We have that

$$\begin{aligned}h^\sigma(\mathbf{x}) &= \operatorname{sign}(\mathbf{w}\cdot\sigma(\mathbf{x})) = \operatorname{sign}(\mathbf{w}\cdot(\pi(\mathbf{x})\oplus\mathbf{v}))\\ &\overset{(a)}{=} \operatorname{sign}((\mathbf{w}\star\mathbf{v})\cdot\pi(\mathbf{x})) \overset{(b)}{=} \operatorname{sign}(\pi^{-1}(\mathbf{w}\star\mathbf{v})\cdot\mathbf{x}),\end{aligned}$$

where $(a)$ holds since $\oplus$ is equivalent to multiplication over $\mathbb{R}$, and $(b)$ holds since

$$\begin{aligned}(\mathbf{w}\star\mathbf{v})\cdot\pi(\mathbf{x}) &= \sum_{i=1}^{n}(\mathbf{w}\star\mathbf{v})_i x_{\pi(i)} = \sum_{i=1}^{n}(\mathbf{w}\star\mathbf{v})_{\pi^{-1}(i)}x_i\\ &= \pi^{-1}(\mathbf{w}\star\mathbf{v})\cdot\mathbf{x}.\end{aligned}$$

□

To prove Lemma 3, we require the following auxiliary claim, which applies to both $\mathcal{H}_{uni}$ and $\mathcal{H}_{vol}$.

**Lemma 8.** *For $\sigma \in \text{Aut}(G)$ such that $\sigma(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{X}$ we have*
$(a)$ *$h^\sigma \in \mathcal{H}$ for every $h \in \mathcal{H}$; and*
$(b)$ *$\Pr(h) = \Pr(h^\sigma)$ for every $h \in \mathcal{H}$.*

*Proof.* Due to Lemma 7, to prove $(a)$ we are only left to show that $h^\sigma(\mathbf{x}_i) = y_i$ for every $i \in [k]$ and $h \in \mathcal{H}$. However, this is clear since $h^\sigma(\mathbf{x}_i) = h(\sigma(\mathbf{x}_i)) = h(\mathbf{x}_i) = y_i$.

Part $(b)$ is obvious for $\mathcal{H}_{uni}$. To prove $(b)$ for $\mathcal{H}_{vol}$, let $h \in \mathcal{H}$, and notice that it suffices to show that $\text{Vol}(\mathcal{V}_h) = \text{Vol}(\mathcal{V}_{h^\sigma})$. For $\mathbf{w} \in \mathbb{R}^n$ and $h \in \mathcal{H}$ let $\mathbb{1}(\mathbf{w}, h)$ be the $(0,1)$-indicator of the event "$\text{sign}(\mathbf{w} \cdot \mathbf{x}) = h(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{F}_2^n$", i.e., $\mathbb{1}(\mathbf{w}, h) = 1$ if and only if $\mathbf{w}$ defines $h$, and otherwise it is zero. Then, we have that

$$\text{Vol}(\mathcal{V}_h) = \int_{\mathcal{V}} \mathbb{1}(\mathbf{w}, h) d\mathbf{w}. \tag{13}$$

We perform the variable substitution $\mathbf{w} = \sigma(\mathbf{u})$, and since $\sigma$ is a linear operator whose determinant is either $1$ or $-1$, it follows that

$$(13) = \int_{\sigma^{-1}(\mathcal{V})} \mathbb{1}(\sigma(\mathbf{u}), h) d\mathbf{u}. \tag{14}$$

To show that (14) equals $\text{Vol}(\mathcal{V}_{h^\sigma})$, it suffices to show that $\sigma^{-1}(\mathcal{V}) = \mathcal{V}$ and that $\mathbb{1}(\sigma(\mathbf{u}), h) = \mathbb{1}(\mathbf{u}, h^\sigma)$ for every $\mathbf{u} \in \mathbb{R}^n$. To show the former, notice that

$$\begin{aligned} \sigma^{-1}(\mathcal{V}) &= \{\sigma^{-1}(\mathbf{w}) | \mathbf{w} \in \mathcal{V}\} = \{\mathbf{w} | \sigma(\mathbf{w}) \in \mathcal{V}\} \\ &= \{\mathbf{w} \in \mathbb{R}^n | y_i(\sigma(\mathbf{w}) \cdot \mathbf{x}_i) \geq 0 \text{ for every } i \in [k] \text{ and } \|\sigma(\mathbf{w})\|_2 \leq 1\}. \end{aligned} \tag{15}$$

Again, since $\sigma$ is a linear transform of determinant $\pm 1$, it follows that $\|\sigma(\mathbf{w})\|_2 = \|\mathbf{w}\|_2$ for every $\mathbf{w} \in \mathbb{R}^n$. In addition, by denoting $\sigma = (\pi, \mathbf{v})$ we have that

$$\begin{aligned} \sigma(\mathbf{w}) \cdot \mathbf{x}_i = (\pi(\mathbf{w}) \star \mathbf{v}) \cdot \mathbf{x}_i &= \pi(\mathbf{w}) \cdot (\mathbf{x}_i \oplus \mathbf{v}) \\ &= \sum_{j=1}^n w_{\pi(j)}(x_{i,j} \oplus v_j) \\ &= \sum_{j=1}^n w_j (x_{i,\pi^{-1}(j)} \oplus v_{\pi^{-1}(j)}) \\ &\overset{(\dagger)}{=} \sum_{j=1}^n w_j (\sigma^{-1}(\mathbf{x}_i))_j = \mathbf{w} \cdot \sigma^{-1}(\mathbf{x}_i), \end{aligned}$$

where $(\dagger)$ follows since $\sigma^{-1}(\mathbf{x}) = \pi^{-1}(\mathbf{x} \oplus \mathbf{v})$ for every $\mathbf{x} \in \mathbb{F}_2^n$. Therefore, it follows that

$$(15) = \{\mathbf{w} \in \mathbb{R}^n | y_i(\mathbf{w} \cdot \sigma^{-1}(\mathbf{x}_i)) \geq 0 \text{ for every } i \in [k] \text{ and } \|\mathbf{w}\|_2 \leq 1\}. \tag{16}$$

Now, since $\sigma(\mathbf{x}_i) = \mathbf{x}_i$, it follows that $\sigma^{-1}(\mathbf{x}_i) = \mathbf{x}_i$, and hence (16) implies that $\sigma^{-1}(\mathcal{V}) = \mathcal{V}$.

To prove that $\mathbb{1}(\sigma(\mathbf{u}), h) = \mathbb{1}(\mathbf{u}, h^\sigma)$ for every $\mathbf{u} \in \mathbb{R}^n$, (i.e., that $\sigma(\mathbf{u})$ defines $h$ if and only if $\mathbf{u}$ defines $h^\sigma$) it is shown that for every $\mathbf{u} \in \mathbb{R}^n$, we have that $h(\sigma(\mathbf{x})) = \text{sign}(\mathbf{u} \cdot \mathbf{x})$ for every $\mathbf{x} \in \mathbb{F}_2^n$ if and only if $h(\mathbf{x}) = \text{sign}(\sigma(\mathbf{u}) \cdot \mathbf{x})$ for every $\mathbf{x} \in \mathbb{F}_2^n$. Let $\mathbf{u} \in \mathbb{R}^n$, and assume that $h(\sigma(\mathbf{x})) = \text{sign}(\mathbf{u} \cdot \mathbf{x})$ for every $\mathbf{x} \in \mathbb{F}_2^n$. Then, (all subsequent expressions hold for every $\mathbf{x} \in \mathbb{F}_2^n$)

$$h(x_{\pi(1)} \oplus v_1, \ldots, x_{\pi(n)} \oplus v_n) = \text{sign}\left(\sum_{j=1}^n u_j x_j\right),$$

which is equivalent to

$$\begin{aligned} h(x_{\pi(1)}, \ldots, x_{\pi(n)}) &= \text{sign}\left(\sum_{j=1}^n u_j (x_j \oplus v_{\pi^{-1}(j)})\right) \\ &= \text{sign}\left(\sum_{j=1}^n (u_j v_{\pi^{-1}(j)}) \cdot x_j\right) \\ &= \text{sign}\left(\sum_{j=1}^n (u_{\pi(j)} v_j) \cdot x_{\pi(j)}\right). \end{aligned}$$

Now, by substituting $x_{\pi(i)}$ with $x_i$, we get

$$\begin{aligned} h(\mathbf{x}) = h(x_1, \ldots, x_n) &= \operatorname{sign}\left(\sum_{j=1}^{n} (u_{\pi(j)} v_j) \cdot x_j\right) \\ &= \operatorname{sign}\left((\pi(\mathbf{u}) \star \mathbf{v}) \cdot \mathbf{x}\right) = \operatorname{sign}(\sigma(\mathbf{u}) \cdot \mathbf{x}), \end{aligned}$$

and hence $\sigma(\mathbf{u})$ defines $h$. The converse is proved by iterating identical steps in a reversed order. Therefore, we have that

$$\int_{\sigma^{-1}(\mathcal{V})} \mathbb{1}(\sigma(\mathbf{u}), h) d\mathbf{u} = \int_{\mathcal{V}} \mathbb{1}(\mathbf{u}, h^{\sigma}) d\mathbf{u} = \operatorname{Vol}(\mathcal{V}_{h^{\sigma}}),$$

and hence $\Pr(h) = \Pr(h^{\sigma})$ in $\mathcal{H}_{vol}$ as well. □

*Proof.* (of Lemma 3) Since $|\mathcal{C}_1| = |\mathcal{C}_2|$, it follows that for every $h_1, h_2 \in \mathcal{H}$, we have that

$$\begin{aligned} d_{\mathcal{C}_1}(h_1^{\sigma}, h_2^{\sigma}) &= \frac{1}{|\mathcal{C}_1|} \sum_{\mathbf{c} \in \mathcal{C}_1} \frac{1 - h_1^{\sigma}(\mathbf{c}) h_2^{\sigma}(\mathbf{c})}{2} \\ &= \frac{1}{|\mathcal{C}_2|} \sum_{\mathbf{c} \in \mathcal{C}_2} \frac{1 - h_1^{\sigma}(\sigma^{-1}(\mathbf{c})) h_2^{\sigma}(\sigma^{-1}(\mathbf{c}))}{2} \\ &= \frac{1}{|\mathcal{C}_2|} \sum_{\mathbf{c} \in \mathcal{C}_2} \frac{1 - h_1(\mathbf{c}) h_2(\mathbf{c})}{2} = d_{\mathcal{C}_2}(h_1, h_2). \end{aligned}$$

Hence, since $h_1^{\sigma}, h_2^{\sigma} \in \mathcal{H}$ by Lemma 8($a$), it follows that for every pair of functions $h_1, h_2 \in \mathcal{H}$ there exists a respective pair of functions $h_1^{\sigma}, h_2^{\sigma} \in \mathcal{H}$ such that $d_{\mathcal{C}_1}(h_1^{\sigma}, h_2^{\sigma}) = d_{\mathcal{C}_2}(h_1, h_2)$. Moreover, it follows from Lemma 8($b$) that

$$\begin{aligned} \mathop{\mathbb{E}}_{h_1, h_2} d_{\mathcal{C}_2}(h_1, h_2) &= \sum_{h_1, h_2 \in \mathcal{H}} \Pr(h_1) \Pr(h_2) d_{\mathcal{C}_2}(h_1, h_2) \\ &= \sum_{h_1, h_2 \in \mathcal{H}} \Pr(h_1^{\sigma}) \Pr(h_2^{\sigma}) d_{\mathcal{C}_1}(h_1^{\sigma}, h_2^{\sigma}), \end{aligned}$$

and since the mapping $h \mapsto h^{\sigma}$ is an injective map from $\mathcal{H}$ to itself, we have

$$\begin{aligned} \sum_{h_1, h_2 \in \mathcal{H}} \Pr(h_1^{\sigma}) \Pr(h_2^{\sigma}) d_{\mathcal{C}_1}(h_1^{\sigma}, h_2^{\sigma}) &= \sum_{h_1, h_2 \in \mathcal{H}} \Pr(h_1) \Pr(h_2) d_{\mathcal{C}_1}(h_1, h_2) \\ &= \mathop{\mathbb{E}}_{h_1, h_2} d_{\mathcal{C}_1}(h_1, h_2), \end{aligned}$$

which concludes the proof. □

## APPENDIX F
## SUBCUBE LEMMAS

**Lemma 9.** *If $\mathcal{X}$ is a $(\boldsymbol{v}, I)$-subcube for some $I = \{i_j\}_{j=1}^{\ell}$ and $\boldsymbol{v} \in \mathbb{F}_2^n$, then $\operatorname{Stab}(\mathcal{X}) = \{\sigma = (\pi, \pi(\boldsymbol{v}) \oplus \boldsymbol{v}) | \pi \in S_I\}$, where $S_I$ is the set of all permutations $\pi$ in $S_n$ such that $\pi(i) = i$ for every $i \in I$.*

*Proof.* Let $\sigma = (\pi, \pi(\mathbf{v}) \oplus \mathbf{v})$ for $\pi \in S_I$, and let $\mathbf{x} = a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v} \in \mathcal{X}$ for some $a_i$'s in $\mathbb{F}_2$. Then,

$$\begin{aligned} \sigma(\mathbf{x}) &= \pi(\mathbf{x}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v} \\ &= \pi(a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v} \\ &\overset{(\dagger)}{=} a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \pi(\mathbf{v}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v} \\ &= a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v} = \mathbf{x}, \end{aligned}$$

where $(\dagger)$ follows since $\pi$ is a linear transform and since $\pi(\mathbf{e}_i) = \mathbf{e}_i$ for every $i \in I$. Therefore, it follows that $\{(\pi, \pi(\mathbf{v}) \oplus \mathbf{v}) | \pi \in S_I\} \subseteq \operatorname{Stab}(\mathcal{X})$.

Conversely, let $\sigma = (\pi, \mathbf{u}) \in \operatorname{Stab}(\mathcal{X})$. If $\pi \notin S_I$ then there exists $i \in I$ and $j \neq i$ such that $\pi(j) = i$. If $j \in I$ then any $\mathbf{x} \in \mathcal{X}$ such that $x_i \neq x_j$ is not mapped to itself by $\sigma$. If $j \notin I$ then any $\mathbf{x} \in \mathcal{X}$ such that $x_i \neq u_j$ is not mapped to itself. Therefore, it must be that $\pi \in S_I$. Now let $\mathbf{x} = a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v} \in \mathcal{X}$ for some $a_i$'s in $\mathbb{F}_2$. Since $\sigma(\mathbf{x}) = \mathbf{x}$, it follows that

$$\pi(a_1 \mathbf{e}_{i_1} \oplus \ldots a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v}) \oplus \mathbf{u} = a_1 \mathbf{e}_{i_1} \oplus \ldots \oplus a_{\ell} \mathbf{e}_{i_{\ell}} \oplus \mathbf{v},$$

and

$$a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell \mathbf{e}_{i_\ell} \oplus \pi(\mathbf{v}) \oplus \mathbf{u} = a_1\mathbf{e}_{i_1} \oplus \ldots a_\ell \mathbf{e}_{i_\ell} \oplus \mathbf{v},$$

and therefore $\mathbf{u} = \pi(\mathbf{v}) \oplus \mathbf{v}$. □

*Proof.* (of Lemma 4) Let $\mathbf{u}, \mathbf{w} \in \mathbb{F}_2^n$ be two vectors with identical Hamming weight on $[n] \setminus I$ and $u_i = w_i = 1$ for every $i \in I$. Therefore, there exists a permutation $\pi \in S_I$ such that $\pi(\mathbf{u}) = \mathbf{w}$. For $\mathcal{C}_\mathbf{u} \triangleq \mathcal{X} \oplus \mathbf{u}$ and $\mathcal{C}_\mathbf{w} \triangleq \mathcal{X} \oplus \mathbf{w}$ we show that $\sigma(\mathcal{C}_\mathbf{u}) = \mathcal{C}_\mathbf{w}$, where $\sigma = (\pi, \pi(\mathbf{v}) \oplus \mathbf{v})$.

For every $\mathbf{c} \in \mathcal{C}_\mathbf{u}$ there exist $a_1, \ldots, a_\ell \in \mathbb{F}_2$ such that $\mathbf{c} = a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \mathbf{v} \oplus \mathbf{u}$. Therefore,

$$\begin{aligned}\sigma(\mathbf{c}) &= \pi(\mathbf{c}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v}\\ &= a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \pi(\mathbf{v}) \oplus \pi(\mathbf{u}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v}\\ &= a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \mathbf{w} \oplus \mathbf{v} \in \mathcal{X} \oplus \mathbf{w} = \mathcal{C}_\mathbf{w},\end{aligned}$$

which readily implies that $\sigma(\mathcal{C}_\mathbf{u}) = \mathcal{C}_\mathbf{w}$. Hence, it follows that any two cosets $\mathcal{C}_\mathbf{u}$ and $\mathcal{C}_\mathbf{w}$ such that $\mathbf{u}$ and $\mathbf{w}$ have identical Hamming weight on $[n] \setminus I$ reside in the same orbit.

We now prove that any $\mathcal{C}_\mathbf{u}$ and $\mathcal{C}_\mathbf{w}$ such that $\mathbf{u}$ and $\mathbf{w}$ differ in their Hamming weight on $[n] \setminus I$ are in *different* orbits. Assuming otherwise, we have some $\sigma = (\pi, \pi(\mathbf{v}) \oplus \mathbf{v}) \in \mathrm{Stab}(\mathcal{X})$ such that $\sigma(\mathcal{C}_\mathbf{u}) = \mathcal{C}_\mathbf{w}$, which implies that for any $\mathbf{c} = a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \mathbf{v} \oplus \mathbf{u} \in \mathcal{C}_\mathbf{u}$ we have $\sigma(\mathbf{c}) \in \mathcal{C}_\mathbf{w}$, i.e.,

$$\begin{aligned}\pi(\mathbf{c}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v} &\in \mathcal{C}_\mathbf{w}\\ a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \pi(\mathbf{v}) \oplus \pi(\mathbf{u}) \oplus \pi(\mathbf{v}) \oplus \mathbf{v} &\in \mathcal{C}_\mathbf{w}\\ a_1\mathbf{e}_{i_1} \oplus \ldots \oplus a_\ell\mathbf{e}_{i_\ell} \oplus \pi(\mathbf{u}) \oplus \mathbf{v} &\in \mathcal{C}_\mathbf{w}.\end{aligned}$$

Now, since $\mathbf{v}$, $\mathbf{u}$, and $\mathbf{w}$ have no $-1$ entries on $I$, and since $\pi \in S_I$, it follows that $\pi(\mathbf{u}) \oplus \mathbf{v} = \mathbf{w} \oplus \mathbf{v}$, i.e., that $\pi(\mathbf{u}) = \mathbf{w}$. However, $\mathbf{w}$ and $\mathbf{u}$ are of different Hamming weights, which is a contradiction.

Hence, we have that the cosets of $\mathcal{X}$ are partitioned according to the weight of their shift vector. That is, there are $n - |I|$ cosets $\mathcal{O}_1, \ldots, \mathcal{O}_{n-|I|}$, and a coset $\mathcal{X} \oplus \mathbf{u}$ lies in $\mathcal{O}_{w_H(\mathbf{u})}$, where $w_H$ denotes Hamming weight. □

## APPENDIX G
## CONCENTRATION OF BINOMIAL COEFFICIENTS

**Lemma 10.** *Let $B \triangleq \{\frac{q}{2} - \lfloor c\sqrt{q} \rfloor, \ldots, \frac{q}{2} + \lfloor c\sqrt{q} \rfloor\}$ for some constant $c > 0$, where $q \triangleq n - |I|$. Then, for large enough $q$ we have that*

$$\sum_{r \in B} \frac{\binom{q}{r}}{2^q} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_r}(h_1,h_2) \leq \mathop{\mathbb{E}}_{h_1,h_2} d(h_1,h_2) \leq \sum_{r \in B} \frac{\binom{q}{r}}{2^q} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_r}(h_1,h_2) + (1 - C(2c) + C(-2c)),$$

*where $C(x) = \frac{1}{2}\left(1 + \mathit{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$ is the cumulative distribution function (CDF) of a standard normal random variable $\mathcal{N}(0,1)$.*

A simple numeric approximation of $C(x)$ shows that $1 - C(2c) + C(-2c)$ approaches zero very fast as $c$ grows. Hence, we have $\mathbb{E}_{h_1,h_2} d(h_1,h_2) \approx \mathbb{E}_{h_1,h_2} \sum_{r \in B} \binom{q}{r} 2^{-q} d_{\mathcal{C}_i}(h_1,h_2)$. In the latter expression the contribution of every sampled $h_1, h_2 \in \mathcal{H}$ can be computed *exactly* in $O(nk\sqrt{n - |I|})$ time.

*Proof.* (of Lemma 10) The lower bound is trivial from Corollary 2. To prove the upper bound, notice that

$$\begin{aligned}\mathop{\mathbb{E}}_{h_1,h_2} d(h_1,h_2) &\leq \sum_{r \in B} \frac{\binom{q}{r}}{2^q} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_r}(h_1,h_2) + \sum_{r \notin B} \frac{\binom{q}{r}}{2^q} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_r}(h_1,h_2)\\ &\leq \sum_{r \in B} \frac{\binom{q}{r}}{2^q} \mathop{\mathbb{E}}_{h_1,h_2} d_{\mathcal{C}_r}(h_1,h_2) + \sum_{r \notin B} \frac{\binom{q}{r}}{2^q}.\end{aligned}$$

According to Lemma 11 which is proved shortly, we have that $\lim_{q \to \infty} \sum_{r \notin B} \frac{\binom{q}{r}}{2^q} = 1 - C(2c) + C(-2c)$, which concludes the claim. □

We are left with an exercise in probability theory, whose proof requires the central limit theorem [14], and a full proof is given for completeness. In what follows, let $\Sigma_q \triangleq \sum_{i=1}^q X_i$ for i.i.d $X_i = \mathit{Bern}(1/2)$. Further, as in Lemma 10, let $B = \{\frac{q}{2} - \lfloor c\sqrt{q} \rfloor, \ldots, \frac{q}{2} + \lfloor c\sqrt{q} \rfloor\}$ for some constant $c$.

**Lemma 11.** $\lim_{q \to \infty} \sum_{r \notin B} \frac{\binom{q}{r}}{2^q} = 1 - C(2c) + C(-2c)$*, where $C$ is the CDF of $\mathcal{N}(0,1)$.*

*Proof.* Clearly, we have that the probability of $\Sigma_q$ to have a value in $B$ is $\sum_{r \in B} \frac{\binom{q}{r}}{2^q}$. Furthermore, this probability can be written as

$$\begin{aligned}\Pr(\Sigma_q \in B) &= \Pr\left(\frac{q}{2} - \lfloor c\sqrt{q} \rfloor \leq \Sigma_q \leq \frac{q}{2} + \lfloor c\sqrt{q} \rfloor\right) \\ &= \Pr\left(\frac{q}{2} - c\sqrt{q} \leq \Sigma_q \leq \frac{q}{2} + c\sqrt{q}\right) \\ &= \Pr\left(-2c \leq \frac{2\Sigma_q - q}{\sqrt{q}} \leq 2c\right).\end{aligned}$$

Since $\mathbb{E}[Bern(1/2)] = \frac{1}{2}$ and $\sigma^2(Bern(1/2)) = \frac{1}{4}$, it follows that

$$\frac{\sum_{i=1}^q X_i - \sum_{i=1}^q \mathbb{E}[X_i]}{\sqrt{\sum_{i=1}^q \sigma^2(X_i)}} = \frac{\Sigma_q - \frac{q}{2}}{\sqrt{\frac{q}{4}}} = \frac{2\Sigma_q - q}{\sqrt{q}}.$$

Therefore, a straightforward application of the central limit theorem implies that

$$\begin{aligned}\lim_{q \to \infty} \sum_{r \in B} \frac{\binom{q}{r}}{2^q} &= \lim_{q \to \infty} \Pr\left(-2c \leq \frac{2\Sigma_q - q}{\sqrt{q}} \leq 2c\right) \\ &= \Pr(-2c \leq \mathcal{N}(0,1) \leq 2c).\end{aligned}$$

Hence, it follows that $\lim_{q \to \infty} \sum_{r \in B} \frac{\binom{q}{r}}{2^q} = C(2c) - C(-2c)$, which implies the claim. □