Learning Geometric Concepts via Gaussian Surface Area

Adam R. Klivans
University of Texas at Austin
klivans@cs.utexas.edu

Ryan O'Donnell
Carnegie Mellon University
odonnell@cs.cmu.edu

Rocco A. Servedio*
Columbia University
rocco@cs.columbia.edu

December 23, 2008

Abstract

We study the learnability of sets in $\mathbb{R}^n$ under the Gaussian distribution, taking Gaussian *surface area* as the "complexity measure" of the sets being learned. Let $\mathcal{C}_S$ denote the class of all (measurable) sets with surface area at most S. We first show that the class $\mathcal{C}_S$ is learnable to any constant accuracy in time $n^{O(S^2)}$, even in the arbitrary noise ("agnostic") model. Complementing this, we also show that any learning algorithm for $\mathcal{C}_S$ information-theoretically requires $2^{\Omega(S^2)}$ examples for learning to constant accuracy. These results together show that Gaussian surface area essentially characterizes the computational complexity of learning under the Gaussian distribution.

Our approach yields several new learning results, including the following (all bounds are for learning to any constant accuracy):

- The class of *all* convex sets can be agnostically learned in time $2^{\tilde{O}(\sqrt{n})}$ (and we prove a $2^{\Omega(\sqrt{n})}$ lower bound for noise-free learning). This is the first subexponential time algorithm for learning general convex sets even in the noise-free (PAC) model.
- Intersections of k halfspaces can be agnostically learned in time $n^{O(\log k)}$ (cf. Vempala's $n^{O(k)}$ time algorithm for learning in the noise-free model [Vem04]).
- Arbitrary cones (with apex centered at the origin), and spheres with arbitrary radius and center, can be agnostically learned in time $\mathrm{poly}(n)$.

*Supported in part by NSF award CCF-0347282 and by NSF award CCF-0523664.

1 Introduction

1.1 Motivation: What is the right measure of complexity for learning? The primary goal of computational learning theory is to understand how the resources required by learning algorithms (running time, number of examples, etc.) scale with the complexity of the functions being learned. For sample complexity our understanding is quite good: it has been known for nearly 20 years that for any class $\mathcal{C}$ of Boolean functions, the Vapnik-Chervonenkis dimension of $\mathcal{C}$ gives essentially matching upper and lower bounds on the number of examples needed for learning $\mathcal{C}$ with respect to an arbitrary (unknown) probability distribution over the space of examples [BEHW89, EHKV89]. Unfortunately, it has proved much more difficult to characterize the *computational* complexity of learning problems. This difficulty is particularly acute in distribution-independent learning models such as Valiant's original PAC learning model [Val84]; as one example of this, our current state of knowledge is consistent both with the possibility that learning an intersection of two n-dimensional halfspaces (under arbitrary distributions) can be done in $O(n^2)$ time, and with the possibility that this learning problem requires time $2^{\Omega(n)}$. In general, research progress on computationally efficient distribution-independent learning has been relatively slow, and for this reason many researchers have considered learning with respect to specific natural distributions such as the uniform distribution on the n-dimensional Boolean hypercube and the uniform distribution on the unit Euclidean sphere in $\mathbb{R}^n$.

In this work we consider learning with respect to the standard Gaussian distribution on $\mathbb{R}^n$. This is arguably one of the most natural, and certainly one of the most studied, distributions on $\mathbb{R}^n$. We note that the commonly studied scenario of learning with respect to the uniform distribution on the n-dimensional Euclidean sphere (see e.g. [BK97, Vem04, Lon94, Lon95, KKMS05]) is essentially equivalent to learning under the standard Gaussian distribution when n is large. (As we shall see in Section 4.5, almost all of our learning results actually hold for arbitrary Gaussian distributions on $\mathbb{R}^n$).

As our main contribution, we propose a new and very natural complexity measure for geometric concepts $A \subset \mathbb{R}^n$, their *Gaussian surface area*, and show that this measure characterizes the computational complexity of learning with respect to the Gaussian distribution in a rather strong sense. We do this by giving essentially matching upper bounds (via an explicit algorithm) and lower bounds (information-theoretic) on the running time of algorithms for learning sets $A \subset \mathbb{R}^n$ in terms of their Gaussian surface area. Furthermore (and perhaps most importantly), this approach yields striking new applications for learning important concept classes such as arbitrary convex sets and intersections of halfspaces.

1.2 The new complexity measure: Gaussian Surface Area. The formal definition of Gaussian surface area is as follows:

Definition 1. *For a Borel set $A \subset \mathbb{R}^n$, its* Gaussian surface area *is*

$$\Gamma(A) \stackrel{def}{=} \liminf_{\delta \to 0} \frac{\mathrm{vol}(A_\delta \setminus A)}{\delta}.$$

Here A_δ denotes the δ-neighborhood of A, $\{x : \mathrm{dist}(x, K) \leq \delta\}$, and $\mathrm{vol}(A)$ denotes the probability mass of A with respect to the standard Gaussian distribution on $\mathbb{R}^n$.

This is very similar to the usual formal definition of surface area, except that Gaussian measure replaces Lebesgue measure. For most "nice" sets A we can take an equivalent definition (see [Naz03b]):

Definition 2. *If $A \subset \mathbb{R}^n$ is sufficiently regular — e.g., has smooth boundary or is convex — then we have*

$$\Gamma(A) = \int_{\partial A} \varphi_n(x)\, d\sigma(x), \tag{1}$$

where $d\sigma(x)$ is the standard surface measure in $\mathbb{R}^n$ and

$$\varphi_n(x) \stackrel{def}{=} \prod_{i=1}^{n} \varphi(x_i), \quad \textit{where } \varphi(x) = \varphi_1(x) \stackrel{def}{=} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$$

is the standard n-dimensional Gaussian density function.

It is straightforward to see from Definition 1 that the Gaussian surface area of a set is smaller than its usual surface area by at least an exponential factor:

Fact 3. *If $A \subset \mathbb{R}^n$ is any measurable set and $\mathrm{surf}(A)$ denotes its usual surface area, then $\Gamma(A) \leq \frac{1}{(2\pi)^{n/2}}\mathrm{surf}(A)$.*

In fact, the Gaussian surface area of A is often far smaller even than this; many natural sets A have *infinite* $\mathrm{surf}(A)$ but small finite Gaussian surface area $\Gamma(A)$. The most notable example is that of halfspaces:

Fact 4. *Every halfspace in $\mathbb{R}^n$ has Gaussian surface area at most $\varphi(0) = \sqrt{2/\pi} \approx 0.8$.*

This is a classical fact because of the "Gaussian isoperimetric inequality" [Bor75, ST78] (see also [Bob97]), which states that halfspaces minimize $\Gamma(A)$ among all sets A with fixed Gaussian volume.

In the remainder of this paper we will use the phrase "surface area" exclusively to mean Gaussian surface area, Γ**.** The following table gives the surface area of some basic geometric sets:

Sets	(Gaussian) Surface area upper bound	Source
Halfspaces	$\sqrt{2/\pi}$	direct computation
Intersections of k halfspaces	$O(\sqrt{\log k})$	Nazarov [Naz03a] (see Section 4)
Arbitrary convex sets	$O(n^{1/4})$	K. Ball [Bal93]
Balls	1	Section 4
Cones with apex at the origin	1	Section 4

We believe that surface area is a very natural complexity measure for sets in $\mathbb{R}^n$. First, it is a universal measure: it assigns a complexity to all sets. Second, is a natural geometric notion befitting geometric sets. Third, surface area seems to address the difficulty of learning sets in a fair way: if a set's boundary is very "wiggly", it is reasonable that many examples will be needed to accurately delineate it. Finally, it is a very stringent measure: as discussed above Gaussian surface area is in general very low.

1.3 Our main results. We give upper and lower bounds for learning sets of surface area S under the Gaussian distribution on $\mathbb{R}^n$. Our algorithmic result is an agnostic learning algorithm running in time $n^{O(S^2)}$. More precisely, the time is $n^{O(S^2/\epsilon^4)}$ for agnostic learning to accuracy ϵ and time $n^{O(S^2/\epsilon^2)}$ for PAC learning. (Agnostic learning may be thought of as a challenging model of learning in the presence of noise.) We give precise definitions of the learning models in Section 2.2, and precise statements of the algorithmic results as Theorem 25 in Section 4.5.

Our lower bound is information-theoretic and applies even to PAC learning algorithms under the Gaussian distribution (no noise) which have membership query access to the function being learned. We show that there is a universal constant $\epsilon_0 > 0$ such that any algorithm for learning sets of surface area at most S to accuracy ϵ_0 requires at least $2^{\Omega(S^2)}$ examples. This holds for any $\sqrt{\log n}/\epsilon_0 \leq S \leq \epsilon_0 n^{1/4}$, and is true even if the sets are promised to be intersections of $2^{\Theta(S^2)}$ halfspaces. We give this lower bound in Section 5.

We believe the main applications of our results are the following two algorithmic consequences, Theorems 5 and 6:

Theorem 5. *The class of all convex sets is agnostically learnable under* any *Gaussian distribution on $\mathbb{R}^n$ in subexponential time: $2^{\tilde{O}(n^{1/2}/\epsilon^4)}$. Further, $2^{\Omega(n^{1/2})}$ examples and hence time is necessary.*

We view Theorem 5 as somewhat surprising, since the general class of convex sets is extremely broad. (We recall that simple VC-dimension arguments can be used to show that for distribution-independent learning, *no* a priori running time bound — $2^n, 2^{2^n}$, etc. — can be given for learning arbitrary convex sets, see e.g. Chapter 4 of [KV94].) Theorem 5 is the first subexponential time algorithm for either agnostically or PAC learning arbitrary convex sets with respect to a non-trivial class of high-dimensional distributions. We note that Theorem 5 can be extended to learn *non-convex* concepts such as finite unions of convex sets; we defer statements of these results to Section 4.

Theorem 6. *Intersections of k halfspaces are agnostically learnable under* any *Gaussian distribution on $\mathbb{R}^n$ in time $n^{O(\log k/\epsilon^4)}$.*

Theorem 6 should be compared with Vempala's $(n/\epsilon)^{O(k)}$ time PAC learning algorithm (under nearly-uniform distributions on the sphere). Vempala's dependence on ϵ is better than ours if $\log(1/\epsilon) \gg \log k$, but otherwise our algorithm has a much better dependence on n, and also works in the agnostic setting.

The fact that Theorems 5 and 6 hold for any Gaussian distribution, as opposed to just the standard one, is an immediate easy consequence of the fact that convex sets and intersections of k halfspaces are closed under linear transformations; see Section C. We give several other new learning results in Section 4 as well.

Uniform Distribution over $\{-1, 1\}^n$. It is natural to ask whether our approach can be translated to the Boolean setting with respect to the uniform distribution on the hypercube. We establish a general connection between Boolean perimeter and learnability, and give tight bounds on the perimeter of Boolean halfspaces (e.g., we show that Boolean halfspaces have Boolean perimeter $\Theta(\sqrt{\log n})$). At this stage, however, this approach does not (yet) lead to new learning results for any well-studied concept classes, so we defer this discussion to Appendix E.

1.4 Our techniques. To broadly outline the proof of our main results, we begin with the result of Kalai et al. [KKMS05], which uses a type of polynomial regression to give an agnostic learning algorithm for functions that can be approximated well by low-degree polynomials. To use this result, we need to understand how well sets in $\mathbb{R}^n$ can be approximated (in ℓ_2) by polynomials with respect to the Gaussian distribution. This task can be separated into two parts:

First, we establish a new connection between the Hermite concentration of the characteristic function of a set (which captures the approximability by low-degree polynomials) and the set's Gaussian surface area. This reduction from learning to bounding surface area makes use of some powerful tools in geometry; especially, the use of semigroup tools in the study of isoperimetry.

Secondly, with this reduction in hand, we can translate bounds on Gaussian surface area to learning results. For example, K. Ball [Bal93] (and subsequently F. Nazarov [Naz03b]) has shown that the surface area of *any* convex set in n dimensions is at most $O(n^{1/4})$. Ball's result, combined with Theorem 25, gives us Theorem 5. We also prove new results on Gaussian surface area for various classes (see the table above) and obtain corresponding learning results for those classes.

Our lower bound is proved by analyzing geometric properties of intersections of randomly chosen halfspaces via concentration inequalities and may be of independent interest.

1.5 Relationship with Fourier-Based Learning. Our main result can be viewed as a statement regarding the approximability of characteristic functions of sets via low-degree orthogonal (Hermite) polynomials with respect to Gaussian distributions. More specifically, we prove that every indicator function of a (Borel) set with surface area S can be approximated (in ℓ_2) by a multivariate polynomial of degree $O(S^2)$.

Since we are considering approximability in ℓ_2 with respect to a family of orthogonal polynomials, our algorithm can be viewed as a Fourier-type algorithm over $\mathbb{R}^n$. A relevant paper for comparison is the work of Klivans et al. [KOS04], which also learned intersections of halfspaces– although with respect to the uniform distribution over $\{-1, 1\}^n$– by showing that these concepts can be approximated well (in ℓ_2) over $\{-1, 1\}^n$ by low-degree polynomials.

The Klivans et al. result [KOS04] bounds the Fourier concentration of a Boolean function (approximability by low-degree polynomials) in terms of the noise stability of that function. They then apply (simple) bounds on the Boolean noise stability of halfspaces to obtain their main algorithmic results.

Similar to the strategy of Klivans et al., as one part of our framework here we bound the Hermite concentration of the characteristic function of a set in $\mathbb{R}^n$ in terms of that function's Gaussian noise stability. In this work, however, we then face a significant stumbling block: we do not know how to directly bound the Gaussian noise stability of any interesting classes of sets in $\mathbb{R}^n$. (In contrast, [KOS04] gives direct and elementary proofs of upper bounds on the Boolean noise stability of halfspaces and intersections of halfspaces.) To get around this, we must appeal to deep theorems in convex geometry to show that the Gaussian noise stability of a set's characteristic function can in fact be bounded by the set's Gaussian surface area. Moreover, some of the actual bounds on Gaussian surface area that we subsequently use are highly non-trivial (e.g. [Bal93]). While we do not establish deep technical results in convex geometry in this paper, we do give the first bounds on Gaussian surface area for simple concept classes, such as balls, that may be of independent interest. We also believe that the link we establish between Gaussian surface area and learnability will likely lead to further algorithmic learning results beyond those presented in this paper.

1.6 Comparison with Previous Work. Let us briefly discuss prior algorithmic results for the specific learning problems we address. We note that learning intersections of halfspaces is one of the most well-studied problems in computational learning theory, see e.g. [Bau90a, BK97, KP98, KOS04, KS04, KS06, Vem04]. In particular, the work of Blum and Kannan [BK97] and subsequently Vempala [Vem04] specifically addressed the problem of PAC learning an intersection of k halfspaces to accuracy ϵ under the uniform distribution on the n-dimensional Euclidean sphere (very similar to the spherical Gaussian distribution). The algorithms of [BK97], [Vem04] are not known to work in the agnostic setting. Kalai et al. [KKMS05] gave the first polynomial-time algorithm for agnostically learning a single halfspace with respect to any Gaussian distribution in $\mathbb{R}^n$. We note here that the Kalai et al. result follows easily from our framework and the classical $O(1)$ bound on the Gaussian surface area of a halfspace.

Learning general convex sets is well known to be a broad and difficult problem, and we are not aware of any prior positive results for learning arbitrary convex sets in $\mathbb{R}^n$. As far as we can tell, our result is the first non-trivial algorithm for learning convex sets with respect to an interesting distribution. Baum [Bau90b] gave a simple algorithm for learning convex subsets of the unit square $[0, 1]^2$ under the uniform distribution based on "gridding"; it is possible to extend this to an algorithm for learning convex subsets of $[0, 1]^n$ under the uniform distribution, but the resulting algorithm has running time at least 2^n.

It is straightforward to see that arbitrary balls (or even ellipsoids) in n dimensions can be PAC learned in polynomial time. The problem of agnostically learning balls, however, is known to be NP-hard if the output hypothesis must also be a ball (that is, the proper agnostic learning problem is NP-hard) [BDEL03, BB05]. We give the first polynomial-time algorithm for agnostically learning balls (of arbitrary radius and center); our output hypothesis is the sign of a low-degree polynomial.

1.7 Organization. In Section 2 we review Fourier and Hermite analysis, learning models, and Gaussian surface area. In Section 3 we establish a connection between Hermite concentration and Gaussian surface area. In Section 4, we show how to bound the surface area of various convex sets and state our new learning results. In Section 4.5 we extend our results to non-spherical Gaussians, and state our most general positive result establishing agnostic learnability of a class of functions in terms of the surface area of the corresponding sets. In Section 5, we prove our main lower bound which shows that several of our positive results are essentially optimal. We give results on the Boolean setting in Appendix E.

2 Preliminaries

2.1 Gaussian distributions, Hermite analysis, Ornstein-Uhlenbeck, Perimeter

Gaussian distributions. We will be working with Gaussian probability distributions on $\mathbb{R}^n$. For the most part we will restrict attention to the standard n-dimensional Gaussian distribution, $\mathcal{N}(0, I_n)$, with mean 0 and independent, variance-1 coordinates. This has density function $\varphi_n(x)$ as defined in Section 1.2. As discussed in Section 4.5, most of our results generalize to *arbitrary* n-variate Gaussian distributions, even with singular covariance matrices. Unless otherwise specified, though, all integrals and expectations in this paper are with respect to the standard distribution, which we abbreviate by $\mathcal{N}^n$.

Hermite analysis. We will work within $L^2(\mathbb{R}^n, \mathcal{N}^n)$, the vector space of all functions $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbf{E}[f^2] < \infty$. This is an inner product space under the inner product $\langle f, g\rangle = \mathbf{E}_{x\sim\mathcal{N}^n}[f(x)g(x)]$. This inner product space has a complete orthonormal basis given by the *Hermite polynomials*. In the case $n = 1$, these are the polynomials $h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = \frac{x^2-1}{\sqrt{2}}, \quad h_3(x) = \frac{x^3-3x}{\sqrt{6}}, \quad \ldots$

For general n, the basis for $L^2(\mathbb{R}^n, \mathcal{N}^n)$ is formed by all products of these polynomials, one for each coordinate. I.e., for each n-tuple $S \in \mathbb{N}^n$ we define the n-variate Hermite polynomial $H_S : \mathbb{R}^n \to \mathbb{R}$ by $H_S(x) = \prod_{i=1}^n h_{S_i}(x_i)$; then the collection $(H_S)_{S\in\mathbb{N}^n}$ is a complete orthonormal basis for the inner product space. All of the "standard" facts of Fourier analysis hold here: every function $f \in L^2$ can be written uniquely as $\sum_{S\in\mathbb{N}^n} \hat{f}(S)H_S(x)$ and we have $\lim_{d\to\infty} \mathbf{E}\left[\left(f(x) - \sum_{|S|\le d} \hat{f}(S)H_S(x)\right)^2\right] = 0$ (here $|S| = \sum_i S_i$ is the total degree of $H_S(x)$ as a polynomial). Each coefficient, $\hat{f}(S)$, is the Fourier or *Hermite* coefficient of f and is equal to $\mathbf{E}_{x\sim\mathcal{N}^n}[f(x)H_S(x)]$. We also have Parseval's and Plancherel's identity. For a few more details see Appendix A.

Ornstein-Uhlenbeck. For each $0 \le t \le \infty$ one can define a (bounded) linear operator P_t on $L^2(\mathbb{R}^n, \mathcal{N}^n)$, the *Ornstein-Uhlenbeck* operator. These operators map a function $f : \mathbb{R}^n \to \mathbb{R}$ to another function $P_t f : \mathbb{R}^n \to \mathbb{R}$ via

$$(P_t f)(x) \stackrel{\text{def}}{=} \mathbf{E}_{y\sim\mathcal{N}^n}[f(e^{-t}x + \sqrt{1 - e^{-2t}}y)].$$

The parameterization here with e^{-t} is traditionally chosen so that the operators form a semigroup: $P_{t_1} \circ P_{t_2} = P_{t_1+t_2}$. Since we will not use this property, we prefer to redefine the operators as follows: For $\rho \in [0, 1]$,

$$(T_\rho f)(x) \stackrel{\text{def}}{=} \mathbf{E}_{y\sim\mathcal{N}^n}[f(\rho x + \sqrt{1 - \rho^2}y)].$$

We thus have $P_t = T_{e^{-r}}$. Alternately stated, $T_\rho f(x)$ is the average value of f under the shifted and scaled Gaussian distribution $\mathcal{N}(\rho x, \sqrt{1 - \rho^2}I_n)$. The fact that T_ρ is a linear operator — i.e., $T_\rho(f+g) = T_\rho f + T_\rho g$ — follows immediately from linearity of expectation.

A key property of T_ρ that we will use is how it operates with respect to the Hermite expansion. Specifically, it can be shown that $T_\rho H_S = \rho^{|S|}H_S$, and hence (by linearity)

$$T_\rho f = \sum_{S\in\mathbb{N}^n} \rho^{|S|}\hat{f}(S)H_S. \qquad (2)$$

For proofs and more details on Hermite analysis and the Ornstein-Uhlenbeck operators, the reader may consult the books of Bakry [Bak94], Janson [Jan97] or Ledoux and Talagrand [LT91].

Gaussian surface area. Given a (Borel) set $K \subseteq \mathbb{R}^n$, the Gaussian *volume* of K is defined to be simply

$$\mathrm{vol}(K) = \mathop{\mathbf{Pr}}_{x\sim\mathcal{N}^n}[x \in K] = \mathop{\mathbf{E}}_{x\sim\mathcal{N}^n}[\mathbf{1}_{x\in K}].$$

We will be especially interested in the Gaussian surface area of K, sometimes referred to as *Gaussian perimeter*, which was defined in Section 1.2. In this paper we will work exclusively with sets K satisfying

$\mathrm{vol}(\partial K) = 0$, where ∂K denotes the boundary of K. We may then make the convenient assumption that our sets K are also always closed; this is no restriction since K and $\overline{K}$ have the same boundary and hence surface area.

2.2 Learning Models We now describe the framework of agnostically learning a class $\mathcal{C}$ with respect to a fixed distribution $\mathcal{D}$ over $\mathbb{R}^n$. In this scenario there is an unknown distribution $\mathcal{D}'$ over $\mathbb{R}^n \times \{-1, 1\}$ whose marginal distribution over $\mathbb{R}^n$ is $\mathcal{D}$. Let $\mathsf{opt} \stackrel{\text{def}}{=} \inf_{f \in \mathcal{C}} \mathbf{Pr}_{(x,y)\sim\mathcal{D}'}[f(x) \neq y]$; i.e. opt is the minimum error of any function from $\mathcal{C}$ in predicting the labels y. The learner must output a hypothesis whose error is within ϵ of opt:

Definition 7. *Let $\mathcal{D}'$ be an arbitrary distribution on $\mathbb{R}^n \times \{-1, 1\}$ whose marginal over $\mathbb{R}^n$ is $\mathcal{D}$, and let $\mathcal{C}$ be a class of Boolean functions $f : \mathbb{R}^n \to \{-1, 1\}$. We say that algorithm B is an agnostic learning algorithm for $\mathcal{C}$ with respect to $\mathcal{D}$ if the following holds: for any $\mathcal{D}'$ as described above, if B is given access to a set of labeled examples (x, y) drawn from $\mathcal{D}'$, then with probability at least $1 - \delta$ algorithm B outputs a hypothesis $h : \mathbb{R}^n \to \{-1, 1\}$ such that $\mathbf{Pr}_{(x,y)\sim\mathcal{D}'}[h(x) \neq y] \leq \mathsf{opt} + \epsilon$.*

Agnostic learning is a challenging model for which, until recently, few nontrivial learning algorithms were known. Intuitively one can think of the unknown distribution $\mathcal{D}'$ over labeled examples as corresponding to an unknown function $f \in \mathcal{C}$ whose outputs are adversarially corrupted with overall probability opt.

The usual (noise-free) model of PAC learning with respect to a distribution $\mathcal{D}$ is the special case of the above definition in which we require that $\mathsf{opt} = 0$, i.e. there is an unknown target function $f \in \mathcal{C}$ such that all examples are labeled according to f.

Agnostic Learning via Hermite Concentration. Here we explain how to learn concept classes that can be approximated well by low-degree polynomials.

Definition 8. *Let $\alpha(\epsilon, n)$ be a function $\alpha : (0, 1/2) \times \mathbb{N} \to \mathbb{N}$. We say that a class of functions $\mathcal{C}$ over $\mathbb{R}^n$ has a* Hermite concentration bound *of $\alpha(\epsilon, n)$ if, for all $n \geq 1$, all $0 < \epsilon < \frac{1}{2}$, and all $f \in \mathcal{C}$ we have $\sum_{|S| \geq \alpha(\epsilon,n)} \hat{f}(S)^2 \leq \epsilon$.*

Our main tool for agnostic learning under $\mathcal{N}^n$ is the L_1 polynomial regression algorithm of Kalai et al. [KKMS05]. To agnostically learn a concept class $\mathcal{C}$, their algorithm approximately minimizes $\mathbf{E}_{(x,y)\sim D}[|p(x) - y|]$ over all multivariate polynomials p of degree d and outputs a thresholded polynomial as its hypothesis. The algorithm runs in time $n^{O(d)}$ where d is chosen according to the Hermite concentration of the concept class $\mathcal{C}$:

Theorem 9 ([KKMS05])**.** *Let $\mathcal{C}$ be a class of functions over $\mathbb{R}^n$ with Hermite concentration bound $\alpha(\epsilon, n)$. The L_1 polynomial regression algorithm is an agnostic learning algorithm for $\mathcal{C}$ with respect to $\mathcal{N}^n$. It runs in time $\mathrm{poly}(n^{\alpha(\epsilon^2/2,n)}, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ to learn to accuracy ϵ with confidence $1 - \delta$.*

PAC Learning via Hermite Concentration. The following theorem is implicit in [KKMS05]:

Theorem 10. *Let $\mathcal{C}$ be a class of ± 1-valued functions over $\mathbb{R}^n$ with Hermite concentration bound $\alpha(\epsilon, n)$. Then there exists an algorithm for learning $\mathcal{C}$ given data labeled according to f and drawn from the standard Gaussian distribution $\mathcal{N}^n$ on $\mathbb{R}^n$ that runs in time $\mathrm{poly}(n^{\alpha(\epsilon/2,n)}, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ and outputs, with probability at least $1 - \delta$, a polynomial p of degree at most $\alpha(\epsilon/2, n)$ such that $\mathbf{Pr}_{x\sim\mathcal{N}^n}[\mathrm{sgn}(p(x)) \neq f(x)] \leq \epsilon$.*

The algorithm of this theorem performs L_2 polynomial regression, i.e. it approximately minimizes $\mathbf{E}_{(x,y)\sim D}[(p(x) - y)^2]$ over all multivariate polynomials p of degree d and outputs a thresholded polynomial as its hypothesis.

To summarize, a concept class $\mathcal{C}$ can be both PAC and agnostically learned in time exponential in the Hermite concentration bounds $\alpha(\epsilon/2, n)$ and $\alpha(\epsilon^2/2, n)$ respectively.

3 Bounding Hermite Concentration in Terms of Surface Area

In this section we give our main connection between Hermite concentration and Surface Area.

Definition 11. *We define* $\mathbb{S}_\rho(f,g) \overset{def}{=} \langle f, T_\rho g\rangle = \langle T_\rho f, g\rangle$. *In the special case* $f = g$ *we write* $\mathbb{S}_\rho(f) \overset{def}{=} \langle f, T_\rho f\rangle$ *and call this the "noise stability of* f *at* ρ."

It is easy to check that the above definition is symmetric in f and g; i.e., $\mathbb{S}_\rho(f,g) = \mathbb{S}_\rho(g,f)$. Further, by combining (2) with Plancherel's identity, we have

$$\mathbb{S}_\rho(f,g) = \sum_{S\in\mathbb{N}^n} \rho^{|S|}\hat{f}(S)\hat{g}(S). \tag{3}$$

We are particularly interested in functions which are indicators of sets $K \subseteq \mathbb{R}^n$; as is usual in learning theory, we use ± 1 indicators. For notational simplicity, we identify a set with its indicator; i.e.,

$$K(x) \overset{\text{def}}{=} \begin{cases} +1 & \text{if } x \in K, \text{ the "positive region"}, \\ -1 & \text{if } x \in K^c, \text{ the "negative region"}. \end{cases}$$

In this case, we define:

Definition 12. *Given* $K \subseteq \mathbb{R}^n$*, the "noise sensitivity of* K *at* $\delta \in [0,1]$*" is*

$$\mathbb{NS}_\delta(K) \overset{def}{=} \tfrac{1}{2} - \tfrac{1}{2}\langle K, T_{1-\delta}K\rangle = \tfrac{1}{2} - \tfrac{1}{2}\mathbb{S}_{1-\delta}(K).$$

By definition of $T_{1-\delta}$, we have that

$$\begin{aligned} \mathbb{NS}_\delta(K) &= \tfrac{1}{2} - \tfrac{1}{2}\langle K, T_{1-\delta}K\rangle \\ &= \tfrac{1}{2} - \tfrac{1}{2}\mathop{\mathbf{E}}_{x,z\sim\mathcal{N}^n}[K(x)K(y)], \qquad \text{where } y \overset{\text{def}}{=} (1-\delta)\,x + \sqrt{2\delta-\delta^2}\,z \\ &= \mathop{\mathbf{Pr}}_{x,z}[K(x) \neq K(y)]; \end{aligned} \tag{4}$$

i.e., $\mathbb{NS}_\delta(K)$ is the probability that two "$(1-\delta)$-correlated" Gaussians land on opposite "sides" of K. From this interpretation, it is intuitive that, at least for small δ, the quantity $\mathbb{NS}_\delta(K)$ should be in some way comparable to the Gaussian surface area of K. The critical theorem we need in this regard was proven by Ledoux [Led94] (who mentioned it was implicitly proven by Pisier [Pis86]):

Theorem 13 (Ledoux-Pisier)**.** *Let* $K \subseteq \mathbb{R}$ *be a set with smooth*[1] *boundary, and let* $t \geq 0$. *Then*

$$\langle \mathbf{1}_K, P_t \mathbf{1}_{K^c}\rangle \leq \frac{\arccos(e^{-t})}{\sqrt{2\pi}}\Gamma(K).$$

We now manipulate Theorem 13 slightly to state it in terms of noise sensitivity. First, we replace P_t by $T_{1-\delta}$ and use the fact that $\arccos(1-\delta) \leq \frac{\pi}{2\sqrt{2}}\sqrt{\delta}$. Next, we compute easily by linearity that $\langle \mathbf{1}_K, T_{1-\delta}\mathbf{1}_{K^c}\rangle = \frac{1}{2}\mathbb{NS}_\delta(K)$. Putting these together we conclude:

Corollary 14. *Let* $K \subseteq \mathbb{R}^n$ *be a Borel set, and let* $\delta \geq 0$. *Then* $\mathbb{NS}_\delta(K) \leq \sqrt{\pi}\sqrt{\delta}\cdot\Gamma(K)$.

[1] A technical remark: We would like to apply Theorem 13 to general convex sets, which need not have smooth boundary. However the arguments in [BH97, proof of "Theorem 1.1, (b) ⇒ (a)"] straightforwardly imply that Theorem 13 extends to all Borel sets (and hence convex sets) [Led06].

Next, using (3) we have the formula $\mathbb{NS}_\delta(K) = \frac{1}{2} - \frac{1}{2}\sum_{S\in\mathbb{N}^n}(1-\delta)^{|S|}\widehat{K}(S)^2$. Using this, and $\sum_S \widehat{K}(S)^2 = 1$ (by Parseval), it is easy to check (see Proposition 16 of [KOS04]) that

$$\sum_{|S|\geq 1/\delta} \widehat{K}(S)^2 \leq \frac{2}{1-1/e^2}\mathbb{NS}_\delta(f).$$

Combining this with Corollary 14 we obtain

$$\sum_{|S|\geq 1/\delta} \widehat{K}(S)^2 \leq 5\cdot\sqrt{\delta}\cdot\Gamma(K),$$

and hence we conclude our main Hermite concentration bound based on surface area:

Theorem 15. *Let $K \subseteq \mathbb{R}^n$ be a Borel set. Then the ± 1 indicator function of K has Hermite concentration bound $\alpha(\epsilon, n) = O(\Gamma(K)^2/\epsilon^2)$.*

4 Gaussian Surface Area Calculations and New Learning Results

Theorems 9, 10 and 15 reduce the problem of PAC and agnostically learning a concept class under the standard Gaussian distribution to the problem of bounding the surface area of the corresponding sets. The specific surface area upper bounds stated in this section for different classes of sets yield a wealth of efficient learning results for the corresponding function classes.

Up through Section 4.4 we consider only the standard spherical Gaussian distribution. In Section 4.5 we show how our learning results for the standard Gaussian distribution extend to arbitrary Gaussian distributions, and state our most general learning results.

We begin by stating a few basic facts about perimeter and recalling the classical example of halfspaces.

4.1 Basic Facts and Examples

Convex sets not containing the origin. In order to upper bound the Gaussian surface area of a convex set, we can always assume it contains the origin, via the following observation (see [Naz03b]):

Fact 16. *Suppose $K \subseteq \mathbb{R}^n$ is a convex set not containing the origin. Then it possible to translate K in such a way that (a) the origin is on the boundary of K, and (b) each point on the boundary of K (in fact, each point in K) moves closer to the origin. Since $\varphi_n(y)$ is a decreasing function of $\|y\|$, this translation only increases the surface area of K (see formula* (1)*).*

Intersections, unions, etc.

Fact 17. *Given sets K_1, K_2 we have $\Gamma(K_1 \cap K_2), \Gamma(K_1 \cup K_2) \leq \Gamma(K_1) + \Gamma(K_2)$.*

This follows from the simple observation that both $\partial(K_1 \cap K_2)$ and $\partial(K_1 \cup K_2)$ are subsets of $\partial K_1 \cup \partial K_2$. More generally, given $K_1, \ldots, K_t$, if $K(x) = f(K_1, \ldots, K_t)$ for any Boolean $f : \{-1,1\}^t \to \{-1,1\}$, then $\Gamma(K) \leq \sum_{i=1}^t \Gamma(K_i)$.

Halfspaces. This is the main classical example. Let $K \subseteq \mathbb{R}^n$ be a halfspace whose boundary is at distance t from the origin. By rotational symmetry of the Gaussian distribution, we may assume that K is the halfspace whose boundary ∂K is the plane $x_1 = t$. This reduces the calculation to a one-dimensional problem, and we immediately obtain $\Gamma(K) = \varphi(t)$. In particular, $\Gamma(K) \leq 1/\sqrt{2\pi} \leq O(1)$ for every halfspace K. The well-known "Gaussian isoperimetric inequality" [Bor75, ST78] (see also [Bob97]) states that among all sets K with $\mathrm{vol}(K)$ fixed, halfspaces *minimize* $\Gamma(K)$.

Applying Theorem 15 and Theorem 9 with the above bound on the surface area of a halfspace, we immediately obtain one of the main results of Kalai et al. [KKMS05], namely that a single halfspace can be agnostically learned with respect to $\mathcal{N}^n$ in time $n^{O(1/\epsilon^4)}$.

4.2 General Convex Sets. Ball gave the following fundamental bound on the surface area of convex sets, solving the "reverse Gaussian isoperimetric inequality":

Theorem 18. *[Bal93] The Gaussian surface area of any convex set in* $\mathbb{R}^n$ *is at most* $4n^{1/4}$.

By applying the above bound with Theorem 15 and Theorem 9 we obtain our main result for learning arbitrary convex sets:

Corollary 19. *The clas of all convex sets in* $\mathbb{R}^n$ *is PAC learnable in time* $n^{O(\sqrt{n})/\epsilon^2}$ *and agnostically learnable in time* $n^{O(\sqrt{n})/\epsilon^4}$ *under* $\mathcal{N}^n$*. The same bound holds for learning any union of* $O(1)$ *many convex sets.*

As we describe in Section 5, Nazarov [Naz03b] later showed that the bound in Theorem 18 is tight (up to a constant factor) by considering the intersection of roughly $\exp(\sqrt{n})$ randomly chosen halfspaces with boundary at distance $n^{1/4}$ from the origin.

4.3 Intersections of k halfspaces. In addition to showing that Ball's estimate is tight, Nazarov also gave a different proof of Ball's upper bound result (with a better constant), and in doing so he proved an inequality that is useful for bounding the Gaussian surface area of convex sets.

To state this bound we introduce some notation from [Naz03b]. Let $K \subseteq \mathbb{R}^n$ be a convex set containing the origin, and let $y \in \partial K$. We write ν_y for the unit normal vector to ∂K at y (which is well-defined except on a set of $(n-2)$-dimensional measure 0) We also write $\alpha(y)$ for $\cos(y \cdot \nu_y)$, and $h(y)$ for $\|y\|\alpha(y)$; in other words, $h(y)$ is the distance from the origin of the tangent (to K) hyperplane containing y. Nazarov's bound is

$$\int_{\partial K} \left(\frac{1}{h(y)+1}\right) \cdot \varphi_n(y)\, d\sigma(y) \leq 1 - \mathrm{vol}(K) \leq 1. \tag{5}$$

Recalling that $\Gamma(K) = \int_{\partial K} \varphi_n(y)\, d\sigma(y)$, for convex sets K, this bound implies that there is little contribution to $\Gamma(K)$ from points y where the tangent hyperplane is near to the origin.

This formula is useful for bounding the Gaussian surface area of intersections of halfspaces. In particular, the following bound on the surface area of the intersection of k halfspaces and proof was communicated to us by Nazarov [Naz03a]:

Theorem 20. *Let* $K \subseteq \mathbb{R}^n$ *be an intersection of up to* k *halfspaces. Then* $\Gamma(K) \leq \sqrt{2\ln k} + 2 \leq O(\sqrt{\log k})$.

To prove this, one first observes that K can be assumed to contain the origin. Then one splits up $\Gamma(K) = \int_{\partial K} \varphi_n(y)\, d\sigma(y)$ into the contribution from those y where $h(y) > \sqrt{2\ln k}$ and those y where $h(y) \leq \sqrt{2\ln k}$. The former parts contribute at most $k \cdot \varphi(\sqrt{2\ln k}) \leq 1$. The latter parts contribute at most $\sqrt{2\ln k} + 1$, using (5). In particular, Theorem 20 implies that any box or parallelopiped in $\mathbb{R}^n$, in any orientation, has Gaussian surface area at most $O(\sqrt{\log n})$. Ball made a similar observation earlier for boxes.

Applying our machinery relating learning to surface area, we obtain

Corollary 21. *Any intersection of up to* k *halfspaces in* $\mathbb{R}^n$ *is PAC learnable in time* $n^{O(\log k)/\epsilon^2}$ *and agnostically learnable in time* $n^{O(\log k)/\epsilon^4}$ *under* $\mathcal{N}^n$.

As noted in the introduction, compared with Vempala's $(n/\epsilon)^{O(k)}$-time PAC learning algorithm (with respect to nearly-uniform distributions on the sphere)[2], his dependence on ϵ is better if $\log(1/\epsilon) \gg \log k$, but otherwise our algorithm has a much better dependence on n and works in the agnostic setting.

We can also use Nazarov's inequality to bound the Gaussian surface area of certain cones:

[2]Vempala [Vem97] claims a running time of $\mathrm{poly}(n)k^k(\frac{1}{\epsilon})^k$ for the algorithm but this was amended to $(n/\epsilon)^{O(k)}$ in [Vem04].

Theorem 22. *Let K be a cone with apex at the origin (i.e. an intersection of arbitrarily many halfspaces all of whose boundaries contain the origin). Then K has Gaussian surface area at most 1.*

This follows immediately from Equation (5) since if K is a cone as described then we have $h(y) = 0$ for every $y \in \partial K$. As a corollary we have that cones with an apex at the origin are PAC and agnostically learnable with respect to $\mathbb{N}^n$ in time $n^{O(1/\epsilon^2)}$ and $n^{O(1/\epsilon^4)}$, respectively.

4.4 Balls. Let B_r^n denote the ball of radius r in $\mathbb{R}^n$, centered at the origin. Ball [Bal93] gave the formula $\Gamma(B_r^n) = \frac{r^{n-1}}{2^{n/2-1}\Gamma(n/2)e^{r^2/2}}$. He noted that this is maximized at $r = \sqrt{n-1}$ where the surface area is asymptotic to $1/\sqrt{\pi}$.

It is tempting to believe that the origin-centered ball has maximum surface area for any radius r, but this is not always true; consider, for example, a ball of radius $r(n)$, where $r(n)$ grows very rapidly relative to n. If such a ball is centered at the origin, its surface area will approach 0 very rapidly (exponentially fast in $r(n)^2$). But, if the ball is displaced so that the origin lies on its surface, then the Gaussian surface area will be nearly that of an origin-centered halfspace, which is an absolute constant $1/\sqrt{2\pi}$ independent of n.

Since Ball's argument uses the radial symmetry of the Gaussian and explicitly computes the integral of the Gaussian density over the surface of the ball, it is not clear how to extend the argument to non-origin centered balls. In Appendix B we give an alternate proof of Ball's result for origin-centered balls that does not rely on computing surface integrals. Instead, we maximize a corresponding probability density function; this approach allows us to show that any ball, origin-centered or not, has surface area at most a constant:

Theorem 23. *The Gaussian surface area of any ball in $\mathbb{R}^n$ is at most* 1.

Applying Theorem 15 and Theorem 9 we have the following corollary:

Corollary 24. *The class of balls in $\mathbb{R}^n$ is agnostically learnable in time $n^{O(1/\epsilon^4)}$ with respect to $\mathcal{N}^n$.*

Again we remark that the same time bound holds even for unions of a constant number of balls.

4.5 Learning under Arbitrary Gaussian Distributions. We can show that (almost all of) our learning results extend to arbitrary Gaussian distributions. The arguments of this section, together with Theorems 15, 10, and 9, give Theorem 25, our most general learning result:

Theorem 25. *Let $\mathcal{C}$ be a class of Borel sets in $\mathbb{R}^n$, each of which has Gaussian surface area at most s. Assume that $\mathcal{C}$ is closed with respect to affine transformations. Then $\mathcal{C}$ is PAC learnable to accuracy ϵ with respect to any Gaussian distribution on $\mathbb{R}^n$ (with nonsingular covariance matrix[3]) in time $n^{O(s^2/\epsilon^2)}$ and agnostically learnable in time $n^{O(s^2/\epsilon^4)}$.*

Due to space considerations we defer this section to Appendix C.

5 Lower Bounds for Learning under Gaussian Distributions

In this section we prove a sample complexity lower bound for learning intersections of 2^ℓ halfspaces under the standard n-dimensional Gaussian distribution $\mathcal{N}^n$ (recall that by Theorem 20, any such intersection of 2^ℓ halfspaces has Gaussian surface area $O(\sqrt{\ell})$).

Theorem 26. *There is a fixed constant $\epsilon_0 > 0$ such that the following holds:*

Let $\ell = \ell(n)$ be such that $\log n \leq \ell \leq n^{1/2}\epsilon_0^2$. Let $\mathcal{H}_\ell$ be the class of all intersections of $2^{O(\ell)}$ halfspaces over $\mathbb{R}^n$. Let A be any algorithm which learns $\mathcal{H}_\ell$ to confidence $\delta = 1/2$ and accuracy ϵ_0 with respect to $\mathcal{N}^n$. Then A must use $2^{\Omega(\ell)}$ examples. This lower bound holds even for algorithms which may make black-box queries to the target function f and suffer no noise.

[3]As discusssed in Section 4.5, if the class $\mathcal{C}$ is closed under intersections with lower-dimensional subspaces then we can drop the requirement that the covariance matrix be nonsingular.

Discussion. This theorem implies that for a wide range of parameters, our algorithm of Corollary 21, which can learn intersections of 2^ℓ halfspaces to constant accuracy ϵ in time $n^{O(\ell/\epsilon^2)}$, is essentially optimal in its dependence on the number of halfspaces. The theorem similarly implies that our positive results for learning general convex sets and learning sets with bounded Gaussian surface area are also essentially optimal. We remind the reader that while the lower bound holds even for learning under the standard Gaussian distribution with membership queries, our positive results for these classes all hold for learning from random examples generated from any Gaussian distribution, without using queries.

We briefly sketch the approach. Given two functions $f, g : \mathbb{R}^n \to \{0,1\}$ we write $d(f,g)$ to denote $\mathbf{Pr}_{X \sim \mathcal{N}^n}[f(X) \neq g(X)]$; we extend the notion to subsets A, B of $\mathbb{R}^n$, writing $d(A,B) = d(\mathbf{1}_A, \mathbf{1}_B)$. We prove Theorem 26 by establishing the following:

Theorem 27. *Let ℓ, ϵ_0 be as in Theorem 26. There exists a set $\mathcal{C}_\ell = \{f_1, \ldots, f_M\}$ of $M = 2^{2^{\Omega(\ell)}}$ many functions $f_i \in \mathcal{H}_\ell$ such that for any $1 \leq i < j \leq M$, we have $d(f_i, f_j) \geq 2\epsilon_0$.*

By results of Benedek and Itai [BI88], this implies that any algorithm (even allowing membership queries) for learning the class $\mathcal{C}_\ell$ under distribution $\mathcal{N}^n$ with confidence parameter $\delta = 1/2$ and accuracy parameter ϵ_0 must have sample complexity at least $\log M = 2^{\Omega(\ell)}$. To prove Theorem 26 it thus suffices to prove Theorem 27.

We prove Theorem 27 using the probabilistic method. The idea is to consider an intersection of N halfspaces (we specify N later) in which each halfspace is chosen uniformly at random from all halfspaces tangent to an origin-centered ball of a certain radius, chosen so that the resulting convex body is likely to have Gaussian volume bounded away from 0 and 1 by a constant.[4] Using the "method of bounded differences" we show that that two convex bodies that are independently generated in this way are extremely likely to be far from each other; together with a union bound, this gives Theorem 27. The proof is given in Appendix D.

6 Acknowledgements

We are grateful to Fedja Nazarov for proving Theorem 20 for us. We also thank Michel Ledoux for a helpful discussion about the conditions required for Theorem 13.

References

[Bak94] D. Bakry. *L'hypercontractivité et son utilisation en théorie des semigroupes*, pages 1–114. Springer, 1994.

[Bal93] K. Ball. The Reverse Isoperimetric Problem for Gaussian Measure. *Discrete and Computational Geometry*, 10:411–420, 1993.

[Bau90a] E. Baum. On learning a union of halfspaces. *Journal of Complexity*, 6(1):67–101, 1990.

[Bau90b] E. Baum. The Perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, 2:248–260, 1990.

[BB05] N. H. Bshouty and L. Burroughs. Maximizing agreements with one-sided error with applications to heuristic learning. *Machine Learning*, 59(1-2):99–123, 2005.

[4] This construction of a convex body formed by randomly intersecting halfspaces in this fashion was first proposed by Nazarov [Naz03b] who showed that a convex body formed in this way may have Gaussian surface area $\Omega(n^{1/4})$ (see [Ben03] for a nice exposition of the proof). It is no coincidence that we follow his construction; since our algorithm gives an $n^{O(\Gamma(A)^2/\epsilon^2)}$ upper bound on the time required to learn a convex body A in terms of its Gaussian surface area, in order to prove a lower bound it is *necessary* to consider convex bodies with large surface area.

[BDEL03] S. Ben-David, N. Eiron, and P. Long. On the difficulty of approximately maximizing agreements. *JCSS: Journal of Computer and System Sciences*, 66, 2003.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[Ben03] V. Bentkus. On the dependence of the Berry-Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113:385–402, 2003.

[BG99] S. Bobkov and F. Götze. Discrete isoperimetric and Poincaré-type inequalities. *Prob. Theory and Related Fields*, 114:245–277, 1999.

[BH97] S. Bobkov and C. Houdre. *Some Connections between Isoperimetric and Sobolev-type Inequalities*. American Mathematical Society, Providence, 1997.

[BI88] G. Benedek and A. Itai. Learnability by fixed distributions. In *First Workshop on Computational Learning Theory*, pages 201–210, 1988.

[BK97] A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces under a uniform distribution. *Journal of Computer and System Sciences*, 54(2):371–380, 1997.

[BKS99] I. Benjamini, G. Kalai, and O. Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90:5–43, 1999.

[Bob97] S. Bobkov. An isoperimetric inequality on the discrete cube and an elementary proof of the isoperimetric inequality in Gauss space. *The Annals of Probability*, 25(1):206–214, 1997.

[Bor75] C. Borell. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30:207–216, 1975.

[BT96] N. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[EHKV89] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[Fel68] W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 1968.

[Gro75] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[Jan97] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, UK, 1997.

[KKMS05] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

[KOS04] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004. Preliminary version in *Proc. of FOCS'02*.

[KP98] S. Kwek and L. Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1/2):53–75, 1998.

[KS04] A. Klivans and R. Servedio. Learning intersections of halfspaces with a margin. In *Proceedings of the 17th Annual Conference on Learning Theory,*, pages 348–362, 2004.

[KS06] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *FOCS*, pages 553–562. IEEE Computer Society, 2006.

[KV94] M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.

[Led94] M. Ledoux. Semigroup proofs of the isoperimetric inequality in Euclidean and Gauss space. *Bull. Sci. Math.*, 118:485–510, 1994.

[Led06] M. Ledoux. Personal communication, 2006.

[Lon94] P. Long. Halfspace learning, linear programming, and nonmalicious distributions. *Information Processing Letters*, 51:245–250, 1994.

[Lon95] P. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. London Mathematical Society Lecture Notes, 1989.

[Naz03a] F. Nazarov. Personal communication, 2003.

[Naz03b] F. Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric aspects of functional analysis (2001-2002)*, pages 169–187. Lecture Notes in Math., Vol. 1807, Springer, 2003.

[Pat49] P.B. Patnaik. The non-central χ^2 and F-distributions and their applications. *Biometrika*, 36:202–232, 1949.

[Per04] Y. Peres. Noise stability of weighted majority, 2004.

[Pis86] G. Pisier. *Probabilistic methods in the geometry of Banach spaces*, pages 167–241. Springer, 1986.

[SBH01] F. Götze S. Bobkov and C. Houdré. On Gaussian and Bernoulli covariance representations. *Bernoulli*, 7(3):439–451, 2001.

[ST78] V. Sudakov and B. Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *J. Soviet Math.*, 9:9–18, 1978. Translated from Zap. Nauchn. Sem. Leningrad. Otdel. Math. Inst. Steklova. 41 (1974), 14–21.

[Tal93] M. Talagrand. Isoperimetry, logarithmic Sobolev inequalities on discrete cube and Margulis' graph connectivity theorem. *GAFA*, 3(3):298–314, 1993.

[Tal96] M. Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[TZ00] J.-P. Tillich and G. Zémor. Discrete isoperimetric inequalities and the probability of a decoding error. *Combinatorics, Probability & Computing*, 9(5), 2000.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vem97] S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pages 508–513, 1997.

[Vem04] S. Vempala. *The Random Projection Method*. American Mathematical Society, DIMACS, 2004.

A Review of Hermite Analysis

We will work within $L^2(\mathbb{R}^n, \mathcal{N}^n)$, the vector space of all functions $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbf{E}[f^2] < \infty$. This is an inner product space under the inner product

$$\langle f, g \rangle = \mathop{\mathbf{E}}_{x \sim \mathcal{N}^n}[f(x)g(x)].$$

This inner product space has a complete orthonormal basis given by the *Hermite polynomials*. In the case $n = 1$, this basis is the sequence of polynomials

$$h_0(x) = 1, \quad h_1(x) = x, \quad h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \quad h_3(x) = \frac{x^3 - 3x}{\sqrt{6}}, \quad \ldots$$

There are several equivalent ways to define this sequence:

$$\exp(\lambda x - \lambda^2/2) =: \sum_{j=0}^{\infty} \frac{\lambda^d}{\sqrt{d!}} h_j(x);$$

$$h_j(x) = \frac{(-1)^d}{\sqrt{d!}\,\varphi(x)} \cdot \frac{d^j}{dx^j}\varphi(x);$$

$$h_j(x) = \frac{\sqrt{j!}}{(j-0)!0!2^0} x^j - \frac{\sqrt{j!}}{(j-2)!1!2^1} x^{j-2} + \frac{\sqrt{j!}}{(j-4)!2!2^2} x^{j-4} - \frac{\sqrt{j!}}{(j-6)!3!2^3} x^{j-6} + \cdots$$

For general n, the basis for $L^2(\mathbb{R}^n, \mathcal{N}^n)$ is formed by all products of these polynomials, one for each coordinate. I.e., for each n-tuple $S \in \mathbb{N}^n$ we define the n-variate Hermite polynomial $H_S : \mathbb{R}^n \to \mathbb{R}$ by

$$H_S(x) = \prod_{i=1}^{n} h_{S_i}(x_i);$$

then the collection $(H_S)_{S \in \mathbb{N}^n}$ is a complete orthonormal basis for the inner product space. By orthonormal we mean that

$$\langle H_S, H_T \rangle = \begin{cases} 1 & \text{if } S = T, \\ 0 & \text{if } S \neq T. \end{cases}$$

By complete, we mean that every function $f \in L^2$ can be uniquely expressed as

$$f(x) = \sum_{S \in \mathbb{N}^n} c_S H_S(x),$$

where the coefficients c_S are real numbers and the infinite sum converges in the sense that

$$\lim_{d \to \infty} \mathbf{E}\left[\left(f(x) - \sum_{|S| \le d} c_S H_S(x)\right)^2\right] = 0;$$

here we have used the notation

$$|S| = \sum_{i=1}^{n} S_i,$$

which is also the total degree of $H_S(x)$ as a polynomial.

Given f, instead of c_S, we will write $\hat{f}(S)$, and call this the *S Hermite coefficient of f*. By orthonormality of the basis $(H_S)_{S\in\mathbb{N}^n}$, we have the following:

$$\hat{f}(S) = \langle f, H_S\rangle = \mathbf{E}[f(x)H_S(x)];$$

$$\|f\|_2^2 \stackrel{\text{def}}{=} \langle f, f\rangle = \sum_{S\in\mathbb{N}^n} \hat{f}(S)^2 \qquad \text{("Parseval's identity")};$$

$$\langle f, g\rangle = \sum_{S\in\mathbb{N}^n} \hat{f}(S)\hat{g}(S) \qquad \text{("Plancherel's identity")}.$$

In particular, if $f : \mathbb{R}^n \to \{-1, 1\}$, then $\sum \hat{f}(S)^2 = 1$ (when no range for a sum over S is specified, we assume $\mathbb{N}^n$).

B Bounding the Gaussian surface area of an arbitrary ball

Our approach to bounding the Gaussian surface area of a ball is by analyzing an appropriate probability density function.

Recall the *chi-square distribution with k degrees of freedom*:

$$\chi_k^2 = \sum_{i=1}^{k} X_i^2$$

where each X_i is a random variable distributed according to $\mathcal{N}(0, 1)$. Notice that for an origin-centered ball K of radius r, the Gaussian volume of K is equal to

$$\mathbf{Pr}[\chi_n^2 \leq r^2].$$

Since the δ-neighborhood of a ball of radius r is a ball of radius $r + \delta$, by the definition of Gaussian surface area we have that the Gaussian surface area of a ball of radius r is equal to

$$\lim_{\delta\to 0} \frac{\mathbf{Pr}[\chi_n^2 \leq (r+\delta)^2] - \mathbf{Pr}[\chi_n^2 \leq r^2]}{\delta}.$$

Consequently, differentiating the cdf and applying the chain rule, we have that the Gaussian surface area of an origin-centered ball is equal to $2r \cdot f_n(r^2)$ where f_n is the pdf of χ_n^2. It is well known [Fel68] that the pdf of χ_n^2 is given by

$$f_n(x) = \frac{x^{n/2-1}}{\Gamma(n/2)2^{n/2}e^{x/2}}.$$

It is straightforward to verify that $2r \cdot f_n(r^2)$ agrees with Ball's formula for the surface area of an origin-centered ball of radius r.

To bound the surface area of non-origin-centered balls, we will need to consider the *non-central chi-square distribution*:

Definition 28. *We say that $Q_{(n,\lambda)}$ is a non-central chi-square distribution with n degrees of freedom and non-centrality parameter λ if $Q_{(n,\lambda)} = \sum_{i=1}^{n} Y_i^2$ where each Y_i is an independent $\mathcal{N}(a_i, 1)$ Gaussian and $\lambda = \sum_{i=1}^{n} a_i^2$.*

To compute the surface area of a ball, we can first assume without loss of generality (due to the rotational symmetry of the Gaussian) that the ball is centered on the x-axis. Next we observe that the Gaussian volume of a ball of radius r centered at distance d from the origin is given by

$$\mathbf{Pr}[Q_{(n,d^2)} \leq r^2].$$

Let $g_{(n,d^2)}$ denote the pdf of the random variable $Q_{(n,d^2)}$. Although there is no simple closed form for $g_{(n,d^2)}$, Patnaik [Pat49] has observed that

$$g_{(n,\lambda)} = \sum_{j=0}^{\infty} \frac{\frac{1}{2}\lambda^j}{j!} \exp(-\lambda/2) f_{n+2j} \tag{6}$$

where each f_{n+2j} is the pdf of χ^2_{n+2j}. This means that the non-central chi-square distribution is a convex combination of standard chi-square distributions, since the weights in the above formula are exactly the probabilities of a Poisson distribution with expected value $\lambda/2$.

We can now bound the surface area of a non-origin-centered ball as follows. From the above discussion it suffices to show that for any r the quantity $2r \cdot g_{(n,d^2)}(r^2)$ is at most 1. From Equation (6), we see that the function $2r \cdot g_{(n,d^2)}(r^2)$ is a convex combination of functions of the form $2r \cdot f_j(r^2)$ across different values of j. It is not difficult to verify that that for all j, the value of $2r \cdot f_j(r^2)$ is always at most 1 (recall that for a given j the maximum is at $r = \sqrt{j-1}$). Thus, $2r \cdot g_{(n,d^2)}(r^2)$ is at most 1 as well.

C Learning with Respect to Arbitrary Gaussians

Here we sketch how (almost all of) our learning results can be extended to arbitrary Gaussian distributions.

Recall that an arbitrary Gaussian distribution $\mathcal{D}$ over $\mathbb{R}^n$ can be generated by first drawing $x \sim \mathcal{N}^n$ and then outputting $\mu + Bx$, where μ is a fixed vector (the mean of $\mathcal{D}$) and B is a fixed square matrix, possibly not of full rank (the matrix square-root of the covariance matrix). Let T denote the affine transformation $x \mapsto \mu + Bx$.

Let us assume for a moment that the matrix B has full rank so that T is invertible. Given a set $K \subseteq \mathbb{R}^n$, let K' denote the set $T^{-1}K$. Now if there is a polynomial p' over $\mathbb{R}^n$ with degree at most d satisfying

$$\mathop{\mathbf{E}}_{x \sim \mathcal{N}^n}[(p'(x) - K'(x))^2] \leq \epsilon$$

(again we identify K' with its ± 1 indicator function), then we immediately have

$$\mathop{\mathbf{E}}_{y \sim \mathcal{D}}[((p' \circ T^{-1})(y) - K(y))^2] \leq \epsilon$$

But T^{-1} is an affine transformation, so $p = p' \circ T^{-1}$ also has degree at most d. In other words, the existence of a good approximating polynomial p' for K' implies the existence of a good approximating polynomial p for K. It follows that our learning algorithms in Section 2.2 will work at least as well when run on K under $\mathcal{D}$ as they do when run on K' under $\mathcal{N}^n$. (Note that we do not have to assume the learning algorithm knows the parameters of the Gaussian distribution $\mathcal{D}$; it always runs the same polynomial regression algorithm.)

In the case when B is not invertible, the distribution $\mathcal{D}$ is equivalent to a nonsingular Gaussian distribution $\mathcal{E}$ supported on an affine subspace H. It is easy to see that the above argument lets us derive approximating polynomials for K under $\mathcal{D}$ that are at least as good as approximating polynomials for $K \cap H$ under

$\mathcal{E}$ (which in turn are at least as good as approximating polynomials for some affine transformation of $K \cap H$ under $\mathcal{N}^m$ for $m = \dim(H)$).

We now observe that many classes $\mathcal{C}$ of subsets of $\mathbb{R}^n$ that we have considered for learning are closed under taking affine transformations and intersections with affine subspaces. For instance, the class of convex sets has this property, as does the class of intersections of k halfspaces. Thus our learning results for these classes immediately extend to all Gaussian distributions. Cones are closed under linear transformations and intersections with subspaces, and thus our learning results for cones extend to all Gaussian distributions so long as the cones have their apex at the Gaussian's mean.

Unfortunately, the class of balls is not closed under linear transformations. We strongly believe that all ellipsoids in $\mathbb{R}^n$ have Gaussian surface area $O(1)$; however we have not yet proved this. If this holds then our learning results for balls would also generalize to all Gaussian distributions.

D Proof of Theorem 27

Let us state Theorem 27 a bit more precisely, revealing the bound on ϵ_0:

Theorem 27. *Fix ϵ_0 to be any constant between 0 and $1/44000$. Let ℓ be as in Theorem 26. There exists a set $\mathcal{C}_\ell = \{f_1, \ldots, f_M\}$ of $M = 2^{2^{\Omega(\ell)}}$ many functions $f_i \in \mathcal{H}_\ell$ such that for any $1 \le i < j \le M$, we have*

$$d(f_i, f_j) \ge 2\epsilon_0. \tag{27}$$

Let $Z^1, Z^2, \ldots$ be independent uniformly distributed random vectors drawn from the unit ball $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. Let $\rho = \ell^{1/2}/\epsilon_0$; observe that by the assumptions on ℓ, ϵ we have $\rho \le n^{1/4}$. Let $A(Z^1, \ldots, Z^N)$ denote the intersection of N halfspaces

$$A(Z^1, \ldots, Z^n) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : x \cdot Z^i \le \rho \text{ for all } i = 1, \ldots, N\}$$

(we will specify N soon). Theorem 27 is proved by showing that if $\{Z^{i,t}\}_{1 \le i \le M, 1 \le t \le N}$ are MN independent uniform random unit vectors as described above, then with nonzero probability, for every $1 \le i < j \le M$ the functions f_i and f_j satisfy (27), where $f_i(x)$ is defined to be the indicator function of $A(Z^{i,1}, \ldots, Z^{i,N})$. We do this by showing that for each pair (i, j) the functions f_i, f_j satisfy (27) with probability at least $1 - 1/M^2$. Since there are fewer than M^2 distinct pairs, a union bound then gives Theorem 27.

So let f_1 be the indicator function of $A(Z^{1,1}, \ldots, Z^{1,N})$ and f_2 be the indicator function of $A(Z^{2,1}, \ldots, Z^{2,N})$ for random $Z^{1,1}, \ldots, Z^{2,N}$ as described above. The key to showing that f_1 and f_2 are w.v.h.p. at least $2\epsilon_0$-far apart is the following lemma showing that the *expected* distance between f_1 and f_2 is large (the expectation is taken over the random choice of $Z^{1,1}, \ldots, Z^{2,N}$):

Lemma 29. $\mathbf{E}[d(f_1, f_2)] \ge \frac{1}{11000}$.

We will prove this lemma later. Now we show how this lower bound on expectation may be combined with the "method of bounded differences" to show that $d(f_1, f_2) < 2\epsilon_0$ holds with probability at most $1/M^2$. Recall McDiarmid's inequality:

McDiarmid bound [McD89]: *Let $X_1, \ldots X_m$ be independent random variables taking values in a set Ω. Let $F \colon \Omega^m \to \mathbb{R}$ be such that for all $i \in [m]$ we have*

$$|F(x_1, \ldots, x_m) - F(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_m)| \le c_i$$

for all $x_1, \ldots, x_m$ and x'_i in Ω. Let $\mu = \mathbf{E}[F(X_1, \ldots, X_m)]$. Then for all $\tau > 0$,

$$\mathbf{Pr}\left[F(X_1, \ldots, X_m) < \mu - \tau\right] < \exp\left(-\frac{\tau^2}{\sum_{i=1}^m c_i^2}\right).$$

We let the $2N$ independent random uniform vectors $Z^{1,1},\ldots,Z^{2,N}$ play the role of $X_1,\ldots,X_m$ in McDiarmid's bound, and we let the function $d(f_1,f_2)$ play the role of $F(X_1,\ldots,X_m)$. Given any fixed setting of $Z^{1,1},\ldots,Z^{2,N}$, the change in magnitude in $d(f_1,f_2)$ that results from replacing some $Z^{i,t}$ by any other unit vector $Z' \in S^{n-1}$ is at most

$$\begin{aligned}
\mathbf{Pr}_{X\sim\mathcal{N}^n}[X\cdot u^1 \geq \rho] + \mathbf{Pr}_{X\sim\mathcal{N}^n}[X\cdot u^2 \geq \rho] &= 2\,\mathbf{Pr}_{X_1\sim N(0,1)}[X_1 \geq \rho] && (7)\\
&\leq 2\varphi(\rho)/\rho = \sqrt{2/\pi}\cdot\rho^{-1}\cdot e^{-\rho^2/2} && (8)
\end{aligned}$$

In (7) the vectors u^1 and u^2 are arbitrary fixed unit vectors, and the equality holds by the spherical symmetry of $\mathcal{N}^n$. The bound (8) follows from the standard bound $1-\Phi(t)\leq\varphi(t)/t$, which holds for $t>0$ where $\Phi(t)$ is the c.d.f. and $\varphi(t)=\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$ is the p.d.f. of $N(0,1)$. We thus may take each c_i in McDiarmid's bound to be the bound (8) above. The mean $\mathbf{E}[d(f_1,f_2)]$ is at least $\frac{1}{11000}$ by Lemma 29. As we show in (16) below, we have $N\leq 12(n^{1/2}/\rho)e^{\rho^2/2}$. Taking $\tau=\frac{1}{22000}$ in McDiarmid's bound, we thus have

$$\begin{aligned}
\mathbf{Pr}[d(f_1,f_2)<2\epsilon_0] &\leq \mathbf{Pr}[d(f_1,f_2)<\frac{1}{22000}]\\
&< \exp\left(\frac{-(1/22000)^2}{24(n^{1/2}/\rho)e^{\rho^2/2}\cdot(\sqrt{2/\pi}\cdot\rho^{-1}\cdot e^{-\rho^2/2})^2}\right)\\
&= \exp\left(-\Theta(1)\cdot(\rho^3/n^{1/2})\cdot e^{\rho^2/2}\right) && (9)
\end{aligned}$$

We define M to be such that $1/M^2 \stackrel{\text{def}}{=} (9)$. Since $\rho^2=\ell/\epsilon_0^2 \gg 2\log n$ by our assumptions on ℓ and ρ, we have that $(9)\leq\exp(-2^{\Omega(\rho^2)})$, and hence $M=2^{2^{\Omega(\ell)}}$. It remains only to prove Lemma 29.

D.1 Proof of Lemma 29 First some notation. We write A^1 to denote $A(Z^{1,1},\ldots,Z^{1,N})$ and A^2 to denote $A(Z^{2,1},\ldots,Z^{2,N})$. Recall that $\varphi_n(x)$ denotes the density function of the standard n-dimensional Gaussian distribution. Let us write "$Z\sim\mathcal{S}$" to indicate that Z is a random unit vector distributed uniformly over the unit ball S^{n-1} in $\mathbb{R}^n$. Given $x\in\mathbb{R}^n$, let us write $b(x)$ to denote

$$b(x) \stackrel{\text{def}}{=} \mathbf{Pr}_{Z\sim\mathcal{S}}[x\cdot Z\leq\rho].$$

Let a_{n-1} denote the surface area of the unit sphere S^{n-1}. It is well known that $a_{n-2}/a_{n-1}=\Theta(n^{1/2})$; for conciseness we write r_n to denote a_{n-2}/a_{n-1}. For $n\geq 3$, for any fixed unit vector $u\in S^{n-1}$, we have

$$\mathbf{Pr}_{Z\sim\mathcal{S}}[\alpha\leq u\cdot Z\leq\beta] = r_n\int_\alpha^\beta\left(\sqrt{1-z^2}\right)^{n-3}dz.$$

Let us write $\mathrm{cap}(t)$ to denote the fractional surface area of the spherical cap $S^{n-1}\cap\{x : x_1\geq t\}$:

$$\mathrm{cap}(t) \stackrel{\text{def}}{=} \mathbf{Pr}_{Z\sim\mathcal{S}}[Z_1\geq t] = r_n\int_t^1\left(\sqrt{1-z^2}\right)^{n-3}dz. \qquad (10)$$

Consequently for all $0\neq x\in\mathbb{R}^n$, we have

$$b(x) = 1-\mathbf{Pr}_{Z\sim\mathcal{S}}\left[\frac{x}{\|x\|}\cdot Z\geq\frac{\rho}{\|x\|}\right] = 1-\mathrm{cap}(\rho/\|x\|) = 1-r_n\int_{\frac{\rho}{\|x\|}}^1\left(\sqrt{1-z^2}\right)^{n-3}dz. \qquad (11)$$

Now we turn to the proof of Lemma 29. We have the following (all expectations are taken over the random choice of $Z^{1,1}, \ldots, Z^{2,N}$):

$$\begin{aligned}
\mathbf{E}[\mathbf{Pr}_{X \sim \mathcal{N}^n}[f_1(X) \neq f_2(X)]] &= 2\,\mathbf{E}[\mathbf{Pr}_{X \sim N^n}[X \in (A^1 \setminus A^2)] \\
&= \mathbf{E}\left[\int_{x \in \mathbb{R}^n} \mathbf{1}_{x \in (A^1 \setminus A^2)} \varphi_n(x) dx\right] \\
&\overset{\text{(Fubini)}}{=} \int_{x \in \mathbb{R}^n} \mathbf{E}[\mathbf{1}_{x \in (A^1 \setminus A^2)}] \varphi_n(x) dx \\
&= \int_{x \in \mathbb{R}^n} \mathbf{Pr}[x \in A^1](1 - \mathbf{Pr}[x \in A^2]) \varphi_n(x) dx \qquad (12) \\
&= \int_{x \in \mathbb{R}^n} (b(x))^N (1 - (b(x))^N) \varphi_n(x) dx. \qquad (13)
\end{aligned}$$

Here equations (12) and (13) are by the independence of the randomly chosen vectors $Z^{1,1}, \ldots, Z^{2,N}$. We shall prove Lemma 29 by showing that

$$b(x)^N (1 - b(x))^N \geq 0.0002244 \text{ for all } x \in \mathbb{R}^n \text{ such that } \|x\| \in [\sqrt{n}, \sqrt{n} + 1]. \qquad (14)$$

For $X \sim \mathcal{N}$, the random variable $\|X\|^2$ is distributed according to a chi-squared distribution χ_n^2 which has mean n and variance $2n$. The Central Limit Theorem implies that as $n \to \infty$, the random variable $(\|X\|^2 - n)/(\sqrt{2n})$ converges to the standard normal distribution $N(0,1)$. Since $N(0,1)$ assigns probability ≈ 0.421 to the interval $[0, \sqrt{2}]$, it follows that for n sufficiently large we have

$$\mathbf{Pr}_{X \sim \mathcal{N}^n}[\|X\|^2 \in [n, n + 2\sqrt{n} + 1]] \geq 0.42$$

which, together with (14), shows that $\mathbf{E}[d(f_1, f_2)] \geq 0.0002244 \cdot 0.42 \geq 0.000094 > \frac{1}{11000}$ and proves Lemma 29.

Now we prove (14). Let

$$N \overset{\text{def}}{=} \frac{1}{\text{cap}(\rho/\sqrt{n})}. \qquad (15)$$

We pause at this point to observe that using the easy bound $\text{cap}(t) \leq e^{-nt^2/2}$, we have $N \geq e^{\rho^2/2}$. In fact, as we now show N is not much larger than this value. We have

$$\begin{aligned}
\text{cap}(\rho/\sqrt{n}) &= r_n \cdot \int_{\rho/\sqrt{n}}^{1} \left(\sqrt{1 - z^2}\right)^{n-3} dz \\
&> r_n \cdot \int_{\rho/\sqrt{n}}^{\rho/(\sqrt{n}-1)} \left(\sqrt{1 - z^2}\right)^{n-3} dz \\
&> r_n \cdot A \cdot B, \text{ where } A = \frac{\rho}{\sqrt{n} - 1} - \frac{\rho}{\sqrt{n}} \text{ and } B = \left(1 - \left(\frac{\rho}{\sqrt{n} - 1}\right)^2\right)^{(n-3)/2}.
\end{aligned}$$

Known bounds give $r_n \geq \frac{1}{3}\sqrt{n}$; an easy computation shows that $A \geq \rho/n$; and some routine asymptotic analysis (using the bound $(1 - 1/m)^m \geq \exp(-1 - \frac{1}{m})$ together with the fact that $\rho \leq n^{1/4}$) gives that $B \geq \frac{1}{4}e^{-\rho^2/2}$. (All these inequalities are for n sufficiently large.) We thus have $\text{cap}(\rho/\sqrt{n}) \geq \frac{1}{12} \cdot (\rho/\sqrt{n}) \cdot e^{-\rho^2/2}$, which implies

$$N \leq 12 \cdot (\sqrt{n}/\rho) \cdot e^{\rho^2/2}. \qquad (16)$$

(Note that with this upper bound on N, we are indeed considering intersections of $2^{O(\ell)}$ halfspaces as claimed in the statement of Theorem 26.) Finally, we assume w.l.o.g. in the sequel that the value N defined by (15) is an integer; the reader can check that there is adequate slack in the bounds to handle rounding N to the nearest integer.

With (15) as our choice of N, for any $\|x\| = \sqrt{n}$ we have $b(x)^N = (1 - 1/N)^N \le e^{-1}$. Since $b(x)$ is a decreasing function of $\|x\|$, we have $b(x)^N \le e^{-1}$ for all $\|x\| \in [\sqrt{n}, \sqrt{n}+1]$. We will show below that

$$\text{for all } \|x\| \in [\sqrt{n}, \sqrt{n}+1], \text{ we have } b(x)^N \ge 0.0002245. \tag{17}$$

(Note that for $\|x\| = \sqrt{n}$, we actually have $b(x)^N \approx e^{-1}$.) Given this, we have

$$\text{for all } \|x\| \in [\sqrt{n}, \sqrt{n}+1], \quad b(x)^N(1 - b(x)^N) \ge 0.0002245 \cdot (1 - .0002245) > .0002244.$$

Since $b(x)$ is decreasing in $\|x\|$, to prove (17) it is enough to give a lower bound on $b(x')$ for $\|x'\| = \sqrt{n}+1$. We will show that

$$\mathrm{cap}\left(\frac{\rho}{\sqrt{n}+1}\right) \le \frac{8.4}{N}. \tag{18}$$

This gives

$$b(x')^N = \left(1 - \mathrm{cap}\left(\frac{\rho}{\sqrt{n}+1}\right)\right)^N \ge (1 - 8.4/N)^N \ge 0.0002245$$

as desired (the last inequality holds for N sufficiently large). Now we prove (18). First recall that

$$\begin{aligned}
\mathrm{cap}\left(\frac{\rho}{\sqrt{n}+1}\right) &= r_n \int_{\frac{\rho}{\sqrt{n}+1}}^{1} (1 - z^2)^{(n-3)/2} dz \\
&= r_n \int_{\frac{\rho}{\sqrt{n}+1}}^{\frac{\rho}{\sqrt{n}}} (1 - z^2)^{(n-3)/2} dz + \mathrm{cap}(\rho/\sqrt{n}) \\
&= r_n \int_{\frac{\rho}{\sqrt{n}+1}}^{\frac{\rho}{\sqrt{n}}} (1 - z^2)^{(n-3)/2} dz + \frac{1}{N}. \qquad (19)
\end{aligned}$$

Now observe that

$$r_n \int_{\frac{\rho}{\sqrt{n}+1}}^{\frac{\rho}{\sqrt{n}}} (1 - z^2)^{(n-3)/2} dz \le r_n \cdot \left(1 - \left(\frac{\rho}{\sqrt{n}+1}\right)^2\right)^{(n-3)/2} \cdot \left(\frac{\rho}{\sqrt{n}} - \frac{\rho}{\sqrt{n}+1}\right) \tag{20}$$

Using Taylor series expansion one can verify that for $0 \le \rho \le n^{1/4}$, we have

$$\lim_{n\to\infty} \frac{\left(1 - \left(\frac{\rho}{\sqrt{n}+1}\right)^2\right)^{(n-3)/2}}{\left(1 - \left(\frac{\rho}{\sqrt{n}-1}\right)^2\right)^{(n-3)/2}} \le e^2 < 7.39$$

(the inequality is an equality for $\rho = n^{1/4}$) and consequently

$$\left(1 - \left(\frac{\rho}{\sqrt{n}+1}\right)^2\right)^{(n-3)/2} \le 7.4 \cdot \left(1 - \left(\frac{\rho}{\sqrt{n}-1}\right)^2\right)^{(n-3)/2} \tag{21}$$

for n sufficiently large. Moreover we trivially have

$$\frac{\rho}{\sqrt{n}} - \frac{\rho}{\sqrt{n}+1} \le \frac{\rho}{\sqrt{n}-1} - \frac{\rho}{\sqrt{n}}. \tag{22}$$

Combining (21) and (22), we have that

$$\begin{aligned}
(20) \quad &\le \quad r_n \cdot 7.4 \cdot \left(1 - \left(\frac{\rho}{\sqrt{n}-1}\right)^2\right)^{(n-3)/2} \cdot \left(\frac{\rho}{\sqrt{n}-1} - \frac{\rho}{\sqrt{n}}\right) \\
&\le \quad 7.4 \cdot r_n \int_{\frac{\rho}{\sqrt{n}}}^{\frac{\rho}{\sqrt{n}-1}} (1-z^2)^{(n-3)/2} dz \\
&< \quad 7.4 \cdot r_n \int_{\frac{\rho}{\sqrt{n}}}^{1} (1-z^2)^{(n-3)/2} dz = 7.4 \cdot \mathrm{cap}(\rho/\sqrt{n}) = \frac{7.4}{N}. \qquad (23)
\end{aligned}$$

Combining (19), (20) and (23) we obtain (18). This concludes the proof of Lemma 29 and hence of Theorem 27, and so our sample complexity bound, Theorem 26, is proved.

E Boolean surface area

E.1 Motivation Given the very useful connection between noise sensitivity and surface area in Gaussian space, Corollary 14, it is natural to wonder if there is a similar connection in the setting of the Boolean cube under the uniform distribution. In some senses the case of $\{-1,1\}^N$ is a *generalization* of the case of $(\mathbb{R}^n, \mathcal{N}^n)$: this is because we can simulate a Gaussian random variable with Boolean ones:

$$\frac{\sum_{i=1}^m x_i}{\sqrt{m}} \approx \mathcal{N}$$

when the string $x \in \{-1,1\}^m$ is drawn from the uniform distribution. There is a long history of proving results in Gaussian space by first deriving them in the Boolean case and then making a limiting argument; notable examples of this include Gross's work on the logarithmic Sobolev and hypercontractive inequalities [Gro75] and Bobkov's proof of the Gaussian isoperimetric inequality [Bob97].

The notions of noise stability and sensitivity for Boolean functions $f : \{-1,1\}^n \to \mathbb{R}$ are well-known [BKS99]. In place of the Ornstein-Uhlenbeck operator we have the *Bonami-Beckner* operator, also denoted T_ρ, acting as

$$(T_\rho f)(x) = \mathop{\mathbf{E}}_{y \sim_\rho x}[f(y)];$$

here $y \sim_\rho x$ means that y is chosen by keeping each bit of x fixed with probability ρ and randomizing it with probability $1-\rho$, independently across coordinates. The noise stability and sensitivity of f are now defined by formally repeating the definitions in the Gaussian case. The analogous expression to (2) for $T_\rho f$ in terms of f's Fourier (Walsh) coefficients continues to hold. As a result, we have the same relationship between low-degree Fourier concentration, noise sensitivity, and learning as in Section 2.2; see [KOS04, KKMS05].

Unfortunately for learning purposes, it is somewhat difficult to prove noise sensitivity upper bounds for natural classes of Boolean functions. The work [KOS04] relied on a clever theorem of Peres [Per04] which states that $\mathbb{NS}_\delta(f) \le O(\sqrt{\delta})$ for any Boolean halfspace f. This immediately implies that an intersection of k halfspaces has noise sensitivity at most $k \cdot O(\sqrt{\delta})$. [KOS04] conjectured that in fact the much better upper bound of $\sqrt{\log k} \cdot O(\sqrt{\delta})$ should hold. We now know, via Corollary 14 and Nazarov's Theorem 20, that the conjecture holds in Gaussian space. This provides significant motivation for seeking a connection between noise sensitivity and "surface area" in the Boolean setting. In fact, we find the desired connection; unfortunately, it does not prove to be quite as useful as hoped.

E.2 Overview of Boolean surface area There is a likely candidate for the proper analogue of "surface area" in the Boolean setting. The proof of Theorem 13 uses $\int |\nabla f|$ as a surrogate for surface area (cf. the discussion after its statement), and there is a well-known notion of "gradient" in the Boolean cube (see e.g. [Bob97]): for $f : \{-1,1\}^n \to \mathbb{R}$, this is:

$$\nabla f(x) = (\mathrm{D}_1 f(x), \ldots, \mathrm{D}_n f(n)),$$

where D_i is the "ith discrete derivative operator", defined by

$$\mathrm{D}_i f(x) = \frac{f(x^{(i=1)}) - f(x^{(i=-1)})}{2},$$

with $x^{(i=b)}$ denoting the string x with its ith coordinate changed to b. Note that when $f : \{-1,1\}^n \to \{-1,1\}$,

$$(\mathrm{D}_i f(x))^2 = \begin{cases} 1 & \text{if } f \text{ is "sensitive" to the } i\text{th coordinate of } x, \\ 0 & \text{else,} \end{cases}$$

and hence the "length of the gradient" is

$$|\nabla f(x)| = \sqrt{\sum_{i=1}^{n} (\mathrm{D}_i f(x))^2} = \sqrt{\text{\# of sensitive coordinates for } f \text{ on } x}.$$

Thus the following definition is natural:

Definition 30. *The "Boolean surface area" of a function $f : \{-1,1\}^n \to \{-1,1\}$ is defined to be*

$$\Gamma(f) = \mathbf{E}[|\nabla f|] = \mathop{\mathbf{E}}_{x}[\sqrt{\textit{\# of sensitive coordinates for } f \textit{ on } x}].$$

Here and throughout this section, $\mathbf{E}[\cdot]$ is with respect to the uniform probability distribution on $\{-1,1\}^n$.

The Boolean surface area appears to have been first introduced and studied by Talagrand [Tal93]. (Actually, Talagrand studied a variant,

$$\mathop{\mathbf{E}}_{x}[\sqrt{\mathbf{1}_{f(x)=1} \cdot \text{\# of sensitive coordinates for } f \text{ on } x}],$$

which is slightly different for f with $|\mathbf{E}[f]|$ very close to 1.) He connected it to various topics, including discrete isoperimetry, logarithmic Sobolev equations, percolation, and Banach space inequalities. It was also used by Bobkov [Bob97] in his proof of the Gaussian isoperimetric inequality and by Tillich and Zémor [TZ00] in the context of coding theory.

One basic fact about Boolean surface area is the following:

$$\Gamma(f) = \mathbf{E}[\sqrt{|\nabla f|^2}] \leq \sqrt{\mathbf{E}[|\nabla f|^2]} = \sqrt{\mathbb{I}(f)},$$

where $\mathbb{I}(f) = \sum_S |S| \hat{f}(S)^2$ is the "total influence" of f. For $f : \{-1,1\}^n \to \{-1,1\}$ this is also called the "average sensitivity" of f, since

$$\mathbb{I}(f) = \mathop{\mathbf{E}}_{x}[\text{\# of sensitive coordinates for } f \text{ on } x].$$

In this case we also write

$$\mathrm{Inf}_i(f) = \mathop{\mathbf{Pr}}_{x}[f \text{ is sensitive to the } i\text{th coordinate of } x],$$

for the "influence" of the ith coordinate on f, and we have $\mathbb{I}(f) = \sum_i \mathrm{Inf}_i(f)$. For monotone functions, $\mathrm{Inf}_i(f) = \hat{f}(i)$.

It is well-known that $\mathbb{I}(f) \le O(\sqrt{n})$ whenever f is a monotone Boolean function, and thus $\Gamma(f) \le O(n^{1/4})$ for monotone f. This seems to be the analogue of Ball's upper bound for Gaussian surface area of convex sets. As further evidence, Talagrand [Tal96] exhibited a monotone Boolean function f with $\Gamma(f) \ge \Omega(n^{1/4})$, and his construction strongly prefigures the lower bound of Nazarov: it can be viewed as the intersection of $2^{\Theta(\sqrt{n})}$ random disjunctions. As more evidence that we are on the right track, the two Boolean halfspaces which are arguably most natural — namely Dictator ($f(x) = x_i$) and Majority — both have $O(1)$ Boolean surface area, just as in the Gaussian case. In Majority's case, this bound holds because a $\Theta(1/\sqrt{n})$ fraction of inputs have sensitivity $n/2$ and the remaining inputs have sensitivity 0. (Bobkov, Götze, and Houdré [SBH01] generalized this to arbitrary symmetric threshold functions.)

E.3 The Boolean version of Corollary 14 In this section we provide the Boolean analogue of Corollary 14:

Theorem 31. *For any $f : \{-1,1\}^n \to \{-1,1\}$ and any $0 \le \delta \le 1$,*

$$\mathbb{NS}_\delta(f) \le \tfrac{\sqrt{\pi}}{2}\sqrt{\delta} \cdot \Gamma(f).$$

The result relies on the following theorem of Bobkov and Götze [BG99]:

Theorem 32. *Let (Ω_i, μ_i) be probability spaces, $i = 1 \dots n$, and write (Ω, μ) for the product probability space. Assuming $f : \Omega \to [0,1]$ is measurable, we have*

$$U(\mathop{\mathbf{E}}_{\mu}[f]) \le \mathop{\mathbf{E}}_{\mu}\left[\sqrt{U(f)^2 + 2\|\nabla f\|_\mu^2}\right]. \tag{24}$$

Here U is the Gaussian isoperimetric function, and

$$\|\nabla f\|_\mu^2 \stackrel{def}{=} \sum_{i=1}^{n} \mathrm{Var}_{\mu_i}[f].$$

We now prove Theorem 31:

Proof. Consider the Bobkov-Götze inequality (24) in the special case that f's range is $\{0,1\}$; since $U(0) = U(1) = 0$, this eliminates the $U(f)^2$ in the right-hand side of (24). We will also eliminate the U on the left-hand side of (24) by using the elementary inequality

$$U(t) \ge \sqrt{2/\pi}(\tfrac{1}{2} - 2(\tfrac{1}{2} - t)^2).$$

Thus for $f : \Omega \to \{0,1\}$ we have

$$\sqrt{2/\pi}(\tfrac{1}{2} - 2(\tfrac{1}{2} - \mathop{\mathbf{E}}_{\mu}[f])^2) \le \mathop{\mathbf{E}}_{\mu}[\sqrt{2}\|\nabla f\|_\mu] \qquad \Rightarrow \qquad \tfrac{1}{2} - 2(\tfrac{1}{2} - \mathop{\mathbf{E}}_{\mu}[f])^2 \le \sqrt{\pi}\mathop{\mathbf{E}}_{\mu}[\|\nabla f\|_\mu] \tag{25}$$

Suppose we fix an $x \in \{-1,1\}^n$ and a $\rho \in [0,1]$. We define $\Omega_i = \{-1,1\}$ and μ_i to be the biased measure which gives probability $\frac{1}{2} + \frac{1}{2}\rho$ to x_i and probability $\frac{1}{2} - \frac{1}{2}\rho$ to $-x_i$. Note that with this choice one can easily check that $\mathbf{E}_\mu[f] = (T_\rho f)(x)$ and that

$$\mathrm{Var}_{\mu_i}[f(y)] = (1 - \rho^2) \cdot (\mathrm{D}_i f(y))^2.$$

Hence

$$\|\nabla f(y)\|_\mu^2 = \sum_{i=1}^n \mathrm{Var}_{\mu_i}[f(y)] = (1-\rho^2)|\nabla f(y)|^2,$$

where on the right side we have the usual, uniform-distribution discrete gradient on $\{-1,1\}^n$. Substituting into (25) we get

$$\tfrac{1}{2} - 2(\tfrac{1}{2} - (T_\rho f)(x))^2 \leq \sqrt{\pi}\sqrt{1-\rho^2}(T_\rho[|\nabla f|])(x).$$

We now revert f's range to $\{-1,1\}$, replacing f by $\frac{1}{2}+\frac{1}{2}f$ in the above. This yields

$$\tfrac{1}{2} - \tfrac{1}{2}((T_\rho f)(x))^2 \leq \tfrac{\sqrt{\pi}}{2}\sqrt{1-\rho^2}(T_\rho[|\nabla f|])(x).$$

Finally, if we take the expectation of this inequality over a uniform choice of $x \in \{-1,1\}^n$, we get precisely

$$\mathbb{NS}_{1-\rho^2}(f) \leq \tfrac{\sqrt{\pi}}{2}\sqrt{1-\rho^2}\,\mathbf{E}[|\nabla f|].$$

Setting $\rho^2 = 1-\delta$ completes the proof. $\square$

By the Fourier concentration method, we now conclude that our main Theorem 5 holds in the Boolean setting:

Theorem 33. *Let $\mathcal{C}$ denote the class of all Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ with $\Gamma(f) \leq s$. Then under the uniform distribution, $\mathcal{C}$ is PAC learnable to accuracy ϵ in time $n^{O(s^2/\epsilon^2)}$ and agnostically learnable in time $n^{O(s^2/\epsilon^4)}$.*

We recall that Bshouty and Tamon [BT96] showed that any Boolean function f has Fourier concentration $\mathbb{I}(f)/\epsilon$, and hence can be learned under the uniform distribution in time $n^{\mathbb{I}(f)/\epsilon}$. Our bound is an improvement on theirs in so far as $\Gamma(f)^2 \leq \mathbb{I}(f)$ for every Boolean function f (and the difference can be substantial, as for the Majority function which has $\Gamma = O(1)$ and $\mathbb{I} = \Theta(\sqrt{n})$); however our bound has an additional factor of $1/\epsilon$ in the exponent.

E.4 Boolean surface area of halfspaces Thus far it seems the Boolean theory is matching the Gaussian theory perfectly. Since halfspaces have Gaussian surface area $O(1)$ it is natural to expect that the same bound holds for Boolean surface area; this would allow us to recover the results of [KOS04]. Bobkov, Götze, and Houdré [SBH01] considered this statement but commented that they did not know how to prove it.

Surprisingly, the statement turns out to be false. The correct answer for the maximum Boolean surface area of any n-variable halfspace is $\Theta(\sqrt{\log n})$, and the halfspace that achieves the maximum — essentially $\mathrm{sgn}(\sum x_i/\sqrt{i})$ — is an unusual example.

Theorem 34.

1. *Every Boolean halfspace $f(x) = \mathrm{sgn}(\sum_i a_i x_i - \theta)$ satisfies $\Gamma(f) \leq O(\sqrt{\log n})$.*
2. *Let η_j denote $\sqrt{j+1} - \sqrt{j}$, so $\eta_j \sim \frac{1}{2\sqrt{j}}$. Then for even n, the Boolean halfspace $f(x) = \mathrm{sgn}(\sum_{i=1}^n \eta_{\lceil i/2 \rceil} x_i)$ satisfies $\Gamma(f) \geq \Omega(\sqrt{\log n})$.*

We expect that the slightly simpler halfspace $\mathrm{sgn}(\sum x_i/\sqrt{i})$ also has Boolean surface area at least $\Theta(\sqrt{\log n})$, but we have not verified this. Since Theorem 34 is somewhat tangential to our main concerns in this paper, we defer its proof to the full version.

We conclude this section by commenting that although we only have $\Gamma(f) \leq O(\sqrt{\log n})$ for Boolean halfspaces, the approach of bounding noise sensitivity by surface area may still prove useful for learning. It may possibly be easier to prove that the intersection of k Boolean halfspaces has surface area $O(\sqrt{\log k}\sqrt{\log n})$ than to prove the conjectured $O(\sqrt{\log k}\sqrt{\delta})$ bound on noise sensitivity. If this surface area bound could be established it would yield an $n^{O(\log k \log n/\epsilon^2)}$-time learning algorithm, which would still be quite strong.