Weak learning convex sets under normal distributions

Anindya De ANINDYAD@CIS.UPENN.EDU
University of Pennsylvania

Rocco A. Servedio ROCCO@CS.COLUMBIA.EDU
Columbia University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper addresses the following natural question: can efficient algorithms weakly learn convex sets under normal distributions? Strong learnability of convex sets under normal distributions is well understood, with near-matching upper and lower bounds given in Klivans et al. (2008), but prior to the current work nothing seems to have been known about weak learning. We essentially answer this question, giving near-matching algorithms and lower bounds.

For our positive result, we give a $\mathrm{poly}(n)$-time algorithm that can weakly learn the class of convex sets to advantage $\Omega(1/\sqrt{n})$ using only random examples drawn from the background Gaussian distribution. Our algorithm and analysis are based on a new "density increment" result for convex sets, which we prove using tools from isoperimetry.

We also give an information-theoretic lower bound showing that $O(\log(n)/\sqrt{n})$ advantage is best possible even for algorithms that are allowed to make $\mathrm{poly}(n)$ many membership queries.

Keywords: weak learning, convex geometry, Gaussian space

1. Introduction

Background and motivation. Several results in Boolean function analysis and computational learning theory suggest an analogy between convex sets in Gaussian space and monotone Boolean functions[1] with respect to the uniform distribution over the hypercube. As an example, Bshouty and Tamon Bshouty and Tamon (1996) gave an algorithm that learns monotone Boolean functions over the n-dimensional hypercube to any constant accuracy in a running time of $n^{O(\sqrt{n})}$. Much later, Klivans, O'Donnell and Servedio Klivans et al. (2008) gave an algorithm that learns convex sets over n-dimensional Gaussian space with the same running time. While the underlying technical tools in the proofs of correctness are different, the algorithms in Klivans et al. (2008) and Bshouty and Tamon (1996) are essentially the same: Bshouty and Tamon (1996) (respectively Klivans et al. (2008)) show that the Fourier spectrum (respectively Hermite spectrum[2]) of monotone functions (respectively convex sets) is concentrated in the first $O(\sqrt{n})$ levels. Other analogies between convex sets and monotone functions are known as well; for example, an old result of Harris Harris (1960) and Kleitman Kleitman (1966) shows that monotone Boolean functions over $\{-1,1\}^n$ are positively correlated. The famous Gaussian correlation conjecture (now a theorem due to Royen Royen (2014)) asserts the same for symmetric convex sets under the Gaussian distribution. We note that while the assertions are analogous, the proof techniques are very different, and indeed the Gaussian

1. Recall that a function $f : \{-1,1\}^n \to \{-1,1\}$ is monotone if $f(x) \leq f(y)$ whenever $x_i \leq y_i$ for all $i \in [n]$.
2. The Hermite polynomials form an orthonormal basis for the space of square-integrable real-valued functions over Gaussian space; the Hermite spectrum of a function over Gaussian space is analogous to the familiar Fourier spectrum of a function over the Boolean hypercube.

correlation conjecture was open for more than half a century while the Harris-Kleitman theorem has a simple one-paragraph inductive proof.

Despite these analogies between convex sets and monotone functions, there are a number of prominent gaps in our algorithmic understanding of convex sets when compared against monotone functions. We list two examples below:

1. Nearly matching $\mathrm{poly}(n)$ upper and lower bounds are known for the query complexity of testing monotone functions over the n-dimensional Boolean hypercube Fischer et al. (2002); Khot et al. (2015); Chakrabarty and Seshadhri (2016); Belovs and Blais (2016); Chen et al. (2015, 2017b). However, the problem of convexity testing over the Gaussian space is essentially wide open, with the best known upper bound (in Chen et al. (2017a)) being $n^{O(\sqrt{n})}$ queries and no nontrivial lower bounds being known.

2. Kearns, Li and Valiant Kearns et al. (1994) showed that the class of all monotone Boolean functions over $\{-1,1\}^n$ is *weakly learnable* under the uniform distribution in polynomial time, meaning that the output hypothesis h satisfies $\mathbf{Pr}_{\boldsymbol{x}\in\{-1,1\}^n}[h(\boldsymbol{x}) = f(\boldsymbol{x})] \geq 1/2 + 1/\mathrm{poly}(n)$, where $f : \{-1,1\}^n \to \{-1,1\}$ is the target monotone function. Kearns et al. (1994) achieved an advantage of $\Omega(1/n)$ over $1/2$; this advantage was improved by Blum, Burch and Langford Blum et al. (1998) to $\Omega(n^{-1/2})$ and subsequently by O'Donnell and Wimmer O'Donnell and Wimmer (2009) to $\Omega(n^{-1/2}\log n)$ which is optimal up to constant factors for $\mathrm{poly}(n)$-time learning algorithms. On the other hand, prior to the current work, nothing non-trivial was known about weak learning convex sets under the Gaussian measure.

The main contribution of this work is in giving upper and lower bounds on the weak learnability of convex sets in Gaussian space, thus addressing item 2 above.

1.1. Learning convex sets in Gaussian space

As mentioned earlier, in Klivans et al. (2008) Klivans et al. showed that convex sets are *strongly learnable* (i.e. learnable to accuracy $1-\varepsilon$ for any $\varepsilon > 0$) in time $n^{O(\sqrt{n}/\varepsilon^2)}$ under the Gaussian distribution, given only random examples drawn from the Gaussian distribution. Up to a mildly better dependence on ε, this matches the running time of the algorithm of Bshouty and Tamon (1996) for learning monotone functions on the hypercube.

However, there is a large gap in the state of the art between monotone Boolean functions on the cube and convex sets in the Gaussian space when it comes to *weak learning*. In particular, while O'Donnell and Wimmer (2009) showed that monotone functions can be weakly learned to accuracy $1/2 + \Omega(n^{-1/2}\log n)$ in polynomial time, prior to this work nothing better than the $n^{\sqrt{n}}$ running time of Klivans et al. (2008) was known for weakly learning convex sets to any nontrivial accuracy (even accuracy $1/2 + \exp(-n)$).

Our main positive result: An algorithm for weak learning convex sets. The main algorithmic contribution of this paper is to bridge this gap and give a polynomial-time weak learning algorithm for convex sets. We prove the following:[3]

3. As stated Theorem 1 only deals with the standard Gaussian distribution $N(0,1^n)$, but since convexity is preserved under affine transformations, the result holds for weak learning with respect to any Gaussian distribution $N(\mu,\Sigma)$.

Theorem 1 (Algorithm for weak learning convex sets) *There is a* $\mathrm{poly}(n)$*-time algorithm which uses only random samples from* $N(0,1)^n$ *and weak learns any unknown convex set* $K \subseteq \mathbb{R}^n$ *to accuracy* $1/2 + \Omega(1/\sqrt{n})$ *under* $N(0,1)^n$.

Our main negative result: A lower bound for weak learning convex sets. We complement Theorem 1 with an information theoretic lower bound. This lower bound shows that any $\mathrm{poly}(n)$-time algorithm, even one which is allowed to query the target function on arbitrary inputs of its choosing, cannot achieve a significantly better advantage than our algorithm achieves even for learning the restricted class of symmetric convex sets (for which $x \in K$ iff $-x \in K$):

Theorem 2 (Lower bound for weak learning symmetric convex sets) *For sufficiently large* n*, for any* $s \geq n$*, there is a distribution* $\mathcal{D}$ *over symmetric convex sets with the following property: for a target convex set* $\boldsymbol{f} \sim \mathcal{D}$*, for any membership-query (black box query) algorithm* A *making at most* s *many queries to* $\boldsymbol{f}$*, the expected error of* A *(the probability over* $\boldsymbol{f} \sim \mathcal{D}$*, over any internal randomness of* A*, and over a random Gaussian* $\boldsymbol{x} \sim N(0,1^n)$*, that the output hypothesis* h *of* A *predicts incorrectly on* $\boldsymbol{x}$*) is at least* $1/2 - \frac{O(\log s)}{n^{1/2}}$.

Theorem 2 shows that the advantage of our weak learner for convex sets (Theorem 1) is tight up to a logarithmic factor for polynomial time algorithms.

1.2. Techniques for our positive result

In this subsection we give a high-level overview of the ideas that underlie our algorithm for weak learning an unknown convex set $K \subseteq \mathbb{R}^n$. To present these ideas we need a few simple definitions:

- The *Gaussian volume* of K is $\mathrm{vol}(K) := \mathbf{Pr}_{\boldsymbol{g} \sim N(0,1)^n}[\boldsymbol{g} \in K]$.
- If K contains the origin, the *inradius* of K is $r_{\mathsf{in}}(K) := \sup\{w \geq 0 : B(0,w) \subseteq K\}$, where $B(0,w)$ is the origin-centered ball of radius w in $\mathbb{R}^n$. If K does not contain the origin then we say that the inradius of K is $-\infty$.

A first easy observation is that if $\mathrm{vol}(K)$ (the Gaussian volume of K under the standard $N(0,1)^n$ distribution) is not very close to $1/2$, then either the constant-0 or constant-1 function is an acceptable weak hypothesis that achieves accuracy significantly greater than $1/2$. Thus we may assume that $\mathrm{vol}(K) \approx 1/2$.

In the rest of the argument we consider two cases depending on whether or not the inradius of K is "large" (where for this intuitive discussion we take "large" to mean "at least some (small) absolute constant independent of n"). The first case is that the inradius is not large: in this case, by the separating hyperplane theorem the convex set K is contained in a halfspace H whose separating hyperplane passes close to the origin. Such a halfspace H must have $\mathrm{vol}(H)$ bounded away from 1; using the fact that $K \subseteq H$, it is not too difficult see that the halfspace H is in fact a weak hypothesis for K (in fact, with accuracy $1/2 + \Theta(1)$). Coupling this with existing results on agnostic learning of halfspaces Awasthi et al. (2013); Kalai et al. (2008); Diakonikolas et al. (2018), a weak learning algorithm for K follows easily (again with accuracy $1/2 + \Theta(1)$).

Thus it remains to handle the second (and more challenging) case, in which the inradius is large. The main technical tool that we use for this case is the following structural result, which together with the above discussion easily yields Theorem 1:

Theorem 3 (Structural result, informal statement) *If $K \subseteq \mathbb{R}^n$ is a convex set with inradius bounded away from 0, then (for sufficiently large n) one of the following three hypotheses h has $\mathbf{Pr}_{\boldsymbol{x} \sim N(0,1)^n}[h(\boldsymbol{x}) = K(\boldsymbol{x})] \geq 1/2 + \Omega(1/\sqrt{n})$: h_0 = the empty set, h_1 = all of $\mathbb{R}^n$, or $h_{1/2}$ = the origin-centered ball of Gaussian volume $1/2$.*

Theorem 3 is analogous to a result of Blum et al. (1998), who showed that for any monotone function $f : \{-1,1\}^n \to \{-1,1\}$ over the Boolean hypercube, one of the following three functions achieves an advantage of $\Omega(n^{-1/2})$ with respect to the uniform distribution: the constant 1 function, the constant -1 function, or the majority function. Our proof of Theorem 3 is inspired by the argument of Blum et al. (1998). The central ingredient of the Blum et al. (1998) proof is the *Kruskal-Katona theorem* Kruskal (1963); Katona (1968); Bollobás and Thomason (1987); Lovász (1981) over the Boolean hypercube; this is a "density increment" result for monotone Boolean functions, which asserts that the density of the 1-set of a monotone function must increases non-trivially over the successive "slices" $\{x \in \{-1,1\}^n : \sum x_i = \ell\}$ of $\{-1,1\}^n$. Similarly, at the heart of our Theorem 3 is a new density increment result for convex subsets of $\mathbb{R}^n$ with positive inradius; we explain this new result below.

A density increment for convex sets with positive inradius. Inspired by the Kruskal-Katona theorem, we begin by identifying an analogue of hypercube slices in the setting of Gaussian space. The most obvious choice is to consider spherical shells; namely, for $r > 0$, define the radius-r spherical shell to be $\mathbb{S}_r^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = r\}$.

Given a convex set $K \subseteq \mathbb{R}^n$, we define the *shell-density function* $\alpha_K : (0,\infty) \to [0,1]$ to be

$$\alpha_K(r) := \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{S}_r^{n-1}}[\boldsymbol{x} \in K]. \tag{1}$$

Having defined $\alpha_K(\cdot)$, the most obvious way to give an analogue of Kruskal-Katona for convex sets is to conjecture that for a convex set K, $\alpha_K(\cdot)$ is a non-increasing function, and further, that as long as $\alpha_K(r)$ is bounded away from 0 and 1, it exhibits a non-trivial rate of decay as r increases. However, a moment's thought shows that this is not quite true because of the following examples:

1. Let $K \subseteq \mathbb{R}^n$ be a convex set with positive Gaussian volume whose closest point to the origin is at some distance $t > 0$. Then the shell density function $\alpha_K(r)$ is zero for $0 < r \leq t$ but subsequently becomes positive. Thus for $\alpha_K(\cdot)$ to be non-increasing, we require $0^n \in K$.

   In fact, it is easy to see that if $0^n \in K$ and K is convex then $\alpha_K(\cdot)$ is in fact non-increasing (since by convexity the intersection of K with any ray extending from the origin is a line segment starting at the origin). However, this does not mean that there is an actual decay in the value of α_K, as witnessed by the next example:

2. Let K be an origin-centered halfspace, i.e. $K = \{x : w \cdot x \geq 0\}$ for some nonzero $w \in \mathbb{R}^n$. K is convex and $0^n \in K$, but $\alpha_K(r) = 1/2$ for all $r > 0$, and hence $\alpha_K(r)$ exhibits no decay as r increases.

The second example above shows that in order for $\alpha_K(\cdot)$ to have decay, it is not enough for the origin to belong to K; rather, what is needed is for K to have a positive inradius. Our density increment result, stated below in simplified form, shows that in fact the above examples are essentially the only obstructions to getting a decay for $\alpha_K(r)$.

In order to avoid a proliferation of parameters at this early stage, for now we only state a corollary of our more general result, Theorem 12 (the more general result does not put any restriction on the value of $\alpha_K(r)$):

Theorem 4 (Density increment for convex sets with positive inradius, informal statement) *Let $K \subseteq \mathbb{R}^n$ be a convex set with inradius $r_{\mathsf{in}} > 0$. Let $r > r_{\mathsf{in}}$ be such that $0.1 \leq \alpha_K(r) \leq 0.9$ and $\alpha_K(\cdot)$ is differentiable at r. Then $\frac{d\alpha_K(r)}{dr} \leq -\Omega(\frac{r_{\mathsf{in}}\sqrt{n}}{r^2})$.*

In a preliminary version of this paper De and Servedio (2019) we gave a self-contained proof of a (quantitatively weaker) version of Theorem 4 by combining elementary geometric arguments with an argument inspired by the central technical lemma of Raz (1999). In the current paper we give a shorter proof which employs the isoperimetric theorem Lévy (1951); Ledoux (2001) for high-dimensional spheres; this approach was suggested by an anonymous reviewer of the earlier version of this paper.

1.3. Techniques for our negative result

As the first stage in our proof of Theorem 2, we construct a "hard" distribution $\mathcal{D}_{\text{ideal}}$ (which is different from the final distribution $\mathcal{D}$, as described below) over symmetric convex subsets of $\mathbb{R}^n$. The distribution $\mathcal{D}_{\text{ideal}}$ is a continuous distribution defined in terms of a Poisson point process; a draw from $\mathcal{D}_{\text{ideal}}$ is essentially a random symmetric polytope with $\mathsf{poly}(s)$ facets where the hyperplane defining each facet is at distance around $O(\sqrt{\log s})$ away from the origin. We analyze the setting in which a learning algorithm is not allowed to make *any* queries to a target function $\boldsymbol{f}$ that is drawn from $\mathcal{D}_{\text{ideal}}$. In this setting, the maximum possible accuracy of any zero-query learning algorithm is achieved by the Bayes optimal classifier for $\mathcal{D}_{\text{ideal}}$ (which we denote by $BO_{\mathcal{D}_{\text{ideal}}}$), which simply labels each $x \in \mathbb{R}^n$ according to whether it is more likely to be labeled positive or negative by a randomly selected target concept $\boldsymbol{f} \sim \mathcal{D}_{\text{ideal}}$. The construction of $\mathcal{D}_{\text{ideal}}$ is well suited to facilitate such an analysis, and indeed we show that the average accuracy of $BO_{\mathcal{D}_{\text{ideal}}}(\boldsymbol{x})$ for $\boldsymbol{x} \sim N(0,1)^n$ is at most $1/2 + \frac{O(\log s)}{n^{1/2}}$.

It becomes tricky to analyze $\mathcal{D}_{\text{ideal}}$ when a learning algorithm is actually allowed to make queries to a target function $\boldsymbol{f} \sim \mathcal{D}_{\text{ideal}}$. To deal with this, in the second stage of the proof we discretize the distribution $\mathcal{D}_{\text{ideal}}$ to construct the actual hard distribution $\mathcal{D}$ (which is finitely supported). The discretization is carefully done to retain some crucial geometric properties, and in particular to ensure that for "most" $\boldsymbol{x}$ (again sampled from $N(0,1)^n$), $\mathbf{Pr}_{\boldsymbol{f}\sim\mathcal{D}_{\text{ideal}}}[\boldsymbol{f}(\boldsymbol{x}) = 1]$ is close to $\mathbf{Pr}_{\boldsymbol{f}\sim\mathcal{D}}[\boldsymbol{f}(\boldsymbol{x}) = 1]$. This implies that the average advantage of the Bayes optimal classifier for $\boldsymbol{f} \sim \mathcal{D}$ (corresponding again to the best possible zero-query learning algorithm), denoted by $BO_{\mathcal{D}}$, remains bounded by $\frac{O(\log s)}{n^{1/2}}$.

In the third and final stage of the proof, we consider the case when the learning algorithm is allowed to makes s queries to an unknown target function $\boldsymbol{f} \sim \mathcal{D}$. We show that for any choice of s query points $\overline{y} = (y_1, \ldots, y_s)$, with high probability over both $\boldsymbol{f} \sim \mathcal{D}$ and $\boldsymbol{x} \sim N(0,1)^n$, the advantage of the optimal classifier is close to that achieved by $BO_{\mathcal{D}}$ (see Appendix A.3). We note that the third stage of our proof, and the general flavor of the analysis used to establish it, follows the lower bound approach of Blum, Burch and Langford Blum et al. (1998), who showed that no s-query algorithm in the membership query model can achieve an advantage of $\omega(\frac{\log s}{\sqrt{n}})$ over random guessing to learn monotone functions under the uniform distribution on $\{-1,1\}^n$. (Stages 1 and 2

of our proof, which are necessary because of the continuous setting of our lower bound, do not have analogues in Blum et al. (1998).)

2. Preliminaries

Background results from geometry. We first recall the definition of the shell density function $\alpha_K(\cdot)$ from Equation (1): for $r \geq 0$, we have $\alpha_K(r) := \mathbf{Pr}_{\boldsymbol{x} \in \mathbb{S}_r^{n-1}}[\boldsymbol{x} \in K]$, so $\alpha_K(r)$ equals the fraction of the origin-centered radius-r sphere which lies in K, i.e. the normalized Haar measure of $K \cap \mathbb{S}_r^{n-1}$. We write $\mu(\cdot)$ to denote the normalized Haar measure (which can be thought of as simply the uniform measure on $\mathbb{S}^{n-1}$), so $\alpha_K(r) = \mu(\mathbb{S}_r^{n-1} \cap K)$. A view which will be useful later is that $\alpha_K(r)$ is the probability that a random Gaussian-distributed point $\boldsymbol{g} \sim N(0,1)^n$ lies in K, conditioned on $\|\boldsymbol{g}\| = r$.

An easy fact about the function $\alpha_K(\cdot)$ is the following:

Fact 5 *If K is convex and $0^n \in K$ then $\alpha_K(\cdot)$ is non-increasing.*

Proof By convexity, if $x \in K$ then $\lambda x \in K$ for any $\lambda \in [0,1]$. This immediately implies that $\mathbf{Pr}_{\boldsymbol{x} \in \mathbb{S}_r^{n-1}}[\boldsymbol{x} \in K] \leq \mathbf{Pr}_{\boldsymbol{x} \in \mathbb{S}_{\lambda r}^{n-1}}[\boldsymbol{x} \in K]$ and consequently $\alpha_K(\cdot)$ is non-increasing. ■

Next, we recall the isoperimetric theorem on the sphere. We use the shorthand $\mathbb{S}^{n-1}$ to denote $\mathbb{S}_1^{n-1}$, the unit sphere in n dimensions. We recall the definition of the geodesic distance on the unit sphere as well as the notion of a spherical cap:

Definition 6 *Let $x, y \in \mathbb{S}^{n-1}$. The* geodesic distance *between x and y, denoted $d_{\mathsf{geo}}(x,y)$, is defined to be $d_{\mathsf{geo}}(x,y) = \arccos(\langle x, y \rangle)$. A set $A \subseteq \mathbb{S}^{n-1}$ is said to be a* spherical cap *if there exists $x^* \in \mathbb{S}^{n-1}$ and $\theta^* \in [0,\pi]$ such that $A = \{x \in \mathbb{S}^{n-1} : d_{\mathsf{geo}}(x,x^*) \leq \theta^*\}$.*

We recall the spherical isoperimetry theorem, which states that spherical caps have the smallest neighborhoods over all measurable sets of a given area:

Theorem 7 (Spherical isoperimetry Lévy (1951); Ledoux (2001)) *For any measurable set $A \subseteq \mathbb{S}^{n-1}$ and any $r \in [0,\pi]$, we define $A_{r,\mathsf{geo}} = \{z \in \mathbb{S}^{n-1} : d_{\mathsf{geo}}(x,z) \leq r \text{ for some } x \in A\}$. Then $\mu(A_{r,\mathsf{geo}}) \geq \mu(H_{r,\mathsf{geo}})$, where H is a spherical cap such that $\mu(H) = \mu(A)$.*

Background results on the Gaussian distribution. We endow $\mathbb{R}^n$ with the standard Gaussian measure $N(0,1)^n$ (i.e. each coordinate is independently distributed as a standard normal). As stated earlier the *Gaussian volume* of a region $K \subseteq \mathbb{R}^n$, denoted $\mathrm{vol}(K)$, is $\mathbf{Pr}_{\boldsymbol{g} \sim N(0,1)^n}[K(\boldsymbol{g}) = 1]$.

We note some basic but crucial properties of the chi-squared distribution with n degrees of freedom. Recall that a non-negative random variable $\boldsymbol{r}^2$ is distributed according to the chi-squared distribution $\chi^2(n)$ if $\boldsymbol{r}^2 = \boldsymbol{g}_1^2 + \cdots + \boldsymbol{g}_n^2$ where $\boldsymbol{g} \sim N(0,1)^n$, and that a draw from the chi distribution $\chi(n)$ is obtained by making a draw from $\chi^2(n)$ and then taking the square root. We recall the following tail bound:

Lemma 8 (Tail bound for the chi-squared distribution Johnstone (2001)) *Let $\boldsymbol{r}^2 \sim \chi^2(n)$. Then we have*

$$\mathbf{Pr}\left[|\boldsymbol{r}^2 - n| \geq tn\right] \leq e^{-(3/16)nt^2} \quad \text{for all } t \in [0,1/2).$$

It follows that for $\boldsymbol{r} \sim \chi(n)$, $\mathbf{Pr}\left[\sqrt{n/2} \leq \boldsymbol{r} \leq \sqrt{3n/2}\right] \geq 1 - e^{-\frac{3n}{64}}$.

The following fact about the anti-concentration of the chi distribution will be useful:

Fact 9 *For $n > 1$, the maximum value of the pdf of the chi distribution $\chi(n)$ is at most 1, and hence for any interval $I = [a, b]$ we have $\mathbf{Pr}_{\boldsymbol{r}^2 \sim \chi^2(n)}[\boldsymbol{r} \in [a, b]] \leq b - a$.*

3. A density increment result for convex sets with positive inradius

In this section we establish our density increment result, Theorem 12. We note that related results of various types can be found in the literature (see e.g. Latala and Oleszkiewicz (1999), which proved the "S-conjecture" due to Shepp, and Theorem 3 of Latala and Oleszkiewicz (2005)), including folklore results such as Lemma 4.4 and Corollary 4.5 of Lovász and Vempala (2007). We were unable to find the exact statement we require in the literature and so we prove it here.

The key technical ingredient we use to prove Theorem 12 is the following consequence of the spherical isoperimetry theorem. (In the lemma below, for A a measurable subset of $\mathbb{S}^{n-1}$ and $\delta > 0$, we define $A_{\delta,\mathsf{Euc}} \subseteq \mathbb{S}^{n-1}$ to be the set of all points of $\mathbb{S}^{n-1}$ at Euclidean distance at most δ from A.)

Lemma 10 *There is an absolute constant $c > 0$ such that for every sufficiently large n, the following holds: Let A be a measurable subset of $\mathbb{S}^{n-1}$. As $\delta \to 0^+$ (independent of n), we have that*

$$\mu(A_{\delta,\mathsf{Euc}} \setminus A) \geq \begin{cases} c\delta\mu(A) & \text{if } \mu(A) \leq 1/2 \\ c\delta(1-\mu(A)) & \text{if } \mu(A) > 1/2 \\ c\delta\sqrt{n}\cdot\mu(A) & \text{if } e^{-n/4} \leq \mu(A) \leq 1/2 \\ c\delta\sqrt{n}\cdot(1-\mu(A)) & \text{if } 1/2 \leq \mu(A) \leq 1-e^{-n/4}. \end{cases} \tag{2}$$

Proof We first note that since the geodesic distance on $\mathbb{S}^{n-1}$ dominates the Euclidean distance, we have that $A_{\delta,\mathsf{geo}} \subseteq A_{\delta,\mathsf{Euc}}$. Thus, to prove Lemma 10, it suffices to lower bound $\mu(A_{\delta,\mathsf{geo}} \setminus A)$.

Let α^* be chosen so that the spherical cap $H = \{v \in \mathbb{S}^{n-1} : \langle v, e_1\rangle \geq \alpha^*\}$ centered around $e_1 \in \mathbb{S}^{n-1}$ has $\mu(H) = \mu(A)$, i.e. $\mathbf{Pr}_{\boldsymbol{v}\in\mathbb{S}^{n-1}}[\langle \boldsymbol{v}, e_1\rangle \geq \alpha^*] = \mu(A)$. Observe that the set $H_{\delta,\mathsf{geo}}$ is given by

$$H_{\delta,\mathsf{geo}} = \{v \in \mathbb{S}^{n-1} : \langle v, e_1\rangle \geq \beta^*\}, \quad \text{where} \quad \arccos(\beta^*) = \arccos(\alpha^*) + \delta.$$

Let us define $\varepsilon = \varepsilon(\delta)$ to be the value such that $\beta^* = \alpha^* - \varepsilon$. Note that $\varepsilon \to 0$ as $\delta \to 0$. It is well-known (see e.g. Baum (1990)) that for any $c \in [0, 1]$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{v}\in\mathbb{S}^{n-1}}[\langle \boldsymbol{v}, e_1\rangle \geq c] = \frac{A_{n-2}}{A_{n-1}} \int_{z=c}^{1} (1-z^2)^{\frac{n-3}{2}} dz, \tag{3}$$

where A_{n-1} is the surface area of $\mathbb{S}^{n-1}$ and $A_{n-2}/A_{n-1} = \Theta(\sqrt{n})$. This equation (3) implies that for $\varepsilon \to 0$,

$$\mu(H_{\delta,\mathsf{geo}} \setminus H) = \frac{A_{n-2}}{A_{n-1}} \cdot (1-\alpha^{*2})^{\frac{n-3}{2}} \cdot \varepsilon.$$

A simple calculus argument shows that for $\delta \to 0$, we have $\varepsilon = \delta \cdot \sqrt{1-\alpha^{*2}}$. Consequently, for $\delta \to 0$, we have that

$$\mu(H_{\delta,\mathsf{geo}} \setminus H) = \frac{A_{n-2}}{A_{n-1}} \cdot \delta \cdot (1-\alpha^{*2})^{\frac{n-2}{2}}. \tag{4}$$

For the rest of the proof we will assume that $\alpha^* \geq 0$ (an entirely similar argument handles the complementary case when $\alpha^* < 0$). We first establish a simple lower bound on $\mu(H_{\delta,\mathsf{geo}} \setminus H)$: to do this, we observe that

$$\begin{aligned}\mu(H) = \mathop{\mathbf{Pr}}_{\boldsymbol{v}\sim\mathbb{S}^{n-1}}[\langle \boldsymbol{v}, e_1\rangle \geq \alpha^*] = \frac{A_{n-2}}{A_{n-1}}\int_{z=\alpha^*}^{1}(1-z^2)^{\frac{n-3}{2}}dz &\leq \frac{A_{n-2}}{A_{n-1}}(1-\alpha^{*2})^{\frac{n-3}{2}}\cdot(1-\alpha^*)\\ &\leq \frac{A_{n-2}}{A_{n-1}}(1-\alpha^{*2})^{\frac{n-2}{2}}. \qquad (5)\end{aligned}$$

Plugging (5) into (4), we get $\mu(H_{\delta,\mathsf{geo}} \setminus H) \geq \delta\mu(H)$, giving the first two bounds of Equation (2).

We now establish the last two lines of Equation (2), which give a much better bound when $\mu(A) = \mu(H)$ is not too close to 0 or 1. Let us assume that $e^{-n/4} \leq \mu(H) \leq 1/2$. We recall the following bound on the surface area of a spherical cap (see e.g. Lemma 2.2 of Ball (1997)):

Fact 11 *For $\beta \in [0,1]$, define the spherical cap F_β to be $F_\beta = \{v \in \mathbb{S}^{n-1} : \langle v, e_1\rangle \geq \beta\}$. Then $\mu(F) \leq e^{-n\beta^2/2}$.*

Fact 11 and the assumption on $\mu(H)$ imply that $\alpha^* \leq 1/\sqrt{2}$. Let us now define J to be the largest integer such that $\alpha^* + J/\sqrt{n} \leq 1$ (note that $J = \Theta(\sqrt{n})$). We have that

$$\begin{aligned}\int_{z=\alpha^*}^{1}(1-z^2)^{\frac{n-3}{2}}dz &\leq \sum_{j=0}^{J}\frac{1}{\sqrt{n}}\cdot\left(1-\left(\alpha^*+\frac{j}{\sqrt{n}}\right)^2\right)^{\frac{n-3}{2}} \leq \sum_{j=0}^{J}\frac{1}{\sqrt{n}}\cdot\left(1-\alpha^{*2}-\frac{j^2}{n}\right)^{\frac{n-3}{2}}\\ &\leq \sum_{j=0}^{J}\frac{1}{\sqrt{n}}\cdot(1-\alpha^{*2})^{\frac{n-3}{2}}\cdot\left(1-\frac{j^2}{n}\right)^{\frac{n-3}{2}}\\ &\leq \sum_{j=0}^{J}\frac{1}{\sqrt{n}}\cdot(1-\alpha^{*2})^{\frac{n-3}{2}}\cdot e^{-j^2/4} = \frac{O(1)}{\sqrt{n}}\cdot(1-\alpha^{*2})^{\frac{n-3}{2}}, \qquad (6)\end{aligned}$$

where the last inequality uses that n is sufficiently large. From (3) and (6), we get that $\mu(H) \leq \frac{A_{n-2}}{A_{n-1}}\cdot\frac{O(1)}{\sqrt{n}}\cdot(1-\alpha^{*2})^{\frac{n-3}{2}}$. Combining with (4) and recalling that $\alpha^* \leq 1/\sqrt{2}$, we have $\mu(H_{\delta,\mathsf{geo}} \setminus H) \geq O(1)\cdot\delta\sqrt{n}\cdot\mu(H)$. Finally, by the spherical isoperimetric theorem Theorem 7, we have

$$\mu(A_{\delta,\mathsf{Euc}} \setminus A) \geq \mu(A_{\delta,\mathsf{geo}} \setminus A) \geq \mu(H_{\delta,\mathsf{geo}} \setminus H) \geq O(1)\cdot\delta\sqrt{n}\cdot\mu(A).$$

This finishes the proof of Lemma 10. ■

With Lemma 10 in hand the desired density increment result is easily obtained:

Theorem 12 (Density increment for convex sets with positive inradius.) *Let $K \subset \mathbb{R}^n$ be a convex set that has inradius $r_{\mathsf{in}} > 0$. Then for $r > r_{\mathsf{in}}$ and $\Delta r \to 0^+$, we have*

$$\alpha_K(r-\Delta r)-\alpha_K(r) \geq \begin{cases}\Omega\left(\frac{r_{\mathsf{in}}\sqrt{n}\Delta r}{r^2}\right)\alpha_K(r)(1-\alpha_K(r)) & \text{if } \min\{\alpha_K(r), 1-\alpha_K(r)\} \geq e^{-n/4},\\ \Omega\left(\frac{r_{\mathsf{in}}\Delta r}{r^2}\right)\alpha_K(r)(1-\alpha_K(r)) & \text{otherwise.}\end{cases}$$

Proof Set $\varepsilon := \Delta r/r$, and for any real number $a > 0$ and convex set K, define $aK := \{ax : x \in K\}$. Let K' denote $(1-\varepsilon)^{-1}K$,and observe that

$$\alpha_K(r - \Delta r) = \alpha_{K'}(r).$$

Now, observe that as $\varepsilon \to 0$, K' approaches the convex set $(1+\varepsilon)K$ (and always contains it). By our inradius assumption we have that $K + B(0, \varepsilon \cdot r_{\mathsf{in}}) \subseteq (1+\varepsilon)K$, so we get that

$$\alpha_K(r - \Delta r) = \alpha_{K'}(r) \geq \alpha_{(1+\varepsilon)K}(r) \geq \alpha_{K+B(0,\varepsilon \cdot r_{\mathsf{in}})}(r).$$

Scaling the results of Lemma 10 to the ball of radius r, the theorem is proved. ■

4. A weak learner for convex sets with large inradius

In this section we formally state and prove Theorem 3. For $t \in (0,1)$ define the function $h_t : \mathbb{R}^n \to \{-1, 1\}$ as $h_t(x) = 1$ if x is in the origin-centered closed ball of Gaussian volume t (and define h_0 to be the constant -1 function and h_1 to be the constant 1 function).

Theorem 3 *(Structural result, formal statement) Fix any constant $r_{\mathsf{in}} > 0$. If $K \subseteq \mathbb{R}^n$ is a convex set with inradius at least r_{in}, then there is some $h \in \{h_0, h_{1/2}, h_1\}$ such that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[h(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Omega(n^{-\frac{1}{2}}). \tag{7}$$

Intuition. Before entering into the detailed analysis of Theorem 3, we give an informal overview of the high level idea. First off, we can assume that the set K is close to being balanced, i.e.,

$$\left| \mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[K(\boldsymbol{g}) = 1] - \frac{1}{2} \right| \leq \frac{1}{\sqrt{n}}, \tag{8}$$

because otherwise either $h = h_0$ or $h = h_1$ satisfies (7).

For the purpose of this intuitive explanation, let us assume that there is a value $r_{1/2}$ such that $\alpha_K(r_{1/2}) = 1/2$.[4] We first argue at a high level why $h_{\mathsf{med}}(x) := \mathrm{sign}\left((r_{1/2})^2 - \sum_{i=1}^n x_i^2\right)$, i.e. the $\{-1,1\}$-valued indicator function of the origin-centered ball of radius $r_{1/2}$, must have some non-negligible correlation with K and can serve as a weak hypothesis.

To see this, we first argue that the advantage of h_{med} is at least non-negative. To see this, first observe that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[K(\boldsymbol{g}) = h_{\mathsf{med}}(\boldsymbol{g})] = \mathop{\mathbf{E}}_{\boldsymbol{r}^2 \sim \chi^2(n)} \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{S}_{\boldsymbol{r}}^{n-1}}[K(\boldsymbol{x}) = h_{\mathsf{med}}(\boldsymbol{x})],$$

and next observe that for each $r > 0$, by the choice of $r_{1/2}$ and the definition of h_{med}, we have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{S}_r^{n-1}}[K(\boldsymbol{x}) = h_{\mathsf{med}}(\boldsymbol{x})] = \begin{cases} \alpha_K(r) & \text{if } r < r_{1/2} \\ 1 - \alpha_K(r) & \text{if } r \geq r_{1/2}, \end{cases}$$

4. In general the function $\alpha_K(\cdot)$ need not be continuous, but it can be made continuous by perturbing K by an arbitrarily small amount, so this is essentially without loss of generality.

which is at least $1/2$ in each case by Fact 5.

Extending this simple reasoning, it is easy to see that if we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{r}^2 \sim \chi^2(n)}[|\overbrace{\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim \mathbb{S}^{n-1}_{\boldsymbol{r}}}[K(\boldsymbol{x}) = 1]}^{=\alpha_K(\boldsymbol{r})} - 1/2| \geq \beta] \geq \gamma, \tag{9}$$

for some $\beta, \gamma > 0$, then h_{med} is a weak hypothesis for K with advantage $\Omega(\gamma\beta)$. Putting it another way, the only way that h_{med} could fail to be a weak hypothesis with non-negligible advantage would be if the function $\alpha_K(\cdot)$ "stayed close to $1/2$" for a "wide range of values around $r_{1/2}$" — but this sort of behavior of $\alpha_K(\cdot)$ is precisely what is ruled out by our density increment result, Theorem 12.

Finally, to establish Theorem 3 we must show that $h_{1/2}$ (rather than h_{med}) has advantage $\Omega(n^{-1/2})$. This can be handled by a slight modification of the above argument that exploits (8).

4.1. Proof of Theorem 3

Let r_{median} denote the median of the $\chi(n)$ distribution. Define the function $r : [0, 1) \to [0, \infty)$ by

$$\mathop{\mathbf{Pr}}_{\boldsymbol{r} \sim \chi(n)}[\boldsymbol{r} \leq r(c)] = c.$$

Observe that since the pdf of $\chi^2(n)$ is always positive, the function $r(c)$ is well-defined, and that we have $r(1/2) = r_{\mathrm{median}}$. Theorem 8 and Theorem 9 together easily yield the following claim:

Claim 13 *The value r_{median} satisfies $|r_{\mathrm{median}} - \sqrt{n}| = O(1)$.*[5] *Further, there exist positive constants $A,\ B \geq 1/4$ such that $r(1/4) = r_{\mathrm{median}} - A$ and $r(3/4) = r_{\mathrm{median}} + B$.*

We now state the two main lemmas, Lemma 14 and Lemma 15. Theorem 3 is an immediate consequence of these two lemmas. To state these lemmas, let us set $c := 1/40$ (the precise value is not important as long as it is positive and sufficiently small).

Lemma 14 *If $|\mathrm{vol}(K) - 1/2| > c \cdot n^{-1/2}$, then either $h = h_0$ or $h = h_1$ achieves*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[h(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Theta(n^{-1/2}).$$

Lemma 15 *If $|\mathrm{vol}(K) - 1/2| \leq c \cdot n^{-1/2}$, then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[h_{1/2}(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Theta(n^{-1/2}).$$

Lemma 14 is immediate, so it remains to prove Lemma 15.

Proof of Theorem 15. We begin by defining the function $\beta : [0, 1) \to [0, 1)$ as follows:

$$\beta(c) := \mathop{\mathbf{Pr}}_{\boldsymbol{x} \in \mathbb{S}^{n-1}_{r(c)}}[\boldsymbol{x} \in K] = \alpha_K(r(c)).$$

5. In fact it is known that $r_{\mathrm{median}} \approx \sqrt{n} \cdot (1 - \frac{2}{9n})^{3/2}$, though we will not need this more precise bound.

Fact 16 *If K is a convex set that contains the origin, then $\beta(\cdot)$ is a non-increasing function.*

Proof This holds since $r(\cdot)$ is strictly increasing and the function $\alpha_K(\cdot)$ is non-increasing when $0^n \in K$ (Theorem 5). ■

Next, we have the following basic claim.

Claim 17 *For convex set K and $\beta(\cdot)$ as defined above, we have*

$$\int_{x\in[0,1)} \beta(x)dx = \mathrm{vol}(K).$$

Proof Let $\chi(n,b)$ denote the pdf of the χ-distribution with n degrees of freedom at b. Then

$$\mathrm{vol}(K) = \mathop{\mathbf{E}}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g})] = \int_{b=0}^{\infty} \chi(n,b)\alpha_K(b)db.$$

Substituting b by $r(\nu)$ (as ν ranges from 0 to 1), we have

$$\mathrm{vol}(K) = \int_{\nu=0}^{1} \chi(n,r(\nu))r'(\nu)\beta(\nu)d\nu. \tag{10}$$

Finally, by definition of $r(\nu)$, we have that $\int_{z=0}^{r(\nu)} \chi(n,z)dz = \nu$. Taking the derivative of this with respect to ν, we get that $\chi(n,r(\nu))r'(\nu) = 1$, and substituting back into (10), we get the claim. ■

Now we are ready to analyze $h_{1/2}$. The following claim says that if $\beta(1/4)$ is "somewhat large", then $h_{1/2}$ is a weak hypothesis with constant advantage:

Claim 18 *If $\beta(1/4) \geq \frac{3}{4}$ then $\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[h_{1/2}(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \frac{1}{24}$.*

Proof Define $s = \int_{x=0}^{1/4} \beta(x)dx$ and $t = \int_{x=1/4}^{1/2} \beta(x)dx$. Using the fact that $\beta(\cdot)$ is non-increasing we have

$$\text{(i) } s = \int_{x=0}^{1/4} \beta(x)dx \geq \frac{3}{4}\cdot\frac{1}{4} = \frac{3}{16}, \quad \text{(ii) } t = \int_{x=1/4}^{1/2} \beta(x)dx \geq \frac{1}{3}\cdot\left(\int_{x=1/4}^{1} \beta(x)dx\right) = \frac{\mathrm{vol}(K)-s}{3} \tag{11}$$

(where Theorem 17 was used for the last inequality of (ii)). We thus get

$$\begin{aligned}\int_{x=0}^{1/2} \beta(x)dx - \int_{x=1/2}^{1} \beta(x)dx &= 2\int_{x=0}^{1/2} \beta(x)dx - \int_{x=0}^{1} \beta(x)dx \\ &= 2s + 2t - \mathrm{vol}(K) \geq \frac{4s}{3} - \frac{\mathrm{vol}(K)}{3} \geq \frac{1}{24}, \end{aligned} \tag{12}$$

where the first inequality above follows by item (ii) of (11) and the second inequality uses item (i) of (11) along with the hypothesis $|\mathrm{vol}(K) - 1/2| \leq c/\sqrt{n} \leq 1/40$. Combining these bounds, we have

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\boldsymbol{g}\in N(0,1)^n}[h_{1/2}(\boldsymbol{g}) = K(\boldsymbol{g})] &= \int_{x=0}^{1/2} \beta(x)dx + \int_{x=1/2}^{1} (1-\beta(x))dx \\ &\geq \frac{1}{2} + \int_{x=0}^{1/2} \beta(x)dx - \int_{x=1/2}^{1} \beta(x)dx \geq \frac{1}{2} + \frac{1}{24}, \qquad \text{(by (12))}\end{aligned}$$

and Theorem 18 is proved. ■

Thus to prove Theorem 15, it remains to consider the case that $\beta(1/4) \leq 3/4$. By the monotonicity of $\beta(\cdot)$, Claim 17, and the hypothesis of Theorem 15, we have that

$$\frac{1}{2} - \frac{1}{40} \leq \int_{x=0}^{1} \beta(x)dx \leq \frac{1}{4} + \frac{3}{4} \cdot \beta(1/4).$$

and hence $\beta(1/4) \geq 3/10$, so we subsequently assume that $3/10 \leq \beta(1/4) \leq 3/4$. Now, recall that $r(1/4) = r_{\mathrm{median}} - A$ and $r(3/4) = r_{\mathrm{median}} + B$, where $1/4 \leq A$ and $r_{\mathrm{median}} = \sqrt{n} \pm O(1)$ by Theorem 13. Thus

$$\beta(1/4) = \alpha_K(r_{\mathrm{median}} - A) \quad \text{and} \quad \beta(3/4) = \alpha_K(r_{\mathrm{median}} + B) \quad \text{where } A, B \geq 1/4. \tag{13}$$

Using the fact that $3/10 \leq \beta(1/4) \leq 3/4$, Equation (13), by Theorem 12 we get that $\beta(1/4) \geq \beta(3/4) + \Omega(r_{\mathsf{in}} \cdot n^{-1/2})$. As $r_{\mathsf{in}} > 0$ is an absolute constant, we get that

$$\beta(1/4) \geq \beta(3/4) + C \cdot n^{-1/2} \tag{14}$$

for an absolute constant $C > 0$. This implies that

$$\begin{aligned}
&\int_{x=0}^{1/2} \beta(x)dx - \int_{x=1/2}^{1} \beta(x)dx \\
= &\int_{x=0}^{1/4} \beta(x)dx - \int_{x=3/4}^{1} \beta(x)dx + \int_{x=1/4}^{1/2} \beta(x)dx - \int_{x=1/2}^{3/4} \beta(x)dx \\
\geq &\frac{C}{4\sqrt{n}} + \int_{x=1/4}^{1/2} \beta(x)dx - \int_{x=1/2}^{3/4} \beta(x)dx \geq \frac{C}{4\sqrt{n}},
\end{aligned} \tag{15}$$

where both inequalities use the monotonicity of $\beta(\cdot)$ and the penultimate inequality additionally uses Equation (14). Applying (15), we get

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\boldsymbol{g} \in N(0,1)^n}[h_{1/2}(\boldsymbol{g}) = K(\boldsymbol{g})] &= \int_{x=0}^{1/2} \beta(x)dx + \int_{x=1/2}^{1} (1 - \beta(x))dx \\
&= \frac{1}{2} + \int_{x=0}^{1/2} \beta(x)dx - \int_{x=1/2}^{1} \beta(x)dx \\
&\geq \frac{1}{2} + \frac{C}{4\sqrt{n}}.
\end{aligned} \tag{16}$$

This proves Theorem 15.

5. A weak learner for general convex sets

In this section we prove Theorem 1. The high level proof strategy is as follows: Note that Theorem 3 gives a weak learner for convex sets K that have $B(0, r_{\mathsf{in}}) \subseteq K$ (for any positive constant r_{in}). Thus, to get a weak learner for general convex sets, it suffices to consider the case when for some positive constant $r_{\mathsf{in}} > 0$, $B(0, r_{\mathsf{in}}) \not\subseteq K$. In particular, we show that there is a positive constant $\zeta > 0$ such that if $B(0, \zeta) \not\subseteq K$ (and K is close to being balanced), then there is an efficient weak learner for K (with constant advantage). Formally, we prove the following theorem.

Theorem 19 *There is a fixed positive constant $\zeta > 0$ and a $\mathrm{poly}(n)$ time algorithm* Learn-halfspace *with the following guarantee: Suppose $K \subseteq \mathbb{R}^n$ is a convex set satisfying:*

1. $B(0,\zeta) \not\subseteq K$;
2. $|\mathrm{vol}(K) - 1/2| \leq \zeta \cdot n^{-1/2}$.

Then, given random labeled samples of the form $(\boldsymbol{g}, K(\boldsymbol{g}))$, the algorithm Learn-halfspace *outputs a halfspace $h_\ell : \mathbb{R}^n \to \{-1,1\}$ such that*

$$\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[h_\ell(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{7}{8}.$$

Before proving Theorem 19, let us first see how it implies Theorem 1.

Proof of Theorem 1 using Theorem 19. For convex set K and positive constant ζ (from Theorem 19),

1. If $|\mathrm{vol}(K) - 1/2| > \zeta \cdot n^{-1/2}$, then using Lemma 14, there is an $h \in \{h_0, h_1\}$ such that
$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[h(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Omega(n^{-1/2}).$$
2. If $B(0,\zeta) \subseteq K$, then using Theorem 3, it follows that for some $h \in \{h_0, h_{1/2}, h_1\}$,
$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[h(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Omega(n^{-1/2}).$$
3. Finally, if $B(0,\zeta) \not\subseteq K$ and $|\mathrm{vol}(K) - 1/2| \leq \zeta \cdot n^{-1/2}$, then we can apply Theorem 19 to obtain a hypothesis h_ℓ such that
$$\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[h_\ell(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{7}{8}.$$

Thus, it follows that for any convex set K, there is a $h \in \{h_0, h_{1/2}, h_1, h_\ell\}$ which satisfies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[h(\boldsymbol{g}) = K(\boldsymbol{g})] \geq \frac{1}{2} + \Omega(n^{-1/2}). \tag{17}$$

Note that the complexity of computing h_ℓ is $\mathrm{poly}(n)$ (both samples and running time). Further, we can identify the "right" h (i.e., an element of the set $\{h_0, h_{1/2}, h_1, h_\ell\}$ satisfying (17)) using a simple hypothesis testing routine using $\mathrm{poly}(n)$ samples and running time. This finishes the proof of Theorem 1.

It remains to prove Theorem 19. To prove this theorem, we essentially use a so-called "agnostic learner" for halfspaces. Several results in the literature, including Kalai et al. (2008); De et al. (2014); Awasthi et al. (2013); Diakonikolas et al. (2018), suffice for our purposes. For the sake of concreteness, we use the following result from Diakonikolas et al. (2018).

Theorem 20 (Theorem 1.2 from Diakonikolas et al. (2018), taking "$d = 1$") *There is an algorithm* Learn-halfspace *with the following guarantee: Let $f : \mathbb{R}^n \to \{-1,1\}$ be a target halfspace such that the algorithm gets access to samples of the form $(\boldsymbol{g}, \Psi(\boldsymbol{g}))$ where $\boldsymbol{g} \sim N(0,1)^n$ and $\Psi : \mathbb{R}^n \to \{-1,1\}$ satisfies $\mathbf{Pr}_{\boldsymbol{g}}[\Psi(\boldsymbol{g}) \neq f(\boldsymbol{g})] \leq \delta$. Then* Learn-halfspace *runs in time $\mathrm{poly}(n, 1/\delta)$ and outputs a hypothesis halfspace $h_\ell : \mathbb{R}^n \to \{-1,1\}$ such that $\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[f(\boldsymbol{g}) \neq h_\ell(\boldsymbol{g})] \leq \delta^B$, where $B > 0$ is an absolute constant.*

Proof of Theorem 1. Let us set ζ so that $(4\zeta) + (4\zeta)^B < 1/8$ where B is the constant appearing in Theorem 20. We run algorithm Learn-halfspace from Theorem 20 with samples $(\boldsymbol{g}, K(\boldsymbol{g}))$ and $\delta := \zeta/4$. We show that the output h_ℓ satisfies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g}) = h_\ell(\boldsymbol{g})] \geq \frac{7}{8}. \tag{18}$$

Towards this, first observe that since $B(0,\zeta) \not\subseteq K$, there must be a point z^* such that $\|z^*\|_2 \leq \zeta$ and $z^* \notin K$. Using the supporting hyperplane theorem (see e.g. page 510 in Luenberger and Ye), it follows that there is a unit vector $\widehat{v} \in \mathbb{S}^{n-1}$ such that the halfspace defined as

$$f(x) = \mathsf{sign}(\widehat{v} \cdot x + \zeta),$$

satisfies $K \subseteq f^{-1}(1)$. Using the fact that the pdf of an $N(0,1)$ Gaussian is everywhere at most 1, we get that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[f(\boldsymbol{g}) = 1] \leq \frac{1}{2} + \zeta.$$

This implies that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g}) = 1 | f(\boldsymbol{g}) = 1] \geq \frac{\frac{1}{2} - \frac{\zeta}{\sqrt{n}}}{\frac{1}{2} + \zeta} \geq 1 - 2\zeta - \frac{2\zeta}{\sqrt{n}}. \tag{19}$$

On the other hand, by construction of $f(\cdot)$, we have that $\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g}) = -1 | f(\boldsymbol{g}) = -1] = 1$. Combining this with (19), we get that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g}) = f(\boldsymbol{g})] \geq 1 - 2\zeta - \frac{2\zeta}{\sqrt{n}} \geq 1 - 4\zeta. \tag{20}$$

If we run the algorithm Learn-halfspace on samples of the form $(\boldsymbol{g}, K(\boldsymbol{g}))$ (where $\delta := 4\zeta$), then by Theorem 20 the output h_ℓ satisfies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[f(\boldsymbol{g}) \neq h_\ell(\boldsymbol{g})] \leq (4\zeta)^B.$$

Combining this with (20), we get

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g}\sim N(0,1)^n}[K(\boldsymbol{g}) = h_\ell(\boldsymbol{g})] \geq 1 - 4\zeta - (4\zeta)^B > \frac{7}{8}.$$

The last inequality follows by our choice of ζ. This finishes the proof of Theorem 1.

Acknowledgments

We thank the anonymous reviewer of an earlier version of this paper De and Servedio (2019). A. D. is supported by NSF CCF-1926872, CCF-1910534 and CCF-2045128. R. A. S. is supported by the Simons Collaboration on Algorithms and Geometry, NSF CCF-1563155, CCF-1814873, and IIS-1838154.

References

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with malicious noise. *CoRR*, abs/1307.8371, 2013.

Keith Ball. *An Elementary Introduction to Modern Convex Geometry*, volume 31. MSRI Publications (Flavors of Geometry), 1997.

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

Aleksandrs Belovs and Eric Blais. A polynomial lower bound for testing monotonicity. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1021–1032. ACM, 2016.

A. Blum, C. Burch, and J. Langford. On learning monotone boolean functions. In *Proceedings of the Thirty-Ninth Annual Symposium on Foundations of Computer Science*, pages 408–415, 1998.

Béla Bollobás and Andrew Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987.

Nader Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

Deeparnab Chakrabarty and Comandur Seshadhri. An $o(n)$ Monotonicity Tester for Boolean Functions over the Hypercube. *SIAM Journal on Computing*, 45(2):461–472, 2016.

Xi Chen, Anindya De, Rocco Servedio, and Li-Yang Tan. Boolean function monotonicity testing requires (almost) $n^{1/2}$ non-adaptive queries. In *Proceedings of the 47th Annual Symposium on Theory of Computing (STOC 2015)*, pages 519–528, 2015.

Xi Chen, Adam Freilich, Rocco A Servedio, and Timothy Sun. Sample-Based High-Dimensional Convexity Testing. In *APPROX/RANDOM 2017*, 2017a.

Xi Chen, Erik Waingarten, and Jinyu Xie. Beyond talagrand functions: New lower bounds for testing monotonicity and unateness. pages 523–536, 2017b.

M. Chiani, D. Dardari, and M. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*, 2 (4):840–845, 2003.

D. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

A. De and R. Servedio. Kruskal-Katona for convex sets, with applications. Available at https://arxiv.org/abs/1911.00178, 2019.

Anindya De, Ilias Diakonikolas, Vitaly Feldman, and Rocco A. Servedio. Nearly optimal solutions for the chow parameters problem and low-weight approximation of halfspaces. *J. ACM*, 61(2): 11:1–11:36, 2014.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073, 2018.

R. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky. Monotonicity testing over general poset domains. In *Proc. 34th Annual ACM Symposium on the Theory of Computing*, pages 474–483, 2002.

T. Harris. A lower bound for the critical probability in a certain percolation process. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 56, pages 13–20, 1960.

Svante Janson. Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135:1–6, 2018.

Iain M. Johnstone. Chi-square oracle inequalities. In *State of the art in probability and statistics*, pages 399–418. Institute of Mathematical Statistics, 2001.

Adam Kalai, Adam Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

G. Katona. A theorem of finite sets. 1968.

M. Kearns, M. Li, and L. Valiant. Learning Boolean formulas. *Journal of the ACM*, 41(6):1298–1328, 1994.

Subhash Khot, Dor Minzer, and Muli Safra. On Monotonicity Testing and Boolean Isoperimetric type Theorems. In *Proceedings of the 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 52–58, 2015.

D. Kleitman. Families of non-disjoint subsets. *Journal of Combinatorial Theory*, 1(1):153–155, 1966.

Adam Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via Gaussian surface area. In *Proceedings of the 49th Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008.

J. Kruskal. The number of simplices in a complex. In R. Bellman, editor, *Mathematical Optimization Techniques*, pages 251–278. University of California, Press, Berkeley, 1963.

Günter Last and Mathew Penrose. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2017.

R. Latala and K. Oleszkiewicz. Gaussian measures of dilatations of convex symmetric sets. *Annals of Probability.*, 27:1922–1938, 1999.

R. Latala and K. Oleszkiewicz. Small ball probability estimate in terms of width. *Studia Math.*, 169:305–314, 2005.

M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2001.

P. Lévy. *Problèmes Concrets D'analyse Fonctionnelle*. Gauthier-Villars, 1951.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

László Lovász. *Combinatorial problems and exercises*, volume 361. American Mathematical Soc., 1981.

D. Luenberger and Y. Ye. *Linear and nonlinear programming*, volume 2. Springer.

R. O'Donnell and K. Wimmer. KKL, Kruskal-Katona, and monotone nets. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2009.

R. Raz. Exponential separation of quantum and classical communication complexity. In *Conference Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 358–367. Association for Computing Machinery (ACM), 1999.

Thomas Royen. A simple proof of the gaussian correlation conjecture extended to multivariate gamma distributions. *Far East Journal of Theoretical Statistics*, pages 139–145, 2014.

Appendix A. A lower bound for weak learning symmetric convex sets

In this section we prove Theorem 2, which we restate here for the convenience of the reader:

Theorem 2 *For sufficiently large n, for any $s \geq n$, there is a distribution $\mathcal{D}_{\mathrm{actual}}$ over centrally symmetric convex sets with the following property: for a target convex set $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{actual}}$, for any membership-query (black box query) algorithm A making at most s many queries to $\boldsymbol{f}$, the expected error of A (the probability over $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{actual}}$, over any internal randomness of A, and over a random Gaussian $\boldsymbol{x} \sim N(0,1^n)$, that the output hypothesis h of A predicts incorrectly on $\boldsymbol{x}$) is at least $1/2 - \frac{O(\log s)}{n^{1/2}}$.*

We note that this lower bound holds even in the membership query (hereafter abbreviated as MQ) model. In this model the learning algorithm has query access to a black-box oracle for the unknown target function $\boldsymbol{f}$; note that a learning algorithm in this model can simulate a learning algorithm in the model where the algorithm receives only random labeled examples of the form $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ (with $\boldsymbol{x} \sim N(0,1)^n$) with no overhead. Thus a lower bound in the MQ model holds *a fortiori* for the random examples model (which is the model that our algorithms use). In particular, by instantiating $s = \mathrm{poly}(n)$ in the above theorem, we get that no algorithm which receives $\mathrm{poly}(n)$ samples (and hence no algorithm running in $\mathrm{poly}(n)$ time) can achieve an advantage of $\frac{\omega(\log n)}{\sqrt{n}}$ over random guessing for learning centrally symmetric convex sets. Thus, our algorithm for weak learning of convex sets, i.e., Theorem 1, achieves an optimal advantage (up to an $O(\log n)$ factor).

Since the proof of Theorem 2 is somewhat involved we begin by explaining its general strategy:

1. We start by constructing a "hard" distribution $\mathcal{D}_{\text{ideal}}$ over centrally symmetric convex subsets of $\mathbb{R}^n$ (note that $\mathcal{D}_{\text{ideal}}$ is different from the final distribution $\mathcal{D}_{\text{actual}}$). We then analyze the case in which the learning algorithm is not allowed to make *any* queries to the target function $\boldsymbol{f} \sim \mathcal{D}_{\text{ideal}}$. It is easy to see that in this setting, the maximum possible accuracy of any zero-query learning algorithm for $\mathcal{D}_{\text{ideal}}$ is achieved by the so-called *Bayes optimal classifier* (which we denote by $BO_{\mathcal{D}_{\text{ideal}}}$) which labels each $x \in \mathbb{R}^n$ as follows:

$$BO_{\mathcal{D}_{\text{ideal}}}(x) = \begin{cases} 1 & \text{if } \mathbf{Pr}_{\boldsymbol{f} \sim \mathcal{D}_{\text{ideal}}}[\boldsymbol{f}(x) = 1] \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

   We show that for "most" $\boldsymbol{x}$ sampled from $N(0,1)^n$, the accuracy of $BO_{\mathcal{D}_{\text{ideal}}}(\boldsymbol{x})$ is close to $1/2$ and in fact, the average advantage over $1/2$ for $\boldsymbol{x} \sim N(0,1)^n$ is bounded by $\frac{O(\log s)}{n^{1/2}}$.

2. The distribution $\mathcal{D}_{\text{ideal}}$ is a continuous distribution defined in terms of a so-called *Poisson point process*. While the construction of $\mathcal{D}_{\text{ideal}}$ is particularly well-suited to the analysis of a zero-query learner, i.e. of the Bayes optimal classifier (indeed this is the motivation for our introducing $\mathcal{D}_{\text{ideal}}$), it becomes tricky to analyze $\mathcal{D}_{\text{ideal}}$ when the learning algorithm is actually allowed to make queries to the target function $\boldsymbol{f}$. To deal with this, we "discretize" the distribution $\mathcal{D}_{\text{ideal}}$ to construct the actual hard distribution $\mathcal{D}_{\text{actual}}$ (which is finitely supported). The discretization is carefully done to ensure that for "most" $\boldsymbol{x}$ (again sampled from $N(0,1)^n$), $\mathbf{Pr}_{\boldsymbol{f} \in \mathcal{D}_{\text{ideal}}}[\boldsymbol{f}(\boldsymbol{x}) = 1]$ is close to $\mathbf{Pr}_{\boldsymbol{f} \in \mathcal{D}_{\text{actual}}}[\boldsymbol{f}(\boldsymbol{x}) = 1]$. This implies that the average advantage of the Bayes optimal classifier for $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ (corresponding to the best possible zero-query learning algorithm), denoted by $BO_{\mathcal{D}_{\text{actual}}}$, remains bounded by $\frac{O(\log s)}{n^{1/2}}$.

3. Finally, we consider the case when the learning algorithm is allowed to makes s queries to the unknown target function $\boldsymbol{f}$. Roughly speaking, we show that for any choice of s query points $\overline{y} = (y_1, \ldots, y_s)$, with high probability over both $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ and $\boldsymbol{x} \sim N(0,1)^n$, the advantage of the optimal classifier is close to that achieved by $BO_{\mathcal{D}_{\text{actual}}}$ (see Appendix A.3). The techniques used to prove this crucially rely on the specific construction of $\mathcal{D}_{\text{actual}}$, so we refrain from giving further details here. However, using this and the upper bound on the advantage of $BO_{\mathcal{D}_{\text{actual}}}$, we obtain Theorem 2.

We note that step 3 and the general flavor of the analysis used to establish that step closely follows the lower bound approach of Blum, Burch and Langford Blum et al. (1998), who showed that no s-query algorithm in the MQ model can achieve an advantage of $\omega(\frac{\log s}{\sqrt{n}})$ over random guessing to learn monotone functions under the uniform distribution on $\{-1,1\}^n$. Of course, the choice of the *hard distribution* is quite different in our work than in Blum et al. (1998); in particular, a draw from $\mathcal{D}_{\text{ideal}}$ is essentially a random symmetric polytope with $\mathsf{poly}(s)$ facets where the hyperplane defining each facet is at distance around $O(\sqrt{\log s})$ away from the origin. The distribution $\mathcal{D}_{\text{actual}}$ is obtained by essentially discretizing $\mathcal{D}_{\text{ideal}}$ while retaining some crucial geometric properties. In contrast, the hard distribution in Blum et al. (1998) is constructed in one step and is essentially a random monotone DNF of width $O(\log s + \log n)$ with roughly s terms. Another significant difference between our argument and that of Blum et al. (1998) is the technical challenges that arise in our case because of dealing with a continuous domain and the resulting discretization that we have to perform.

Finally, we note that in the proof of Theorem 2, which we give below, we may assume that $s = 2^{O(\sqrt{n})}$, since otherwise the claimed bound trivially holds.

A.1. The idealized distribution $\mathcal{D}_{\text{ideal}}$ and the Bayes optimal classifier for it

We will define the distribution $\mathcal{D}_{\text{actual}}$ by first defining a related distribution $\mathcal{D}_{\text{ideal}}$. As mentioned earlier, the distribution $\mathcal{D}_{\text{actual}}$ will be obtained by discretization of $\mathcal{D}_{\text{ideal}}$. To define $\mathcal{D}_{\text{ideal}}$, we need to recall the notion of a spatial Poisson point process; we specialize this notion to the unit sphere $\mathbb{S}^{n-1}$, though it is clear that an analogue of the definition we give below can be given over any bounded measurable set $B \subseteq \mathbb{R}^n$.

Definition 21 *A* point process $\boldsymbol{X}$ *on the* carrier space $\mathbb{S}^{n-1}$ *is a stochastic process such that a draw from this process is a sequence of points* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\boldsymbol{N}} \in \mathbb{S}^{n-1}$. *(Note that each individual point* $\boldsymbol{x}_i$ *as well as the number of points* $\boldsymbol{N}$ *are all random variables as described below.)*

A spatial Poisson point process with parameter λ on $\mathbb{S}^{n-1}$ *is a point process on* $\mathbb{S}^{n-1}$ *with the following two properties:*

1. *For any measurable subset* $B \subseteq \mathbb{S}^{n-1}$, *let* $\boldsymbol{N}(B)$ *denote the number of points which fall in* B. *Then, the distribution of* $\boldsymbol{N}(B)$ *follows* $\mathsf{Poi}(\lambda\mu(B))$ *where* $\mu(B)$ *is the fractional density of* B *inside* $\mathbb{S}^{n-1}$.
2. *If* $B_1, \ldots, B_k \subseteq \mathbb{S}^{n-1}$ *are pairwise disjoint measurable sets, then* $\boldsymbol{N}(B_1), \ldots, \boldsymbol{N}(B_k)$ *are mutually independent.*

Finally, we note that the spatial Poisson point process with parameter λ *can be realized as follows: Sample* $\boldsymbol{N} \sim \mathsf{Poi}(\lambda)$, *and output* $\boldsymbol{N}$ *points* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\boldsymbol{N}}$ *that are chosen uniformly and independently at random from* $\mathbb{S}^{n-1}$.

We refer the reader to Last and Penrose (2017) and Daley and Vere-Jones (2007) for details about Poisson point processes.

We next choose $d > 0$ so that for any unit vector v,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{u} \sim \mathbb{S}^{n-1}}\left[|v \cdot \boldsymbol{u}| \geq \frac{d}{\sqrt{n}}\right] = \frac{1}{s^{100}}. \tag{21}$$

Note that by symmetry the choice of v is immaterial. We also recall the following fundamental fact about inner products with random unit vectors (which is easy to establish using e.g. Equation (36)):

Claim 22 *Let* $v \in \mathbb{S}^{n-1}$. *For any* $0 < t < 1/2$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{u} \in \mathbb{S}^{n-1}}[|v \cdot \boldsymbol{u}| \geq t] = e^{-\Theta(t^2 n)}.$$

Since we have $s = 2^{O(\sqrt{n})}$, it follows from this fact that $d = \Theta(\sqrt{\log s})$ in (21). Next, for any unit vector $z \in \mathbb{S}^{n-1}$, we define the "slab" function $\mathrm{slab}_z : \mathbb{R}^n \to \{0, 1\}$,

$$\mathrm{slab}_z(x) \coloneqq \mathbb{1}\left[-d \leq z \cdot x \leq d\right].$$

It is clear that for any unit vector z the function $\mathrm{slab}_z(\cdot)$ defines a centrally symmetric convex set. Finally, we define the parameter Λ to be

$$\Lambda := s^{100} \cdot \ln 2. \tag{22}$$

Now we are ready to define the distribution $\mathcal{D}_{\text{ideal}}$. A function $\boldsymbol{f}$ is sampled from $\mathcal{D}_{\text{ideal}}$ as follows:

- Sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{\boldsymbol{N}}$ from the spatial Poisson point process on $\mathbb{S}^{n-1}$ with parameter Λ.
- Set $\boldsymbol{f}$ to be

$$\boldsymbol{f}(x) = \bigwedge_{i=1}^{\boldsymbol{N}} \mathrm{slab}_{\boldsymbol{z}_i}(x).$$

We have the following observation (whose proof is immediate from the construction):

Observation 23

1. *Any $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{ideal}}$ defines a centrally symmetric convex set.*
2. *For any point $x \in \mathbb{R}^n$, the value of $\mathcal{D}_{\mathrm{ideal}}(x) := \mathbf{Pr}_{\boldsymbol{f} \sim \mathcal{D}_{\mathrm{ideal}}}[\boldsymbol{f}(x) = 1]$ is completely determined by $\|x\|_2$, the distance of x from the origin.*

A.1.1. ANALYZING THE BAYES OPTIMAL CLASSIFIER FOR $\mathcal{D}_{\mathrm{ideal}}$

We now bound the advantage of the Bayes optimal classifier (denoted by $BO_{\mathcal{D}_{\mathrm{ideal}}}$) for $\mathcal{D}_{\mathrm{ideal}}$, which, as stated earlier, corresponds to the best possible learning algorithm that makes zero queries to the unknown target function $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{ideal}}$. Observe that on input $x \in \mathbb{R}^n$, the classifier $BO_{\mathcal{D}_{\mathrm{ideal}}}(x)$ outputs 1 if $\mathcal{D}_{\mathrm{ideal}}(x) \geq 1/2$ and outputs 0 on x if $\mathcal{D}_{\mathrm{ideal}}(x) < 1/2$. Thus, the expected error of $BO_{\mathcal{D}_{\mathrm{ideal}}}$ is

$$\mathrm{opt}(\mathcal{D}_{\mathrm{ideal}}) := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)^n}[\min\{\mathcal{D}_{\mathrm{ideal}}(\boldsymbol{x}), 1 - \mathcal{D}_{\mathrm{ideal}}(\boldsymbol{x})\}],$$

and the expected advantage of $BO_{\mathcal{D}_{\mathrm{ideal}}}$ is $1/2 - \mathrm{opt}(\mathcal{D}_{\mathrm{ideal}})$.

The next lemma bounds $\mathrm{opt}(\mathcal{D}_{\mathrm{ideal}})$ and completes Step 1 of the proof outline given earlier:

Lemma 24 *We have*

$$\frac{1}{2} - \mathrm{opt}(\mathcal{D}_{\mathrm{ideal}}) = \frac{O(\log s)}{\sqrt{n}}.$$

Proof For $c = 1, 2, \ldots, 4\sqrt{\ln n}$, define the set

$$S_c := \{x \in \mathbb{R}^n : |\|x\|_2^2 - n| \in [2(c-1)\sqrt{n}, 2c\sqrt{n})\}$$

and further define the set

$$\mathrm{S}_{\mathrm{extreme}} := \{x \in \mathbb{R}^n : |\|x\|_2^2 - n| > 8\sqrt{n \ln n}\}.$$

We observe that by Theorem 8, we have that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{g} \sim N(0,1)^n}[\boldsymbol{g} \in \mathrm{S}_{\mathrm{extreme}}] \leq \frac{1}{n^5}. \tag{23}$$

We will show that for each $c = 1, \ldots, 4\sqrt{\ln n}$ the value of $\mathcal{D}_{\mathrm{ideal}}(x)$ is "close" to $1/2$ for every $x \in \mathrm{S}_{\mathrm{med}}$ (in a quantitative sense that depends on c), and combining the resulting bounds will easily

yield the lemma. To do this, we define $\mathrm{Region}(x)$ to be the set of those unit vectors z such that x does *not* lie within the slab defined by z, i.e.

$$\mathrm{Region}(x) := \{z \in \mathbb{S}^{n-1} : |z \cdot x| > d\}.$$

Observe that the fractional density of $\mathrm{Region}(x)$ inside $\mathbb{S}^{n-1}$, which we denote by $\mu_1(\mathrm{Region}(x))$, is determined by $\|x\|_2$.

Fix a value of c; we would like to analyze $\mu_1(\mathrm{Region}(x))$ for all points $x \in \mathrm{S}_c$. To do this, we first analyze it for points at distance exactly $\sqrt{n}$ from the origin. So choose any point $a_0 \in \mathbb{R}^n$ such that $\|a_0\|_2 = \sqrt{n}$. By the definition of d in (21) and observing that $a_0/\sqrt{n}$ is a unit vector, we have

$$\mu_1(\mathrm{Region}(a_0)) = \mathop{\mathbf{Pr}}_{\boldsymbol{u}\sim\mathbb{S}^{n-1}}\left[\frac{a_0}{\sqrt{n}} \cdot \boldsymbol{u} \geq \frac{d}{\sqrt{n}}\right] = \frac{1}{s^{100}}. \tag{24}$$

Next, consider any $b_0 \in \mathrm{S}_c$, and note that $\|b_0\|_2 = \sqrt{n}(1+\delta)$ where $|\delta| = O(\sqrt{\frac{c}{n}})$. Hence

$$\mu_1(\mathrm{Region}(b_0)) = \mathop{\mathbf{Pr}}_{\boldsymbol{u}\sim\mathbb{S}^{n-1}}\left[\frac{b_0}{\sqrt{n}(1+\delta)} \cdot \boldsymbol{u} \geq \frac{d}{\sqrt{n}(1+\delta)}\right],$$

where $\frac{b_0}{\sqrt{n}(1+\delta)}$ is a unit vector. Recalling that we can assume $\log s \leq c_0\sqrt{n}$ for a sufficiently small positive constant $c_0 > 0$ and that $d = \Theta(\sqrt{\log s})$, we can apply Theorem 40 to get that

$$\left|\frac{\mu_1(\mathrm{Region}(a_0))}{\mu_1(\mathrm{Region}(b_0))} - 1\right| = O\left(d^2 \cdot \frac{c}{\sqrt{n}}\right) = O\left(\frac{c\log s}{\sqrt{n}}\right). \tag{25}$$

From (25) and (24), we get that every $x \in \mathrm{S}_{\mathrm{med}}$ satisfies

$$\mu_1(\mathrm{Region}(x)) = \frac{1}{s^{100}} \cdot \left(1 + O\left(\frac{c\log s}{\sqrt{n}}\right)\right). \tag{26}$$

To finish the proof, we observe that sampling $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{ideal}}$ is equivalent to sampling $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{\boldsymbol{N}}$ from the spatial Poisson point process on $\mathbb{S}^{n-1}$ with parameter Λ. Let $\mathbf{Num}_x$ be the random variable defined as $|\{\boldsymbol{z}_i\}_{i=1}^{\boldsymbol{N}} \cap \mathrm{Region}(x)|$. Observe that

1. $\boldsymbol{f}(x) = 1$ iff $\mathbf{Num}_x = 0$;
2. $\mathbf{Num}_x$ is distributed as $\mathsf{Poi}(\Lambda \cdot \mu_1(\mathrm{Region}(x)))$.

Putting these two items together with (26) and (22), we get that for $x \in \mathrm{S}_c$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{f}\sim\mathcal{D}_{\mathrm{ideal}}}[\boldsymbol{f}(x) = 1] = \mathbf{Pr}[\mathsf{Poi}(\Lambda \cdot \mu(\mathrm{Region}(x))) = 0] = e^{-\Lambda\cdot\mu_1(\mathrm{Region}(x))} = \frac{1}{2} \pm O\left(\frac{c\log s}{\sqrt{n}}\right),$$

so recalling that $\mathcal{D}_{\mathrm{ideal}}(x) = \mathbf{Pr}_{\boldsymbol{f}\sim\mathcal{D}_{\mathrm{ideal}}}[\boldsymbol{f}(x) = 1]$, we have that

$$\min\{\mathcal{D}_{\mathrm{ideal}}(x), 1 - \mathcal{D}_{\mathrm{ideal}}(x)\} = \frac{1}{2} \pm O\left(\frac{c\log s}{\sqrt{n}}\right).$$

Recalling (23) and observing that by Theorem 8 we have that $\mathbf{Pr}_{\boldsymbol{g}\sim N(0,1)^n}[\boldsymbol{g} \in \mathrm{S}_c] \leq e^{-\Omega(c^2)}$, we get that

$$\frac{1}{2} - \mathrm{opt}(\mathcal{D}_{\mathrm{ideal}}) \leq \frac{1}{n^5} + \sum_{c=1}^{4\sqrt{\ln n}} \frac{c\log s}{\sqrt{n}} \cdot e^{-\Omega(c^2)} \leq \frac{1}{n^5} + \frac{\log s}{\sqrt{n}} \sum_{c=1}^{\infty} c \cdot e^{-\Omega(c^2)} \leq \frac{O(\log s)}{\sqrt{n}},$$

and Theorem 24 is proved. ∎

A.2. Discretizing $\mathcal{D}_{\text{ideal}}$ to obtain $\mathcal{D}_{\text{actual}}$, and the Bayes optimal classifier for $\mathcal{D}_{\text{actual}}$

We now discretize the distribution $\mathcal{D}_{\text{ideal}}$ to construct the distribution $\mathcal{D}_{\text{actual}}$. We begin by recalling some results which will be useful for this construction.

Definition 25 *Let $\mathcal{X}_1$, $\mathcal{X}_2$ be two distributions supported on $\mathbb{R}^n$. The* Wasserstein distance *between $\mathcal{X}_1$ and $\mathcal{X}_2$, denoted by $\mathrm{d}_{\mathrm{W},1}(\mathcal{X}_1, \mathcal{X}_2)$ is defined to be*

$$\mathrm{d}_{\mathrm{W},1}(\mathcal{X}_1, \mathcal{X}_2) = \min_{\mathcal{Z}} \mathbf{E}_{\mathcal{Z}}[\|\mathcal{Z}_1 - \mathcal{Z}_2\|_1],$$

where $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$ is a coupling of $\mathcal{X}_1$ and $\mathcal{X}_2$.

The following fundamental result is due to Dudley Dudley (1969):

Theorem 26 *Let $\mathcal{X}$ be any compactly supported measure on $\mathbb{R}^n$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ be M random samples from $\mathcal{X}$ and let $\boldsymbol{X}_M$ be the resulting empirical measure. Then*

$$\mathbf{E}[\mathrm{d}_{\mathrm{W},1}(\mathcal{X}, \boldsymbol{X}_M)] = O(M^{-1/n}).$$

Let $\mathrm{U}_{\mathbb{S}^{n-1}}$ denote the Haar measure (i.e., the uniform measure) on $\mathbb{S}^{n-1}$. Instantiating Theorem 26 with $\mathrm{U}_{\mathbb{S}^{n-1}}$, we get the following corollary:

Corollary 27 *For any error parameter $\zeta > 0$, there exists $M_{n,\zeta}$ such that for any $M \geq M_{n,\zeta}$, there is a distribution $U_{M,\mathrm{emp}}$ which satisfies the following:*

1. $\mathrm{d}_{\mathrm{W},1}(U_{M,\mathrm{emp}}, \mathrm{U}_{\mathbb{S}^{n-1}}) \leq \zeta$.
2. *The distribution $U_{M,\mathrm{emp}}$ is uniform over its M-element support, which we denote by S_{actual}.*

We are now ready to construct the distribution $\mathcal{D}_{\text{actual}}$. We fix parameters ζ, p and M as follows:

$$\zeta\sqrt{\log(1/\zeta)} := \frac{1}{\Lambda \cdot \sqrt{n}}, \qquad M := \max\left\{M_{n,\zeta}, \frac{\Lambda^2}{\zeta}\right\}, \qquad p := \frac{\Lambda}{M}. \tag{27}$$

Definition 28 *A draw of a function $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ is sampled as follows: For each z in S_{actual}, define an independent Bernoulli random variable $\boldsymbol{W}_z$ which is 1 with probability p. The function $\boldsymbol{f}$ is*

$$\boldsymbol{f}(x) := \bigwedge_{z \in S_{\text{actual}} : \boldsymbol{W}_z = 1} \mathrm{slab}_z(x).$$

Given such a $\boldsymbol{f}$, we define $\mathrm{Rel}(\boldsymbol{f}) := \{z \in S_{\text{actual}} : \boldsymbol{W}_z = 1\}$

For intuition, $\mathrm{Rel}(\boldsymbol{f})$ can be viewed as the set of those elements of S_{actual} that are "relevant" to $\boldsymbol{f}$. With the definition of $\mathcal{D}_{\text{actual}}$ in hand, we define $\mathcal{D}_{\text{actual}}(x)$ (analogous to $\mathcal{D}_{\text{ideal}}(x)$) as follows:

$$\mathcal{D}_{\text{actual}}(x) = \mathop{\mathbf{Pr}}_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}}[\boldsymbol{f}(x) = 1].$$

Similar to $\mathcal{D}_{\text{ideal}}$, we now consider the Bayes optimal classifier $BO_{\mathcal{D}_{\text{actual}}}(x)$, which corresponds to the output of the best zero-query learning algorithm for an unknown target function $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$. The expected error of $BO_{\mathcal{D}_{\text{actual}}}$ is given by

$$\mathrm{opt}(\mathcal{D}_{\text{actual}}) := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)^n}[\min\{\mathcal{D}_{\text{actual}}(\boldsymbol{x}), 1 - \mathcal{D}_{\text{actual}}(\boldsymbol{x})\}].$$

The next lemma is the main result of this subsection and the rest of this subsection is devoted to its proof. It relates $\mathrm{opt}(\mathcal{D}_{\text{actual}})$ to $\mathrm{opt}(\mathcal{D}_{\text{ideal}})$ and completes Step 2 of the outline given earlier:

Lemma 29 *For $\mathcal{D}_{\text{actual}}$ and $\mathcal{D}_{\text{ideal}}$ as defined above and parameters ζ, M and p as set in (27),*

$$|\text{opt}(\mathcal{D}_{\text{actual}}) - \text{opt}(\mathcal{D}_{\text{ideal}})| = O(n^{-1/2}).$$

The proof of Theorem 29 requires several claims.

Claim 30 *Let vectors $z, z' \in \mathbb{S}^{n-1}$ satisfy $\|z - z'\|_2 \leq 1/3$. Then*

$$\Pr_{\boldsymbol{x} \sim N(0,1)^n}[\text{slab}_z(\boldsymbol{x}) \neq \text{slab}_{z'}(\boldsymbol{x})] \leq 5\|z - z'\|_2 \sqrt{\ln\left(\frac{1}{\|z - z'\|_2}\right)}.$$

Proof Define $\mathsf{Bd}_\kappa := \{y \in \mathbb{R} : ||y| - d| \leq \kappa\}$. For any parameter $t > 0$ and any $x \in \mathbb{R}^n$, observe that

$$\text{slab}_z(x) \neq \text{slab}_{z'}(x) \text{ only if } (|(z - z') \cdot x| \geq t\|z - z'\|_2) \text{ and } (z \cdot x \in \mathsf{Bd}_{t\|z-z'\|_2}). \tag{28}$$

Let us write $\mathsf{erfc}(t)$ to denote $\mathbf{Pr}_{\boldsymbol{x} \sim N(0,1)^n}[|\boldsymbol{x}| \geq t]$. Recalling that $\mathsf{erfc}(t) \leq (e^{-t^2} + e^{-2t^2})/2$ (e.g., see equation 10 in Chiani et al. (2003)), we have that

$$\Pr_{\boldsymbol{x} \sim N(0,1)^n}[|(z - z') \cdot \boldsymbol{x}| \geq t\|z - z'\|_2] \leq \frac{e^{-t^2} + e^{-2t^2}}{2}.$$

Likewise, using the fact that the density of the standard normal is bounded by 1 everywhere, we have that

$$\Pr_{\boldsymbol{x} \sim N(0,1)^n}[z \cdot \boldsymbol{x} \in \mathsf{Bd}_{t\|z-z'\|_2}] \leq 4t\|z - z'\|_2.$$

Plugging the last two equations back into (28), we have that

$$\begin{aligned}\Pr_{\boldsymbol{x} \sim N(0,1)^n}[\text{slab}_z(\boldsymbol{x}) \neq \text{slab}_{z'}(\boldsymbol{x})] &\leq \min_{t>0}\left\{\frac{e^{-t^2} + e^{-2t^2}}{2} + 4t\|z - z'\|_2\right\} \\ &\leq 5\|z - z'\|_2 \sqrt{\ln\left(\frac{1}{\|z - z'\|_2}\right)},\end{aligned}$$

giving Theorem 30. ■

The next (standard) claim relates the Poisson point process over a finite set $\mathcal{A}$ to the process of sampling each element independently (with a fixed probability) from $\mathcal{A}$.

Claim 31 *Let $\mathcal{A}$ be any set of size M and let $\Lambda > 0$. Consider the following two stochastic processes (a draw from the first process outputs a subset of $\mathcal{A}$ while a draw from the second process outputs a multiset of elements from $\mathcal{A}$):*

1. *The process $\mathsf{Indsample}(\mathcal{A}, \Lambda)$ produces a subset $\mathcal{B}_b \subseteq \mathcal{A}$ where each element $a \in \mathcal{A}$ is included independently with probability $p = \Lambda/|\mathcal{A}|$.*
2. *The process $\mathsf{Poisample}(\mathcal{A}, \Lambda)$ produces a multiset $\mathcal{B}_p$ of elements from $\mathcal{A}$ where we first draw $\boldsymbol{L} \sim \mathsf{Poi}(\Lambda)$ and then set $\mathcal{B}_p$ to be a multiset consisting of $\boldsymbol{L}$ independent uniform random elements from $\mathcal{A}$ (drawn with replacement).*

Then the statistical distance $\|\mathsf{Indsample}(\mathcal{A},\Lambda) - \mathsf{Poisample}(\mathcal{A},\Lambda)\|_1$ *is at most* $2\Lambda^2/M$.

Proof A draw of $\mathcal{B}_p$ from $\mathsf{Poisample}(\mathcal{A},\Lambda)$ can equivalently be generated as follows: for each $a \in A$, sample $\boldsymbol{x}_a \sim \mathsf{Poi}(p)$ independently at random and then include $\boldsymbol{x}_a$ many copies of a in $\mathcal{B}_p$. For $0 \le q \le 1$, let $\mathsf{Bern}(q)$ denote a Bernoulli random variable with expectation q. Recalling that $\|\mathsf{Poi}(q) - \mathsf{Bern}(q)\|_1 \le 2q^2$, applying this bound to every $a \in A$ and taking a union bound, we have that

$$\|\mathsf{Indsample}(\mathcal{A},\Lambda) - \mathsf{Poisample}(\mathcal{A},\Lambda)\|_1 \le \sum_{a\in A} 2p^2 = 2\frac{\Lambda^2}{M^2}\cdot M = \frac{2\Lambda^2}{M}.$$

■

Finally, to prove Theorem 29, we will use an intermediate distribution of functions defined as follows:

Definition 32 *For the parameter* Λ *defined earlier, we define the distribution* $\mathcal{D}_{\mathrm{inter}}$ *as follows: to sample a draw* $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{inter}}$*, we (i) first sample* $\boldsymbol{L} \sim \mathsf{Poi}(\Lambda)$*, and (ii) then sample* $\boldsymbol{z}_1,\dots,\boldsymbol{z}_{\boldsymbol{L}} \sim U_{M,\mathrm{emp}}$*. The function* $\boldsymbol{f}$ *is*

$$\boldsymbol{f}(x) := \bigwedge_{i=1}^{\boldsymbol{L}} \mathrm{slab}_{\boldsymbol{z}_i}(x).$$

As with $\mathcal{D}_{\mathrm{actual}}$ and $\mathcal{D}_{\mathrm{ideal}}$, we define $\mathcal{D}_{\mathrm{inter}}(x)$ and $\mathrm{opt}(\mathcal{D}_{\mathrm{inter}})$ as

$$\mathcal{D}_{\mathrm{inter}}(x) := \mathop{\mathbf{Pr}}_{\boldsymbol{f}\sim\mathcal{D}_{\mathrm{inter}}}[\boldsymbol{f}(x)=1], \qquad \mathrm{opt}(\mathcal{D}_{\mathrm{inter}}) := \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}[\min\{\mathcal{D}_{\mathrm{inter}}(\boldsymbol{x}), 1-\mathcal{D}_{\mathrm{inter}}(\boldsymbol{x})\}].$$

Now we are ready for the proof of Theorem 29:

Proof of Theorem 29. We begin with the following easy claim which shows that $\mathcal{D}_{\mathrm{inter}}(x)$ is very close to $\mathcal{D}_{\mathrm{actual}}(x)$ for every x:

Claim 33 *For any* $x \in \mathbb{R}^n$*,*

$$|\mathcal{D}_{\mathrm{inter}}(x) - \mathcal{D}_{\mathrm{actual}}(x)| \le \frac{2\Lambda^2}{M}.$$

Proof Observe that $\boldsymbol{f}_{\mathrm{inter}} \sim \mathcal{D}_{\mathrm{inter}}$ ($\boldsymbol{f}_{\mathrm{actual}} \sim \mathcal{D}_{\mathrm{actual}}$, respectively) can be sampled as follows: Sample $(\boldsymbol{z}_1,\dots,\boldsymbol{z}_{\boldsymbol{L}}) \sim \mathsf{Poisample}(S_{\mathrm{actual}},\Lambda)$ ($(\boldsymbol{y}_1,\dots,\boldsymbol{y}_{\boldsymbol{Q}}) \sim \mathsf{Indsample}(S_{\mathrm{actual}},\Lambda)$, respectively), and set

$$\boldsymbol{f}_{\mathrm{inter}}(x) = \bigwedge_{i=1}^{\boldsymbol{L}} \mathrm{slab}_{\boldsymbol{z}_i}(x) \qquad \text{and} \qquad \boldsymbol{f}_{\mathrm{actual}}(x) = \bigwedge_{i=1}^{\boldsymbol{Q}} \mathrm{slab}_{\boldsymbol{y}_i}(x).$$

It follows from Theorem 31 that $\|\mathsf{Poisample}(S_{\mathrm{actual}},\Lambda) - \mathsf{Indsample}(S_{\mathrm{actual}},\Lambda)\|_1 \le 2\Lambda^2/M$ and consequently $\|\mathcal{D}_{\mathrm{inter}} - \mathcal{D}_{\mathrm{actual}}\|_1 \le 2\Lambda^2/M$. This implies that

$$|\mathcal{D}_{\mathrm{inter}}(x) - \mathcal{D}_{\mathrm{actual}}(x)| = \left| \mathop{\mathbf{Pr}}_{\boldsymbol{f}_{\mathrm{inter}}\sim\mathcal{D}_{\mathrm{inter}}}[\boldsymbol{f}_{\mathrm{actual}}(x)=1] - \mathop{\mathbf{Pr}}_{\boldsymbol{f}_{\mathrm{actual}}\sim\mathcal{D}_{\mathrm{actual}}}[\boldsymbol{f}_{\mathrm{actual}}(x)=1] \right| \le \frac{2\Lambda^2}{M}.$$

■

Next we relate the average value of $\mathcal{D}_{\mathrm{inter}}$ (for $\boldsymbol{x} \sim N(0,1)^n$) to the average value of $\mathcal{D}_{\mathrm{ideal}}$:

Claim 34

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}[|\mathcal{D}_{\mathrm{inter}}(\boldsymbol{x}) - \mathcal{D}_{\mathrm{ideal}}(\boldsymbol{x})|] = O(\Lambda\zeta\sqrt{\log(1/\zeta)}).$$

Proof Recall that by Theorem 27 there exists a coupling $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$ between $U_{M,\mathrm{emp}}$ and $\mathrm{U}_{\mathbb{S}^{n-1}}$ such that $\mathbf{E}[\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1] \le \zeta$. We consider the following coupling between $\mathcal{D}_{\mathrm{inter}}$ and $\mathcal{D}_{\mathrm{ideal}}$:

1. Sample $\boldsymbol{L} \sim \mathsf{Poi}(\Lambda)$.
2. Sample $\{(\boldsymbol{z}_1^{(j)}, \boldsymbol{z}_2^{(j)})\}_{1\le j\le \boldsymbol{L}}$ independently from $\boldsymbol{Z}^{\boldsymbol{L}}$.
3. Define

$$\boldsymbol{f}_{\mathsf{in}}(x) = \bigwedge_{j=1}^{\boldsymbol{L}} \mathrm{slab}_{\boldsymbol{z}_1^{(j)}}(x) \qquad \text{and} \qquad \boldsymbol{f}_{\mathsf{id}}(x) = \bigwedge_{j=1}^{\boldsymbol{L}} \mathrm{slab}_{\boldsymbol{z}_2^{(j)}}(x).$$

Observe that $\boldsymbol{f}_{\mathsf{in}}$ follows the distribution $\mathcal{D}_{\mathrm{inter}}$ and $\boldsymbol{f}_{\mathsf{id}}$ follows the distribution $\mathcal{D}_{\mathrm{ideal}}$. Thus, the process above indeed describes a coupling between $\mathcal{D}_{\mathrm{inter}}$ and $\mathcal{D}_{\mathrm{ideal}}$. We consequently have

$$\begin{aligned}
|\mathrm{opt}(\mathcal{D}_{\mathrm{ideal}}) - \mathrm{opt}(\mathcal{D}_{\mathrm{inter}})| &\le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\left|\mathop{\mathbf{Pr}}_{\boldsymbol{f}_{\mathsf{id}}}[\boldsymbol{f}_{\mathsf{id}}(\boldsymbol{x}) = 1] - \mathop{\mathbf{Pr}}_{\boldsymbol{f}_{\mathsf{in}}}[\boldsymbol{f}_{\mathsf{in}}(\boldsymbol{x}) = 1]\right|\right] \\
&= \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\mathop{\mathbf{E}}_{\boldsymbol{Z}^{\boldsymbol{L}}}\left[\bigwedge_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_1^{(i)}}(\boldsymbol{x}) - \bigwedge_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_2^{(i)}}(\boldsymbol{x})\right]\right|\right] \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{Z}^{\boldsymbol{L}}}\left[\bigwedge_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_1^{(i)}}(\boldsymbol{x}) - \bigwedge_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_2^{(i)}}(\boldsymbol{x})\right]\right|\right] \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\left[\left|\mathop{\mathbf{E}}_{\boldsymbol{Z}^{\boldsymbol{L}}}\left[\sum_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_1^{(i)}}(\boldsymbol{x}) - \sum_{i=1}^{\boldsymbol{L}}\mathrm{slab}_{\boldsymbol{z}_2^{(i)}}(\boldsymbol{x})\right]\right|\right] \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\mathop{\mathbf{E}}_{\boldsymbol{Z}^{\boldsymbol{L}}}\sum_{i=1}^{\boldsymbol{L}}\left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\left|\mathrm{slab}_{\boldsymbol{z}_1^{(i)}}(\boldsymbol{x}) - \mathrm{slab}_{\boldsymbol{z}_2^{(i)}}(\boldsymbol{x})\right|\right]\right). \qquad (29)
\end{aligned}$$

Now, by Theorem 30, we have that

$$\left(\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\left|\mathrm{slab}_{\boldsymbol{z}_1^{(i)}}(\boldsymbol{x}) - \mathrm{slab}_{\boldsymbol{z}_2^{(i)}}(\boldsymbol{x})\right|\right]\right) \le 5\|\boldsymbol{z}_1^{(i)} - \boldsymbol{z}_2^{(i)}\|_2\sqrt{\log\left(\frac{1}{\|\boldsymbol{z}_1^{(i)} - \boldsymbol{z}_2^{(i)}\|_2}\right)}.$$

Plugging this back into (29), we have that

$$\begin{aligned}
|\mathrm{opt}(\mathcal{D}_{\mathrm{ideal}}) - \mathrm{opt}(\mathcal{D}_{\mathrm{inter}})| &\le \mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\mathop{\mathbf{E}}_{\boldsymbol{Z}^{\boldsymbol{L}}}\sum_{i=1}^{\boldsymbol{L}}\left[5\|\boldsymbol{z}_1^{(i)} - \boldsymbol{z}_2^{(i)}\|_2\sqrt{\log\left(\frac{1}{\|\boldsymbol{z}_1^{(i)} - \boldsymbol{z}_2^{(i)}\|_2}\right)}\right] \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{L}\sim\mathsf{Poi}(\Lambda)}\sum_{i=1}^{\boldsymbol{L}}\left[5\cdot\zeta\sqrt{\log(1/\zeta)}\right] \\
&\le 5\Lambda\zeta\sqrt{\log(1/\zeta)}, \qquad (30)
\end{aligned}$$

where the penultimate inequality used $\mathbf{E}[\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1] \le \zeta$ and the concavity of the function $x\sqrt{\log(1/x)}$. ■

Theorem 29 follows from Theorem 33 and Theorem 34, recalling the values of the parameters set in (27).

A.3. Analyzing query algorithms

Theorem 29 and Theorem 24 together imply a bound on the accuracy of the Bayes optimal classifier for $\mathcal{D}_{\text{actual}}$ when the algorithm makes zero queries to the target function $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$. To analyze the effect of queries, it will be useful to first consider an alternate combinatorial formulation of $\mathcal{D}_{\text{actual}}(x)$. For any point $x \in \mathbb{S}^n$, define $S_{\text{actual}}(x) = \{z \in S_{\text{actual}} : \text{slab}_z(x) = 0\}$. By definition of $\mathcal{D}_{\text{actual}}$, recalling the definition of p from Equation (27), we have that for any $x \in \mathbb{R}^n$,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}}[\boldsymbol{f}(x) = 1] = (1-p)^{|S_{\text{actual}}(x)|}. \tag{31}$$

Restated in these terms, Theorem 24 and Theorem 29 give us that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)^n}\left[\left|(1-p)^{|S_{\text{actual}}(\boldsymbol{x})|} - 1/2\right|\right] = O\left(\frac{\log s}{\sqrt{n}}\right) \tag{32}$$

We return to our overall goal of analyzing the Bayes optimal classifier when the learning algorithm makes at most s queries to the unknown target $\boldsymbol{f}$. While the actual MQ oracle, when invoked on $x \in \mathbb{R}^n$, returns the binary value of $\boldsymbol{f}(x)$, for the purposes of our analysis we consider an augmented oracle which provides more information and is described below.

A.3.1. AN AUGMENTED ORACLE, AND ANALYZING LEARNING ALGORITHMS THAT USE THIS ORACLE

Similar to Blum et al. (1998), to keep the analysis as clean as possible it is helpful for us to consider an augmented version of the MQ oracle. (Note that this is in the context of $\mathcal{D}_{\text{actual}}$, so the set S_{actual} is involved in what follows.) Fix an ordering of the elements in S_{actual}, and let f be a function in the support of $\mathcal{D}_{\text{actual}}$. Recalling the definition of $\text{Rel}(f)$ from Theorem 28, we observe that for any point $x \in \mathbb{R}^n$,

$$f(x) = 1 \text{ if and only if } S_{\text{actual}}(x) \cap \text{Rel}(f) = \emptyset.$$

This motivates the definition of our "augmented oracle" for f. Namely,

1. On input x, if $f(x) = 1$ then the oracle returns 1 (thereby indicating that $S_{\text{actual}}(x) \cap \text{Rel}(f) = \emptyset$).

2. On input x, if $f(x) = 0$ then the oracle returns the first $z \in S_{\text{actual}}$ (according to the above-described ordering on S_{actual}) for which $z \in S_{\text{actual}}(x) \cap \text{Rel}(f)$.[6]

It is clear that on any query string x, the augmented oracle for f provides at least as much information as the standard oracle for f. Thus, it suffices to prove a query lower bound for learning algorithms which have access to this augmented oracle.

At any point in the execution of the s-query learning algorithm, let X represent the list of query-answer pairs that have been received thus far from this augmented oracle. Let $\mathcal{D}_{\text{actual},X}$ denote the conditional distribution of $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ conditioned on the query-answer list given by X. As in Blum et al. (1998), the distribution $\mathcal{D}_{\text{actual},X}$ is quite clean and easy to describe. To do so, consider

6. We note that the need to define this "first z" is the main reason that we do not work with $\mathcal{D}_{\text{ideal}}$ directly and instead discretized it to obtain $\mathcal{D}_{\text{actual}}$.

a vector V_X whose entries are indexed by the elements of S_{actual}. For $z \in S_{\text{actual}}$, we define $V_X(z)$ as

$$V_X(z) := \mathop{\mathbf{Pr}}_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual},X}} [z \in \text{Rel}(\boldsymbol{f})].$$

Let us also define the Bernoulli random variables $\{\boldsymbol{W}_{X,z}\}_{z \in S_{\text{actual}}}$, where $\boldsymbol{W}_{X,z}$ is 1 if $z \in \text{Rel}(\boldsymbol{f})$ for $\boldsymbol{f} \sim \mathcal{D}_{\text{actual},X}$.

We begin by making the following observation:

Claim 35 *When X is the empty list (i.e. when zero queries have been made), each $V_X(z)$ is equal to p, and the Bernoulli random variables $\{\boldsymbol{W}_{X,z}\}_{z \in S_{\text{actual}}}$ are mutually independent.*

Let us consider what happens when the "current" query-answer list X is extended with a new query x. We can view the augmented oracle as operating as follows: it proceeds over each entry z in $S_{\text{actual}}(x)$ (according to the specified ordering), and:

1. If $V_X(z) = 0$, this means that the query-answer pairs already in X imply that $z \notin \text{Rel}(\boldsymbol{f})$. Then the augmented oracle proceeds to the next z.

2. If $V_X(z) = 1$, this means that the query-answer pairs already in X imply that $z \in \text{Rel}(\boldsymbol{f})$. In this case, the oracle stops and returns z (recall that this is a vector in $\mathbb{R}^n$, specifically an element of S_{actual}) to the algorithm. Note that this z is the first $z \in S_{\text{actual}}$ (in order) such that $\text{slab}_z(x) = 0$.

3. Finally, if $V_X(z) = p$, then the oracle fixes $\boldsymbol{W}_{X,z}$ to 1 with probability p and to 0 with probability $1 - p$. (Recall that the random variable $\boldsymbol{W}_{X,z}$ corresponds to the event that $z \in \text{Rel}(\boldsymbol{f})$.) If $\boldsymbol{W}_{X,z}$ is fixed to 0 then the oracle moves on to the next z, and if it is fixed to 1 then the oracle stops and returns z. As in the previous case, this is then the first z in S_{actual} such that $\text{slab}_z(x) = 0$.

Finally, we augment X with the query x and the above-defined response from the oracle. Based on the above description of the oracle, it is easy to see that the following holds:

Claim 36 *For any X, each entry of $V_X(z)$ is either 0, 1 or p. Further, for any X, the random variables $\boldsymbol{W}_{X,z}$ are mutually independent. Consequently, we can sample $\boldsymbol{f} \sim \mathcal{D}_{\text{actual},X}$ as*

$$\boldsymbol{f}(x) = \bigwedge_{z \in S_{\text{actual}} : \boldsymbol{W}_{X,z} = 1} \text{slab}_z(x).$$

Next, we have the following two claims (which correspond respectively to Claim 1 and Claim 2 of Blum et al. (1998)):

Claim 37 *If the learning algorithm makes s queries, then the number of entries in $V_X(\cdot)$ which are set to 1 is at most s.*

Theorem 37 is immediate from the above description of the oracle. The next claim is also fairly straightforward:

Claim 38 *If the learning algorithm makes s queries, then with probability at least $1 - e^{-\frac{s}{4}}$, the number of zero entries in V_X is bounded by $2s/p$.*

Proof Given any X, on a new query x the oracle iterates over all $z \in S_{\text{actual}}(x)$ and sets $V_X(z)$ to 0 with probability $1-p$ and 1 with probability p, stopping this process as soon as (a) it sets the first 1, or (b) it has finished iterating over all $z \in S_{\text{actual}}(x)$, or (c) the current $V_X(z)$ was already set to 1 in a previous round.

Thus, given any X, the number of new zeros added to V_X on a new query x is stochastically dominated by $\mathsf{Geom}(p)$, the geometric random variable with parameter p. It follows that the (random variable corresponding to the) total number of zeros in V_X is stochastically dominated by a sum of s independent variables, each following $\mathsf{Geom}(p)$. We now recall the following standard tail bound for sums of geometric random variables Janson (2018):

Theorem 39 *Let* $\mathbf{R}_1, \ldots, \mathbf{R}_s$ *be independent* $\mathsf{Geom}(p)$ *random variables. For* $\lambda \geq 1$,

$$\mathbf{Pr}\left[\mathbf{R}_1 + \ldots + \mathbf{R}_s \geq \frac{\lambda s}{p}\right] \leq e^{-s(\lambda - 1 - \ln \lambda)}.$$

Substituting $\lambda = 2$, we get that the number of zeros in V_X is bounded by $2s/p$ with probability at least $1 - e^{-s/4}$. This finishes the proof. ■

A.4. Proof of Theorem 2

All the pieces are now in place for us to finish our proof of Theorem 2. The high-level idea is that thanks to Theorem 37 and Theorem 38, the distribution $\mathcal{D}_{\text{actual},X}$ cannot be too different from $\mathcal{D}_{\text{actual}}$ as far as the accuracy of the Bayes optimal classifier is concerned; this, together with Theorem 29 and Theorem 24, gives the desired result.

Let $\mathcal{E}$ be the event (defined on the space of all possible outcomes of X, the list of at most s query-answer pairs) that the number of zero entries in V_X is at most $2s/p$. Observe that $\mathbf{Pr}[\overline{\mathcal{E}}] \leq e^{-s/4}$ by Theorem 38. We now bound the performance of the Bayes optimal estimator for $\mathcal{D}_{\text{actual},X}$ conditioned on the event $\mathcal{E}$.

Let $\mathcal{A}_1 = \{z \in S_{\text{actual}} : V_X(z) = 1\}$ and $\mathcal{A}_0 = \{z \in S_{\text{actual}} : V_X(z) = 0\}$. Using Theorem 36 and Theorem 37, we have the following observations:

- If $x \in \mathbb{R}^n$ is such that $S_{\text{actual}}(x) \cap \mathcal{A}_1 \neq \emptyset$, then $\mathop{\mathbf{Pr}}_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual},X}}[\boldsymbol{f}(x) = 0] = 1$.
- If $x \in \mathbb{R}^n$ is such that $S_{\text{actual}}(x) \cap \mathcal{A}_1 = \emptyset$, then $\mathop{\mathbf{Pr}}_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual},X}}[\boldsymbol{f}(x) = 1] = (1-p)^{|S_{\text{actual}}(x) \setminus \mathcal{A}_0|}$.
- $$\mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim N(0,1)^n}[S_{\text{actual}}(\boldsymbol{x}) \cap \mathcal{A}_1 \neq \emptyset] \leq \sum_{z \in \mathcal{A}_1} \mathbf{Pr}[\text{slab}_z(\boldsymbol{x}) = 0] \leq \frac{|\mathcal{A}_1|}{s^{100}} \leq \frac{1}{s^{99}}. \tag{33}$$

The last inequality uses Theorem 37 to bound the size of $|\mathcal{A}_1|$ and the definition of $\text{slab}_z(\cdot)$. Next, for any $z \in \mathcal{A}_0$, observe that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)^n}[\mathbb{1}[z \in S_{\text{actual}}(\boldsymbol{x})]] = \mathop{\mathbf{Pr}}_{\boldsymbol{x} \sim N(0,1)^n}[\text{slab}_z(\boldsymbol{x}) = 1] = \frac{1}{s^{100}}.$$

This immediately implies that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)^n}\left[\sum_{z \in \mathcal{A}_0} \mathbb{1}[z \in S_{\text{actual}}(\boldsymbol{x})]\right] = \frac{|\mathcal{A}_0|}{s^{100}} \leq \frac{2}{p \cdot s^{99}}.$$

By Markov's inequality, this implies that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\sum_{z\in\mathcal{A}_0}\mathbb{1}[z\in S_{\text{actual}}(\boldsymbol{x})]\geq\frac{2}{ps^{98}}\right]\leq\frac{1}{s}. \tag{34}$$

Let us say that $x\in\mathbb{R}^n$ is *good* if $S_{\text{actual}}(x)\cap\mathcal{A}_1=\emptyset$ and

$$\sum_{z\in\mathcal{A}_0}\mathbb{1}[z\in S_{\text{actual}}(x)]\geq\frac{2}{ps^{98}}.$$

By (34), we have that $\mathbf{Pr}_{\boldsymbol{x}\sim N(0,1)^n}[\boldsymbol{x}\text{ is good}]\leq 1/s$. We observe that for any good x, we have

$$|S_{\text{actual}}(x)|-\frac{2}{ps^{98}}\leq|S_{\text{actual}}(x)\setminus\mathcal{A}_0|\leq|S_{\text{actual}}(x)|.$$

It follows that

$$(1-p)^{|S_{\text{actual}}(x)|}\cdot(1-p)^{-\frac{2}{ps^{98}}}\geq(1-p)^{|S_{\text{actual}}(x)\setminus\mathcal{A}_0|}\geq(1-p)^{|S_{\text{actual}}(x)|}.$$

Using the fact that $(1-p)^{-\frac{2}{ps^{98}}}\leq 1+\frac{4}{s^{98}}$, we have that

$$(1-p)^{|S_{\text{actual}}(x)|}\cdot\left(1+\frac{4}{s^{98}}\right)\geq(1-p)^{|S_{\text{actual}}(x)\setminus\mathcal{A}_0|}\geq(1-p)^{|S_{\text{actual}}(x)|}.$$

This implies that for any $x\in\mathbb{R}^n$ which is good,

$$\left|\mathcal{D}_{\text{actual},X}(x)-\frac{1}{2}\right|=\left|(1-p)^{|S_{\text{actual}}(x)\setminus\mathcal{A}_0|}-\frac{1}{2}\right|\leq\left|(1-p)^{|S_{\text{actual}}(x)|}-\frac{1}{2}\right|+\frac{4}{s^{98}}. \tag{35}$$

Combining this with (32), (34) and (33), we get that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)^n}\left[\left|\mathcal{D}_{\text{actual},X}(x)-\frac{1}{2}\right|\right]\leq\frac{1}{s}+\frac{4}{s^{98}}+\frac{1}{s^{99}}+O\left(\frac{\log s}{\sqrt{n}}\right).$$

This bounds the error of the Bayes optimal classifier for $\mathcal{D}_{\text{actual},X}$ conditioned on $\mathcal{E}$ to be at least $\frac{1}{2}-O(\frac{\log s}{\sqrt{n}})$. Observing that $\mathbf{Pr}[\mathcal{E}]\geq 1-e^{-s/4}$ and $s\geq n$, the proof of Theorem 2 is complete.

Appendix B. Correlation of a fixed vector with a random unit vector

In this section, we prove the following lemma.

Lemma 40 *Let $v\in\mathbb{S}^{n-1}$ be a fixed vector and $\boldsymbol{u}\in\mathbb{S}^{n-1}$ be a uniformly drawn element of $\mathbb{S}^{n-1}$. For $0<\varepsilon<1$ and $1/2\geq\beta\geq\alpha>\frac{1}{\sqrt{n}}$ such that $\beta=(1+\varepsilon)\alpha$, we have*

$$1\leq\frac{\mathbf{Pr}[|\langle v,\boldsymbol{u}\rangle|\geq\alpha]}{\mathbf{Pr}[|\langle v,\boldsymbol{u}\rangle|\geq\beta]}\leq 1+O(n\alpha^2\varepsilon)$$

provided that $n\cdot\alpha^2\cdot\varepsilon\leq\frac{1}{8e^2}$.

Proof It is well-known (see Baum (1990)) and easy to verify that

$$\mathbf{Pr}[\langle v, \boldsymbol{u}\rangle \geq \alpha] = \frac{A_{n-2}}{A_{n-1}} \int_{z=\alpha}^{1} (1-z^2)^{\frac{n-3}{2}} dz. \tag{36}$$

Here A_{n-1} is the surface area of the sphere $\mathbb{S}^{n-1}$. By symmetry, this implies that

$$\frac{\mathbf{Pr}[|\langle v, \boldsymbol{u}\rangle| \geq \alpha]}{\mathbf{Pr}[|\langle v, \boldsymbol{u}\rangle| \geq \beta]} = \frac{\int_{z=\alpha}^{1}(1-z^2)^{\frac{n-3}{2}} dz}{\int_{z=\beta}^{1}(1-z^2)^{\frac{n-3}{2}} dz.} \tag{37}$$

Define $F(\alpha)$ as

$$F(\alpha) = (1-\alpha^2)^{\frac{n-3}{2}}.$$

Define $\Delta = \frac{1}{n\alpha}$. Observe that $\Delta \leq \alpha$ (for our choice of α) and $\Delta\alpha = \frac{1}{n}$. Using this, we have

$$(1-(\alpha+\Delta)^2) \geq (1-\alpha^2)(1-4\alpha\Delta).$$

This implies

$$F(\alpha+\Delta) = (1-(\alpha+\Delta)^2)^{\frac{n-3}{2}} \geq (1-\alpha^2)^{\frac{n-3}{2}} \cdot (1-4\alpha\Delta)^{\frac{n-3}{2}} \geq F(\alpha) \cdot \frac{1}{e^2}. \tag{38}$$

Then, using (38),

$$\int_{z=\alpha}^{1} (1-z^2)^{\frac{n-3}{2}} dz \geq \int_{z=\alpha}^{z=\alpha+\Delta} (1-z^2)^{\frac{n-3}{2}} dz \geq \frac{\Delta}{e^2} F(\alpha). \tag{39}$$

On the other hand,

$$\int_{z=\alpha}^{z=\beta} (1-z^2)^{\frac{n-3}{2}} dz \leq (\beta-\alpha)F(\alpha) = \varepsilon \cdot \alpha \cdot F(\alpha) \tag{40}$$

Note that the assumption $n\alpha^2\varepsilon \leq 1/(8e^2)$ translates to $\varepsilon\alpha \leq \frac{\Delta}{8e^2}$. Combining (40), (39) and this observation, we get

$$1 \leq \frac{\mathbf{Pr}[|\langle v, \boldsymbol{u}\rangle| \geq \alpha]}{\mathbf{Pr}[|\langle v, \boldsymbol{u}\rangle| \geq \beta]} \leq 1 + \frac{\varepsilon\alpha}{\Delta/e^2 - \varepsilon\alpha} \leq 1 + O(n\alpha^2\varepsilon).$$

■