Sample and Computationally Efficient Learning Algorithms under S-Concave Distributions

Maria-Florina Balcan
Machine Learning Department
Carnegie Mellon University, USA
ninamf@cs.cmu.edu

Hongyang Zhang*
Machine Learning Department
Carnegie Mellon University, USA
hongyanz@cs.cmu.edu

Abstract

We provide new results for noise-tolerant and sample-efficient learning algorithms under s-concave distributions. The new class of s-concave distributions is a broad and natural generalization of log-concavity, and includes many important additional distributions, e.g., the Pareto distribution and t-distribution. This class has been studied in the context of efficient sampling, integration, and optimization, but much remains unknown about the geometry of this class of distributions and their applications in the context of learning. The challenge is that unlike the commonly used distributions in learning (uniform or more generally log-concave distributions), this broader class is not closed under the marginalization operator and many such distributions are fat-tailed. In this work, we introduce new convex geometry tools to study the properties of s-concave distributions and use these properties to provide bounds on quantities of interest to learning including the probability of disagreement between two halfspaces, disagreement outside a band, and the disagreement coefficient. We use these results to significantly generalize prior results for margin-based active learning, disagreement-based active learning, and passive learning of intersections of halfspaces. Our analysis of geometric properties of s-concave distributions might be of independent interest to optimization more broadly.

1 Introduction

Developing provable learning algorithms is one of the central challenges in learning theory. The study of such algorithms has led to significant advances in both the theory and practice of passive and active learning. In the passive learning model, the learning algorithm has access to a set of labeled examples sampled i.i.d. from some unknown distribution over the instance space and labeled according to some underlying target function. In the active learning model, however, the algorithm can access unlabeled examples and request labels of its own choice, and the goal is to learn the target function with significantly fewer labels. In this work, we study both learning models in the case where the underlying distribution belongs to the class of s-concave distributions.

Prior work on noise-tolerant and sample-efficient algorithms mostly relies on the assumption that the distribution over the instance space is log-concave [2, 22, 9, 57]. A distribution is *log-concave* if the logarithm of its density is a concave function. The assumption of log-concavity has been made for a few purposes: for computational efficiency reasons and for sample efficiency reasons. For computational efficiency reasons, it was made to obtain a noise-tolerant algorithm even for seemingly simple decision surfaces like linear separators. These simple algorithms exist for noiseless scenarios, e.g., via linear programming [51], but they are notoriously hard once we have noise [25, 42, 32]; This is why progress on noise-tolerant algorithms has focused on uniform [36, 43] and log-concave distributions [6]. Other concept spaces, like intersections of halfspaces, even have no

*Corresponding author.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

computationally efficient algorithm in the noise-free settings that works under general distributions, but there has been nice progress under uniform and log-concave distributions [44]. For sample efficiency reasons, in the context of active learning, we need distributional assumptions in order to obtain label complexity improvements [26]. The most concrete and general class for which prior work obtains such improvements is when the marginal distribution over instance space satisfies log-concavity [59, 9]. In this work, we provide a broad generalization of all above results, showing how they extend to s-concave distributions ($s < 0$). A distribution with density $f(x)$ is *s-concave* if $f(x)^s$ is a concave function. We identify key properties of these distributions that allow us to simultaneously extend all above results.

How general and important is the class of s-concave distributions? The class of s-concave distributions is very broad and contains many well-known (classes of) distributions as special cases. For example, when $s \to 0$, s-concave distributions reduce to *log-concave* distributions. Furthermore, the s-concave class contains infinitely many fat-tailed distributions that do not belong to the class of log-concave distributions, e.g., Cauchy, Pareto, and t distributions, which have been widely applied in the context of theoretical physics and economics, but much remains unknown about how the provable learning algorithms, such as active learning of halfspaces, perform under these realistic distributions. We also compare s-concave distributions with nearly-log-concave distributions, a slightly broader class of distributions than log-concavity. A distribution with density $f(x)$ is nearly-log-concave if for any $\lambda \in [0,1]$, $x_1, x_2 \in \mathbb{R}^n$, we have $f(\lambda x_1 + (1-\lambda)x_2) \geq e^{-0.0154} f(x_1)^{\lambda} f(x_2)^{1-\lambda}$ [9]. The class of s-concave distributions includes many important extra distributions which do not belong to the nearly-log-concave distributions: a nearly-log-concave distribution must have sub-exponential tails (see Theorem 11, [9]), while the tail probability of an s-concave distribution might decay much slower (see Theorem 1 (6)). We also note that efficient sampling, integration and optimization algorithms for s-concave distributions have been well understood [23, 37]. Our analysis of s-concave distributions bridges these algorithms to the strong guarantees of noise-tolerant and sample-efficient learning algorithms.

1.1 Our Contributions

Structural Results. We study various geometric properties of s-concave distributions. These properties serve as the structural results for many provable learning algorithms, e.g., margin-based active learning [9], disagreement-based active learning [56, 35], learning intersections of halfspaces [44], etc. When $s \to 0$, our results exactly reduce to those for log-concave distributions [9, 4, 6]. Below, we state our structural results informally:

Theorem 1 (Informal)**.** *Let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$. Then there exist **closed-form** functions $\gamma(s,m)$, $f_1(s,n)$, $f_2(s,n)$, $f_3(s,n)$, $f_4(s,n)$, and $f_5(s,n)$ such that*

1. (Weakly Closed under Marginal) *The marginal of $\mathcal{D}$ over m arguments (or cumulative distribution function, CDF) is isotropic $\gamma(s,m)$-concave. (Theorems 3, 4)*
2. (Lower Bound on Hyperplane Disagreement) *For any two unit vectors u and v in $\mathbb{R}^n$, $f_1(s,n)\theta(u,v) \leq \Pr_{x\sim\mathcal{D}}[\text{sign}(u\cdot x) \neq \text{sign}(v\cdot x)]$, where $\theta(u,v)$ is the angle between u and v. (Theorem 12)*
3. (Probability of Band) *There is a function $d(s,n)$ such that for any unit vector $w \in \mathbb{R}^n$ and any $0 < t \leq d(s,n)$, we have $f_2(s,n)t < \Pr_{x\sim\mathcal{D}}[|w\cdot x| \leq t] \leq f_3(s,n)t$. (Theorem 11)*
4. (Disagreement outside Margin) *For any absolute constant $c_1 > 0$ and any function $f(s,n)$, there exists a function $f_4(s,n) > 0$ such that $\Pr_{x\sim\mathcal{D}}[\text{sign}(u\cdot x) \neq \text{sign}(v\cdot x)$ and $|v\cdot x| \geq f_4(s,n)\theta(u,v)] \leq c_1 f(s,n)\theta(u,v)$. (Theorem 13)*
5. (Variance in 1-D Direction) *There is a function $d(s,n)$ such that for any unit vectors u and a in $\mathbb{R}^n$ such that $\|u-a\| \leq r$ and for any $0 < t \leq d(s,n)$, we have $\mathbb{E}_{x\sim\mathcal{D}_{u,t}}[(a\cdot x)^2] \leq f_5(s,n)(r^2+t^2)$, where $\mathcal{D}_{u,t}$ is the conditional distribution of $\mathcal{D}$ over the set $\{x : |u\cdot x| \leq t\}$. (Theorem 14)*
6. (Tail Probability) *We have $\Pr[\|x\| > \sqrt{n}t] \leq \left[1 - \frac{cst}{1+ns}\right]^{(1+ns)/s}$. (Theorem 5)*

If $s \to 0$ (i.e., the distribution is log-concave), then $\gamma(s,m) \to 0$ and the functions $f(s,n)$, $f_1(s,n)$, $f_2(s,n)$, $f_3(s,n)$, $f_4(s,n)$, $f_5(s,n)$, and $d(s,n)$ are all absolute constants.

To prove Theorem 1, we introduce multiple new techniques, e.g., *extension of Prekopa-Leindler theorem* and *reduction to baseline function* (see the supplementary material for our techniques), which might be of independent interest to optimization more broadly.

Table 1: Comparisons with prior distributions for margin-based active learning, disagreement-based active learning, and Baum's algorithm.

	Prior Work		Ours
Margin (Efficient, Noise)	uniform [5]	log-concave [6]	s-concave
Disagreement	uniform [34]	nearly-log-concave [9]	s-concave
Baum's	symmetric [11]	log-concave [44]	s-concave

Margin Based Active Learning: We apply our structural results to margin-based active learning of a halfspace w^* under any isotropic s-concave distribution for both *realizable* and *adversarial* noise models. In the realizable case, the instance X is drawn from an isotropic s-concave distribution and the label $Y = \text{sign}(w^* \cdot X)$. In the adversarial noise model, an adversary can corrupt any η ($\leq O(\epsilon)$) fraction of labels. For both cases, we show that there exists a *computationally efficient* algorithm that outputs a linear separator w_T such that $\Pr_{x\sim\mathcal{D}}[\text{sign}(w_T \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$ (see Theorems 15 and 16). The label complexity w.r.t. $1/\epsilon$ improves exponentially over the passive learning scenario under s-concave distributions, though the underlying distribution might be fat-tailed. To the best of our knowledge, this is the first result concerning the *computationally-efficient, noise-tolerant* margin-based active learning *under the broader class of s-concave distributions*. Our work solves an open problem proposed by Awasthi et al. [6] about exploring wider classes of distributions for provable active learning algorithms.

Disagreement Based Active Learning: We apply our results to agnostic disagreement-based active learning under s-concave distributions. The key to the analysis is estimating the disagreement coefficient, a distribution-dependent measure of complexity that is used to analyze certain types of active learning algorithms, e.g., the CAL [24] and A^2 algorithm [7]. We work out the disagreement coefficient under isotropic s-concave distribution (see Theorem 17). By composing it with the existing work on active learning [27], we obtain a bound on label complexity under the class of s-concave distributions. As far as we are aware, this is the first result concerning disagreement-based active learning that goes beyond log-concave distributions. Our bounds on the disagreement coefficient match the best known results for the much less general case of log-concave distributions [9]; Furthermore, they apply to the s-concave case where we allow an arbitrary number of discontinuities, a case not captured by [31].

Learning Intersections of Halfspaces: Baum's algorithm is one of the most famous algorithms for learning the intersections of halfspaces. The algorithm was first proposed by Baum [11] under symmetric distribution, and later extended to log-concave distribution by Klivans et al. [44] as these distributions are almost symmetric. In this paper, we show that approximate symmetry also holds for the case of s-concave distributions. With this, we work out the label complexity of Baum's algorithm under the broader class of s-concave distributions (see Theorem 18), and advance the state-of-the-art results (see Table 1).

We provide lower bounds to partially show the tightness of our analysis. Our results can be potentially applied to other provable learning algorithms as well [38, 58, 13, 57, 10], which might be of independent interest. We discuss our techniques and other related papers in the supplementary material.

2 Preliminary

Before proceeding, we define some notations and clarify our problem setup in this section.

Notations: We will use capital or lower-case letters to represent random variables, $\mathcal{D}$ to represent an s-concave distribution, and $\mathcal{D}_{u,t}$ to represent the conditional distribution of $\mathcal{D}$ over the set $\{x : |u \cdot x| \leq t\}$. We define the *sign* function as $\text{sign}(x) = +1$ if $x \geq 0$ and -1 otherwise. We denote by $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ the beta function, and $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ the gamma function. We will consider a single norm for the vectors in $\mathbb{R}^n$, namely, the 2-norm denoted by $\|x\|$. We will frequently use μ (or μ_f, $\mu_\mathcal{D}$) to represent the measure of the probability distribution $\mathcal{D}$ with density function f. The notation $\mathsf{ball}(w^*, t)$ represents the set $\{w \in \mathbb{R}^n : \|w - w^*\| \leq t\}$. For convenience, the symbol $\oplus$ slightly differs from the ordinary addition $+$: For $f = 0$ or $g = 0$, $\{f^s \oplus g^s\}^{1/s} = 0$; Otherwise, $\oplus$ and $+$ are the same. For $u, v \in \mathbb{R}^n$, we define the angle between them as $\theta(u, v)$.

2.1 From Log-Concavity to S-Concavity

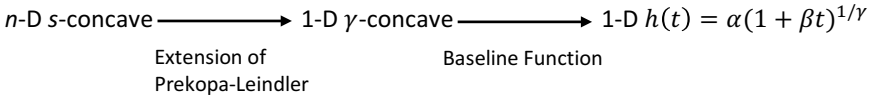
We begin with the definition of s-concavity. There are slight differences among the definitions of s-concave density, s-concave distribution, and s-concave measure.

Definition 1 (S-Concave (Density) Function, Distribution, Measure). *A function f: $\mathbb{R}^n \to \mathbb{R}_+$ is s-concave, for $-\infty \le s \le 1$, if $f(\lambda x + (1-\lambda)y) \ge (\lambda f(x)^s + (1-\lambda) f(y)^s)^{1/s}$ for all $\lambda \in [0,1]$, $\forall x, y \in \mathbb{R}^n$.[2] A probability distribution $\mathcal{D}$ is s-concave, if its density function is s-concave. A probability measure μ is s-concave if $\mu(\lambda A + (1-\lambda)B) \ge [\lambda \mu(A)^s + (1-\lambda)\mu(B)^s]^{1/s}$ for any sets $A, B \subseteq \mathbb{R}^n$ and $\lambda \in [0,1]$.*

Special classes of s-concave functions include *concavity* ($s = 1$), *harmonic-concavity* ($s = -1$), *quasi-concavity* ($s = -\infty$), etc. The conditions in Definition 1 are progressively weaker as s becomes smaller: s_1-concave densities (distributions, measures) are s_2-concave if $s_1 \ge s_2$. Thus one can verify [23]: concave ($s = 1$) $\subsetneq$ log-concave ($s = 0$) $\subsetneq$ s-concave ($s < 0$) $\subsetneq$ quasi-concave ($s = -\infty$).

3 Structural Results of S-Concave Distributions: A Toolkit

In this section, we develop geometric properties of s-concave distribution. The challenge is that unlike the commonly used distributions in learning (uniform or more generally log-concave distributions), this broader class is not closed under the marginalization operator and many such distributions are fat-tailed. To address this issue, we introduce several new techniques. We first introduce the extension of the Prekopa-Leindler inequality so as to reduce the high-dimensional problem to the one-dimensional case. We then reduce the resulting one-dimensional s-concave function to a well-defined baseline function, and explore the geometric properties of that baseline function. We summarize our high-level proof ideas briefly by the following figure.



3.1 Marginal Distribution and Cumulative Distribution Function

We begin with the analysis of the marginal distribution, which forms the basis of other geometric properties of s-concave distributions ($s \le 0$). Unlike the (nearly) log-concave distribution where the marginal remains (nearly) log-concave, the class of s-concave distributions is not closed under the marginalization operator. To study the marginal, our primary tool is the theory of convex geometry. Specifically, we will use an extension of the Prékopa-Leindler inequality developed by Brascamp and Lieb [20], which allows for a characterization of the integral of s-concave functions.

Theorem 2 ([20], Thm 3.3). *Let $0 < \lambda < 1$, and H_s, G_1, and G_2 be non-negative integrable functions on $\mathbb{R}^m$ such that $H_s(\lambda x + (1-\lambda)y) \ge [\lambda G_1(x)^s \oplus (1-\lambda) G_2(y)^s]^{1/s}$ for every $x, y \in \mathbb{R}^m$. Then $\int_{\mathbb{R}^m} H_s(x)dx \ge \left[\lambda \left(\int_{\mathbb{R}^m} G_1(x)dx\right)^\gamma + (1-\lambda)\left(\int_{\mathbb{R}^m} G_2(x)dx\right)^\gamma\right]^{1/\gamma}$ for $s \ge -1/m$, with $\gamma = s/(1+ms)$.*

Building on this, the following theorem plays a key role in our analysis of the marginal distribution.

Theorem 3 (Marginal). *Let $f(x,y)$ be an s-concave density on a convex set $K \subseteq \mathbb{R}^{n+m}$ with $s \ge -\frac{1}{m}$. Denote by $K|_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m \text{ s.t. } (x,y) \in K\}$. For every x in $K|_{\mathbb{R}^n}$, consider the section $K(x) \triangleq \{y \in \mathbb{R}^m : (x,y) \in K\}$. Then the marginal density $g(x) \triangleq \int_{K(x)} f(x,y)dy$ is γ-concave on $K|_{\mathbb{R}^n}$, where $\gamma = \frac{s}{1+ms}$. Moreover, if $f(x,y)$ is isotropic, then $g(x)$ is isotropic.*

Similar to the marginal, the CDF of an s-concave distribution might not remain in the same class. This is in sharp contrast to log-concave distributions. The following theorem studies the CDF of an s-concave distribution.

Theorem 4. *The CDF of s-concave distribution in $\mathbb{R}^n$ is γ-concave, where $\gamma = \frac{s}{1+ns}$ and $s \ge -\frac{1}{n}$.*

Theorem 3 and 4 serve as the bridge that connects high-dimensional s-concave distributions to one-dimensional γ-concave distributions. With them, we are able to reduce the high-dimensional problem to the one-dimensional one.

[2] When $s \to 0$, we note that $\lim_{s\to 0}(\lambda f(x)^s + (1-\lambda) f(y)^s)^{1/s} = \exp(\lambda \log f(x) + (1-\lambda)\log f(y))$. In this case, $f(x)$ is known to be log-concave.

3.2 Fat-Tailed Density

Tail probability is one of the most distinct characteristics of s-concave distributions compared to (nearly) log-concave distributions. While it can be shown that the (nearly) log-concave distribution has an exponentially small tail (Theorem 11, [9]), the tail of an s-concave distribution is fat, as proved by the following theorem.

Theorem 5 (Tail Probability)**.** *Let x come from an isotropic distribution over $\mathbb{R}^n$ with an s-concave density. Then for every $t \geq 16$, we have $\Pr[\|x\| > \sqrt{n}t] \leq \left[1 - \frac{cst}{1+ns}\right]^{(1+ns)/s}$, where c is an absolute constant.*

Theorem 5 is almost tight for $s < 0$. To see this, consider X that is drawn from a one-dimensional Pareto distribution with density $f(x) = (-1 - \frac{1}{s})^{-\frac{1}{s}} x^{\frac{1}{s}}$ $(x \geq \frac{s+1}{-s})$. It can be easily seen that $\Pr[X > t] = \left[\frac{-s}{s+1}t\right]^{\frac{s+1}{s}}$ for $t \geq \frac{s+1}{-s}$, which matches Theorem 5 up to an absolute constant factor.

3.3 Geometry of S-Concave Distributions

We now investigate the geometry of s-concave distributions. We first consider one-dimensional s-concave distributions: We provide bounds on the density of centroid-centered halfspaces (Lemma 6) and range of the density function (Lemma 7). Building upon these, we develop geometric properties of high-dimensional s-concave distributions by reducing the distributions to the one-dimensional case based on marginalization (Theorem 3).

3.3.1 One-Dimensional Case

We begin with the analysis of one-dimensional halfspaces. To bound the probability, a normal technique is to bound the centroid region and the tail region separately. However, the challenge is that the s-concave distribution is fat-tailed (Theorem 5). So while the probability of a one-dimensional halfspace is bounded below by an absolute constant for log-concave distributions, such a probability for s-concave distributions decays as s (≤ 0) becomes smaller. The following lemma captures such an intuition.

Lemma 6 (Density of Centroid-Centered Halfspaces)**.** *Let X be drawn from a one-dimensional distribution with s-concave density for $-1/2 \leq s \leq 0$. Then $\Pr(X \geq \mathbb{E}X) \geq (1+\gamma)^{-1/\gamma}$ for $\gamma = s/(1+s)$.*

We also study the image of a one-dimensional s-concave density. The following condition for $s > -1/3$ is for the existence of second-order moment.

Lemma 7. *Let $g : \mathbb{R} \to \mathbb{R}_+$ be an isotropic s-concave density function and $s > -1/3$. (a) For all x, $g(x) \leq \frac{1+s}{1+3s}$; (b) We have $g(0) \geq \sqrt{\frac{1}{3(1+\gamma)^{3/\gamma}}}$, where $\gamma = \frac{s}{s+1}$.*

3.3.2 High-Dimensional Case

We now move on to the high-dimensional case $(n \geq 2)$. In the following, we will assume $-\frac{1}{2n+3} \leq s \leq 0$. Though this working range of s vanishes as n becomes larger, it is almost the broadest range of s that we can hopefully achieve: Chandrasekaran et al. [23] showed a lower bound of $s \geq -\frac{1}{n-1}$ if one require the s-concave distribution to have good geometric properties. In addition, we can see from Theorem 3 that if $s < -\frac{1}{n-1}$, the marginal of an s-concave distribution might even not exist; Such a case does happen for certain s-concave distributions with $s < -\frac{1}{n-1}$, e.g., the Cauchy distribution. So our range of s is almost tight up to a $1/2$ factor.

We start our analysis with the density of centroid-centered halfspaces in high-dimensional spaces.

Lemma 8 (Density of Centroid-Centered Halfspaces)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an s-concave density function, and let H be any halfspace containing its centroid. Then $\int_H f(x)dx \geq (1+\gamma)^{-1/\gamma}$ for $\gamma = s/(1+ns)$.*

Proof. W.L.O.G., we assume H is orthogonal to the first axis. By Theorem 3, the first marginal of f is $s/(1+(n-1)s)$-concave. Then by Lemma 6, $\int_H f(x)dx \geq (1+\gamma)^{-1/\gamma}$, where $\gamma = s/(1+ns)$. □

The following theorem is an extension of Lemma 7 to high-dimensional spaces. The proofs basically reduce the n-dimensional density to its first marginal by Theorem 3, and apply Lemma 7 to bound the image.

Theorem 9 (Bounds on Density). *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an isotropic s-concave density. Then*

(a) Let $d(s,n) = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$, where $\beta = \frac{s}{1+(n-1)s}$ and $\gamma = \frac{\beta}{1+\beta}$. For any $u \in \mathbb{R}^n$ such that $\|u\| \le d(s,n)$, we have $f(u) \ge \left(\frac{\|u\|}{d}((2-2^{-(n+1)s})^{-1}-1)+1\right)^{1/s} f(0)$.

(b) $f(x) \le f(0)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s}$ for every x.

(c) There exists an $x \in \mathbb{R}^n$ such that $f(x) > (4e\pi)^{-n/2}$.

(d) $(4e\pi)^{-n/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+\frac{1}{s}}\right)^s - 1\right]^{-\frac{1}{s}} < f(0) \le (2-2^{-(n+1)s})^{1/s}\frac{n\Gamma(n/2)}{2\pi^{n/2}d^n}$.

(e) $f(x) \le (2-2^{-(n+1)s})^{1/s}\frac{n\Gamma(n/2)}{2\pi^{n/2}d^n}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s}$ for every x.

(f) For any line ℓ through the origin, $\int_\ell f \le (2-2^{-ns})^{1/s}\frac{(n-1)\Gamma((n-1)/2)}{2\pi^{(n-1)/2}d^{n-1}}$.

Theorem 9 provides uniform bounds on the density function. To obtain more refined upper bound on the image of s-concave densities, we have the following lemma. The proof is built upon Theorem 9.

Lemma 10 (More Refined Upper Bound on Densities). *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an isotropic s-concave density. Then $f(x) \le \beta_1(n,s)(1-s\beta_2(n,s)\|x\|)^{1/s}$ for every $x \in \mathbb{R}^n$, where*

$$\beta_1(n,s) = \frac{(2-2^{-(n+1)s})^{\frac{1}{s}}}{2\pi^{n/2}d^n}(1-s)^{-1/s}n\Gamma(n/2)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s},$$

$$\beta_2(n,s) = \frac{2\pi^{(n-1)/2}d^{n-1}}{(n-1)\Gamma((n-1)/2)}(2-2^{-ns})^{-\frac{1}{s}}\frac{[(a(n,s)+(1-s)\beta_1(n,s)^s)^{1+\frac{1}{s}} - a(n,s)^{1+\frac{1}{s}}]s}{\beta_1(n,s)^s(1+s)(1-s)},$$

$a(n,s) = (4e\pi)^{-ns/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1}$, $\gamma = \frac{\beta}{1+\beta}$, $\beta = \frac{s}{1+(n-1)s}$, *and* $d = (1+\gamma)^{-\frac{1}{\gamma}}\frac{1+3\beta}{3+3\beta}$.

We also give an *absolute* bound on the measure of band.

Theorem 11 (Probability inside Band). *Let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$. Denote by $f_3(s,n) = 2(1+ns)/(1+(n+2)s)$. Then for any unit vector w, $\Pr_{x\sim\mathcal{D}}[|w\cdot x| \le t] \le f_3(s,n)t$. Moreover, if $t \le d(s,n) \triangleq \left(\frac{1+2\gamma}{1+\gamma}\right)^{-\frac{1+\gamma}{\gamma}}\frac{1+3\gamma}{3+3\gamma}$ where $\gamma = \frac{s}{1+(n-1)s}$, then $\Pr_{x\sim\mathcal{D}}[|w\cdot x| \le t] > f_2(s,n)t$, where $f_2(s,n) = 2(2-2^{-2\gamma})^{-1/\gamma}(4e\pi)^{-1/2}\left(2\left(\frac{1+\gamma}{1+3\gamma}\sqrt{3}\left(\frac{1+2\gamma}{1+\gamma}\right)^{\frac{3+3\gamma}{2\gamma}}\right)^\gamma - 1\right)^{-1/\gamma}$.*

To analyze the problem of learning linear separators, we are interested in studying the disagreement between the hypothesis of the output and the hypothesis of the target. The following theorem captures such a characteristic under s-concave distributions.

Theorem 12 (Probability of Disagreement). *Assume $\mathcal{D}$ is an isotropic s-concave distribution in $\mathbb{R}^n$. Then for any two unit vectors u and v in $\mathbb{R}^n$, we have $d_{\mathcal{D}}(u,v) = \Pr_{x\sim\mathcal{D}}[\text{sign}(u\cdot x) \ne \text{sign}(v\cdot x)] \ge f_1(s,n)\theta(u,v)$, where $f_1(s,n) = c(2-2^{-3\alpha})^{-\frac{1}{\alpha}}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{1+1/\alpha}\right)^\alpha - 1\right]^{-\frac{1}{\alpha}}(1+\gamma)^{-2/\gamma}\left(\frac{1+3\beta}{3+3\beta}\right)^2$, c is an absolute constant, $\alpha = \frac{s}{1+(n-2)s}$, $\beta = \frac{s}{1+(n-1)s}$, $\gamma = \frac{s}{1+ns}$.*

Due to space constraints, all missing proofs are deferred to the supplementary material.

4 Applications: Provable Algorithms under S-Concave Distributions

In this section, we show that many algorithms that work under log-concave distributions behave well under s-concave distributions by applying the above-mentioned geometric properties. For simplicity, we will frequently use the notations in Theorem 1.

4.1 Margin Based Active Learning

We first investigate margin-based active learning under isotropic s-concave distributions in both *realizable* and *adversarial noise* models. The algorithm (see Algorithm 1) follows a localization technique: It proceeds in rounds, aiming to cut the error down by half in each round in the margin [8].

Algorithm 1 Margin Based Active Learning under S-Concave Distributions

Input: Parameters b_k, τ_k, r_k, m_k, κ, and T as in Theorem 16.
1: Draw m_1 examples from $\mathcal{D}$, label them and put them into W.
2: **For** $k = 1, 2, ..., T$
3: Find $v_k \in \mathsf{ball}(w_{k-1}, r_k)$ to approximately minimize the hinge loss over W s.t. $\|v_k\| \leq 1$: $\ell_{\tau_k} \leq \min_{w \in \mathsf{ball}(w_{k-1}, r_k) \cap \mathsf{ball}(0,1)} \ell_{\tau_k}(w, W) + \kappa/8$.
4: Normalize v_k, yielding $w_k = \frac{v_k}{\|v_k\|}$; Clear the working set W.
5: **While** m_{k+1} additional data points are not labeled
6: Draw sample x from $\mathcal{D}$.
7: **If** $|w_k \cdot x| \geq b_k$, reject x; **else** ask for label of x and put into W.
Output: Hypothesis w_T.

4.1.1 Relevant Properties of S-Concave Distributions

The analysis requires more refined geometric properties as below. Theorem 13 basically claims that the error mostly concentrates in a band, and Theorem 14 guarantees that the variance in any 1-D direction cannot be too large. We defer the detailed proofs to the supplementary material.

Theorem 13 (Disagreement outside Band)**.** *Let u and v be two vectors in $\mathbb{R}^n$ and assume that $\theta(u, v) = \theta < \pi/2$. Let $\mathcal{D}$ be an isotropic s-concave distribution. Then for any absolute constant $c_1 > 0$ and any function $f_1(s, n) > 0$, there exists a function $f_4(s, n) > 0$ such that $\Pr_{x \sim \mathcal{D}}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x)$ and $|v \cdot x| \geq f_4(s, n)\theta] \leq c_1 f_1(s, n)\theta$, where $f_4(s, n) = \frac{4\beta_1(2,\alpha)B(-1/\alpha-3,3)}{-c_1 f_1(s,n)\alpha^3 \beta_2(2,\alpha)^3}$, $B(\cdot, \cdot)$ is the beta function, $\alpha = s/(1 + (n-2)s)$, $\beta_1(2, \alpha)$ and $\beta_2(2, \alpha)$ are given by Lemma 10.*

Theorem 14 (1-D Variance)**.** *Assume that $\mathcal{D}$ is isotropic s-concave. For d given by Theorem 9 (a), there is an absolute C_0 such that for all $0 < t \leq d$ and for all a such that $\|u - a\| \leq r$ and $\|a\| \leq 1$, $\mathbb{E}_{x \sim \mathcal{D}_{u,t}}[(a \cdot x)^2] \leq f_5(s, n)(r^2 + t^2)$, where $f_5(s, n) = 16 + C_0 \frac{8\beta_1(2,\eta)B(-1/\eta-3,2)}{f_2(s,n)\beta_2(2,\eta)^3(\eta+1)\eta^2}$, $(\beta_1(2, \eta), \beta_2(2, \eta))$ and $f_2(s, n)$ are given by Lemma 10 and Theorem 11, and $\eta = \frac{s}{1+(n-2)s}$.*

4.1.2 Realizable Case

We show that margin-based active learning works under s-concave distributions in the realizable case.

Theorem 15. *In the realizable case, let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$. Then for $0 < \epsilon < 1/4$, $\delta > 0$, and absolute constants c, there is an algorithm (see the supplementary material) that runs in $T = \lceil \log \frac{1}{c\epsilon} \rceil$ iterations, requires $m_k = O\Big(\frac{f_3 \min\{2^{-k} f_4 f_1^{-1}, d\}}{2^{-k}} \Big(n \log \frac{f_3 \min\{2^{-k} f_4 f_1^{-1}, d\}}{2^{-k}} + \log \frac{1+s-k}{\delta}\Big)\Big)$ labels in the k-th round, and outputs a linear separator of error at most ϵ with probability at least $1 - \delta$. In particular, when $s \to 0$ (a.k.a. log-concave), we have $m_k = O\left(n + \log(\frac{1+s-k}{\delta})\right)$.*

By Theorem 15, we see that the algorithm of margin-based active learning under s-concave distributions works almost as well as the log-concave distributions in the resizable case, improving exponentially w.r.t. the variable $1/\epsilon$ over passive learning algorithms.

4.1.3 Efficient Learning with Adversarial Noise

In the adversarial noise model, an adversary can choose any distribution $\widetilde{\mathcal{P}}$ over $\mathbb{R}^n \times \{+1, -1\}$ such that the marginal $\mathcal{D}$ over $\mathbb{R}^n$ is s-concave but an η fraction of labels can be flipped adversarially. The analysis builds upon an induction technique where in each round we do hinge loss minimization in the band and cut down the 0/1 loss by half. The algorithm was previously analyzed in [5, 6] for the special class of log-concave distributions. In this paper, we analyze it for the much more general class of s-concave distributions.

Theorem 16. *Let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$ over x and the label y obey the adversarial noise model. If the rate η of adversarial noise satisfies $\eta < c_0\epsilon$ for some absolute constant c_0, then for $0 < \epsilon < 1/4$, $\delta > 0$, and an absolute constant c, Algorithm 1 runs in $T = \lceil \log \frac{1}{c\epsilon} \rceil$ iterations, outputs a linear separator w_T such that $\Pr_{x \sim \mathcal{D}}[\text{sign}(w_T \cdot x) \neq \text{sign}(w^* \cdot x)] \leq \epsilon$ with probability at least $1 - \delta$. The label complexity in the k-th round is $m_k = O\left(\frac{[b_{k-1}s + \tau_k(1+ns)[1-(\delta/(\sqrt{n}(k+k^2)))^{s/(1+ns)}] + \tau_k s]^2}{\kappa^2 \tau_k^2 s^2} n \left(n + \log \frac{k+k^2}{\delta}\right)\right)$, where $\kappa = \max\left\{\frac{f_3 \tau_k}{f_2 \min\{b_{k-1}, d\}}, \frac{b_{k-1}\sqrt{f_5}}{\tau_k \sqrt{f_2}}\right\}$, $\tau_k = \Theta\left(f_1^{-2} f_2^{-1/2} f_3 f_4^2 f_5^{1/2} 2^{-(k-1)}\right)$, and $b_k = \min\{\Theta(2^{-k} f_4 f_1^{-1}), d\}$. In particular, if $s \to 0$, $m_k = O\left(n \log(\frac{n}{\epsilon\delta})(n + \log(\frac{k}{\delta}))\right)$.*

By Theorem 16, the label complexity of margin-based active learning improves exponentially over that of passive learning w.r.t. $1/\epsilon$ even under fat-tailed s-concave distributions and challenging adversarial noise model.

4.2 Disagreement Based Active Learning

We apply our results to the analysis of disagreement-based active learning under s-concave distributions. The key is estimating the disagreement coefficient, a measure of complexity of active learning problems that can be used to bound the label complexity [34]. Recall the definition of the disagreement coefficient w.r.t. classifier w^*, precision ϵ, and distribution $\mathcal{D}$ as follows. For any $r > 0$, define $\mathsf{ball}_{\mathcal{D}}(w, r) = \{u \in \mathcal{H} : d_{\mathcal{D}}(u, w) \leq r\}$ where $d_{\mathcal{D}}(u, w) = \Pr_{x\sim\mathcal{D}}[(u \cdot x)(w \cdot x) < 0]$. Define the disagreement region as $\text{DIS}(\mathcal{H}) = \{x : \exists u, v \in \mathcal{H} \text{ s.t. } (u \cdot x)(v \cdot x) < 0\}$. Let the Alexander capacity $\text{cap}_{w^*,\mathcal{D}} = \frac{\Pr_{\mathcal{D}}(\text{DIS}(\mathsf{ball}_{\mathcal{D}}(w^*,r)))}{r}$. The disagreement coefficient is defined as $\Theta_{w^*,\mathcal{D}}(\epsilon) = \sup_{r\geq\epsilon}[\text{cap}_{w^*,\mathcal{D}}(r)]$. Below, we state our results on the disagreement coefficient under isotropic s-concave distributions.

Theorem 17 (Disagreement Coefficient)**.** *Let $\mathcal{D}$ be an isotropic s-concave distribution over $\mathbb{R}^n$. For any w^* and $r > 0$, the disagreement coefficient is $\Theta_{w^*,\mathcal{D}}(\epsilon) = O\left(\frac{\sqrt{n}(1+ns)^2}{s(1+(n+2)s)f_1(s,n)}(1-\epsilon^{\frac{s}{1+ns}})\right)$. In particular, when $s \to 0$ (a.k.a. log-concave), $\Theta_{w^*,\mathcal{D}}(\epsilon) = O(\sqrt{n}\log(1/\epsilon))$.*

Our bounds on the disagreement coefficient match the best known results for the much less general case of log-concave distributions [9]; Furthermore, they apply to the s-concave case where we allow arbitrary number of discontinuities, a case not captured by [31]. The result immediately implies concrete bounds on the label complexity of disagreement-based active learning algorithms, e.g., CAL [24] and A^2 [7]. For instance, by composing it with the result from [27], we obtain a bound of $\widetilde{O}\left(n^{3/2}\frac{(1+ns)^2}{s(1+(n+2)s)f(s)}(1-\epsilon^{s/(1+ns)})\left(\log^2\frac{1}{\epsilon}+\frac{OPT^2}{\epsilon^2}\right)\right)$ for *agnostic* active learning under an isotropic s-concave distribution $\mathcal{D}$. Namely, it suffices to output a halfspace with error at most $OPT + \epsilon$, where $OPT = \min_w \mathsf{err}_{\mathcal{D}}(w)$.

4.3 Learning Intersections of Halfspaces

Baum [11] provided a polynomial-time algorithm for learning the intersections of halfspaces w.r.t. symmetric distributions. Later, Klivans [44] extended the result by showing that the algorithm works under any distribution $\mathcal{D}$ as long as $\mu_{\mathcal{D}}(E) \approx \mu_{\mathcal{D}}(-E)$ for any set E. In this section, we show that it is possible to learn intersections of halfspaces under the broader class of s-concave distributions.

Theorem 18. *In the PAC realizable case, there is an algorithm (see the supplementary material) that outputs a hypothesis h of error at most ϵ with probability at least $1-\delta$ under isotropic s-concave distributions. The label complexity is $M(\epsilon/2, \delta/4, n^2) + \max\{2m_2/\epsilon, (2/\epsilon^2)\log(4/\delta)\}$, where $M(\epsilon, \delta, m)$ is defined by $M(\epsilon, \delta, n) = O\left(\frac{n}{\epsilon}\log\frac{1}{\epsilon}+\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, $m_2 = M(\max\{\delta/(4eKm_1), \epsilon/2\}, \delta/4, n)$, $K = \beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3}\frac{3+1/\kappa}{h(\kappa)d^{3+1/\kappa}}$, $d = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$, $h(\kappa) = \left(\frac{1}{d}((2-2^{-4\kappa})^{-1}-1)+1\right)^{\frac{1}{\kappa}}(4e\pi)^{-\frac{3}{2}}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{2+\frac{1}{\kappa}}\right)^{\kappa}-1\right]^{-1/\kappa}$, $\beta = \frac{\kappa}{1+2\kappa}$, $\gamma = \frac{\kappa}{1+\kappa}$, and $\kappa = \frac{s}{1+(n-3)s}$. In particular, if $s \to 0$ (a.k.a. log-concave), K is an absolute constant.*

5 Lower Bounds

In this section, we give information-theoretic lower bounds on the label complexity of passive and active learning of homogeneous halfspaces under s-concave distributions.

Theorem 19. *For a fixed value $-\frac{1}{2n+3} \leq s \leq 0$ we have: (a) For any s-concave distribution $\mathcal{D}$ in $\mathbb{R}^n$ whose covariance matrix is of full rank, the sample complexity of learning origin-centered linear separators under $\mathcal{D}$ in the passive learning scenario is $\Omega\left(nf_1(s,n)/\epsilon\right)$; (b) The label complexity of active learning of linear separators under s-concave distributions is $\Omega\left(n\log\left(f_1(s,n)/\epsilon\right)\right)$.*

If the covariance matrix of $\mathcal{D}$ is not of full rank, then the intrinsic dimension is less than d. So our lower bounds essentially apply to all s-concave distributions. According to Theorem 19, it is possible to have an exponential improvement of label complexity w.r.t. $1/\epsilon$ over passive learning by active sampling, even though the underlying distribution is a fat-tailed s-concave distribution. This observation is captured by Theorems 15 and 16.

6 Conclusions

In this paper, we study the geometric properties of s-concave distributions. Our work advances the state-of-the-art results on the margin-based active learning, disagreement-based active learning, and

learning intersections of halfspaces w.r.t. the distributions over the instance space. When $s \to 0$, our results reduce to the best-known results for log-concave distributions. The geometric properties of s-concave distributions can be potentially applied to other learning algorithms, which might be of independent interest more broadly.

Acknowledgements. This work was supported in part by grants NSF-CCF 1535967, NSF CCF-1422910, NSF CCF-1451177, a Sloan Fellowship, and a Microsoft Research Fellowship.

References

[1] M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[2] D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *ACM Symposium on Theory of Computing*, pages 156–163, 1991.

[3] P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Annual Conference on Learning Theory*, pages 167–190, 2015.

[4] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.

[5] P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *ACM Symposium on Theory of Computing*, pages 449–458, 2014.

[6] P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):50, 2017.

[7] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[8] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Annual Conference on Learning Theory*, pages 35–50, 2007.

[9] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Annual Conference on Learning Theory*, pages 288–316, 2013.

[10] M.-F. Balcan and H. Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 2955–2963, 2016.

[11] E. B. Baum. A polynomial time algorithm that learns two hidden unit nets. *Neural Computation*, 2(4):510–522, 1990.

[12] D. Bertsimas and S. Vempala. Solving convex programs by random walks. *Journal of the ACM*, 51(4):540–556, 2004.

[13] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *International Conference on Machine Learning*, pages 49–56, 2009.

[14] A. Beygelzimer, D. J. Hsu, J. Langford, and C. Zhang. Search improves label for active learning. In *Advances in Neural Information Processing Systems*, pages 3342–3350, 2016.

[15] A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, pages 199–207, 2010.

[16] A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. In *IEEE Symposium on Foundations of Computer Science*, pages 330–338, 1996.

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

[18] S. G. Bobkov. Large deviations and isoperimetry over convex probability measures with heavy tails. *Electronic Journal of Probability*, 12:1072–1100, 2007.

[19] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: A survey of recent advances. *ESAIM: Probability and Statistics*, 9(9):323–375, 2005.

[20] H. J. Brascamp and E. H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

[21] C. Caramanis and S. Mannor. An inequality for nearly log-concave distributions with applications to learning. In *Annual Conference on Learning Theory*, pages 534–548, 2004.

[22] C. Caramanis and S. Mannor. An inequality for nearly log-concave distributions with applications to learning. *IEEE Transactions on Information Theory*, 53(3):1043–1057, 2007.

[23] K. Chandrasekaran, A. Deshpande, and S. Vempala. Sampling s-concave functions: The limit of convexity based isoperimetry. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 420–433. 2009.

[24] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[25] A. Daniely. Complexity theoretic limitations on learning halfspaces. In *ACM Symposium on Theory of computing*, pages 105–117, 2016.

[26] S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, volume 17, pages 337–344, 2004.

[27] S. Dasgupta, D. J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, pages 353–360, 2007.

[28] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Annual Conference on Learning Theory*, pages 249–263, 2005.

[29] J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *ACM Symposium on Theory of computing*, pages 315–320, 2004.

[30] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Information, prediction, and query by committee. *Advances in Neural Information Processing Systems*, pages 483–483, 1993.

[31] E. Friedman. Active learning for smooth problems. In *Annual Conference on Learning Theory*, 2009.

[32] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[33] Q. Han and J. A. Wellner. Approximation and estimation of s-concave densities via Rényi divergences. *The Annals of Statistics*, 44(3):1332–1359, 2016.

[34] S. Hanneke. A bound on the label complexity of agnostic active learning. In *International Conference on Machine Learning*, pages 353–360, 2007.

[35] S. Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309, 2014.

[36] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[37] A. T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

[38] D. M. Kane, S. Lovett, S. Moran, and J. Zhang. Active classification with comparison queries. In *IEEE Symposium on Foundations of Computer Science*, pages 355–366, 2017.

[39] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[40] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[41] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[42] A. Klivans and P. Kothari. Embedding hard learning problems into gaussian space. *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems*, 28:793–809, 2014.

[43] A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[44] A. R. Klivans, P. M. Long, and A. K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 588–600. 2009.

[45] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. In *IEEE Symposium on Foundations of Computer Science*, pages 177–186, 2002.

[46] S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993.

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

[48] P. M. Long. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

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

[50] M. Minsky and S. Papert. Perceptrons–extended edition: An introduction to computational geometry, 1987.

[51] R. A. Servedio. *Efficient algorithms in computational learning theory*. PhD thesis, Harvard University, 2001.

[52] S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the zero-one loss. *arXiv preprint arXiv:1005.3681*, 2010.

[53] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, pages 135–166, 2004.

[54] V. Vapnik. *Estimations of dependences based on statistical data*. Springer, 1982.

[55] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[56] L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(Jul):2269–2292, 2011.

[57] Y. Xu, H. Zhang, A. Singh, A. Dubrawski, and K. Miller. Noise-tolerant interactive learning using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2428–2437, 2017.

[58] S. Yan and C. Zhang. Revisiting perceptron: Efficient and label-optimal active learning of halfspaces. *arXiv preprint arXiv:1702.05581*, 2017.

[59] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.

A Our Techniques

In this section, we introduce the techniques used for obtaining our results.

Marginalization: Our results are inspired by isoperimetric inequality for s-concave distributions by the work of Chandrasekaran et al. [23]. Roughly, the isoperimetry states that if two sets K_1 and K_2 are well-separated, then the area B between them has large measure *relative to the measure of the two sets* (see Figure 1). Results of this kind are particularly useful for margin-based active learning of halfspace [5, 4, 6]: The algorithm proceeds in rounds, aiming to cut down the error by half in each round in the band. Since the measure of the band is large or even dominates, the error over the whole space decreases almost by half in each round, resulting in exponentially fast convergence rate. However, in order to make the analysis of such algorithms work for s-concave distribution, we typically require more refined geometric properties than the isoperimetry as the isoperimetric inequality states nothing about the *absolute* measure of band under s-concave distributions.

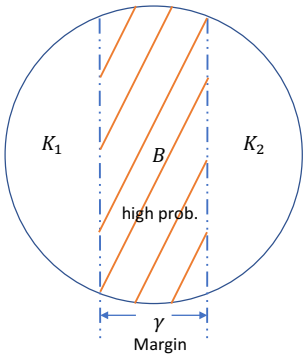


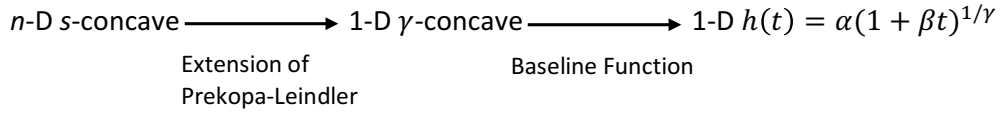
Figure 1: Isoperimetry.

The insight behind the isoperimetry is a collection of properties concerning the geometry of probability density. While the geometric properties of some classic paradigms, such as log-concave distributions (for the case of $s = 0$), are well-studied [49], it is typically hard to generalize those results to the s-concave distribution, for broader range of $s < 0$. This is due to the fact that the class of s-concave functions is not closed under marginalization: The marginal of an s-concave function may not be s-concave any more. This directly restricts the possibility of applying the prior proof techniques for log-concave distribution to the s-concave one. Furthermore, previous proofs heavily depend on the assumption that the density is light-tailed (see Theorem 11 in [9]), which is not applicable for probably fat-tailed s-concave distribution.

To mitigate the above concerns, we begin with a powerful tool from convex geometry by Brascamp and Lieb [20]. This result can be viewed as an extension of celebrated Prékopa-Leindler inequality, an integral inequality that is closely related to a number of classical inequalities in analysis and serves as the building block of isoperimetry under the log-concave distributions [21, 22]. With this, we can show that the marginal of any s-concave function is γ-concave, with a closed-form γ that is related to the parameter s and the dimension of marginalization. Our analysis is tight as there exists an s-concave function with a γ-concave marginal.

Reduction to 1-D Baseline Function: It is in general hard to study a high-dimensional s-concave distribution. Instead, we build on the marginalization technique described above to reduce each n-dimensional s-concave function to the one-dimensional case. Thus it suffices to investigate the geometry of one-dimensional γ-concave functions. But there are still infinitely many such functions in this class.

Our proofs take a novel analysis by reducing *all* one-dimensional γ-concave density to a certain baseline function. The baseline function should meet two goals: (a) It represents the worst case in the class of γ-concave functions, namely, such functions should achieve the bounds of geometric properties of our interest; (b) The function should be easy to analyze, e.g., with closed-form moments or integrations. Note that choosing a baseline function at the "boundary" between γ-concavity and non-γ-concavity classes readily achieves goal (a). To achieve goal (b), we set the "template" function as easy as $h(t) = \alpha(1 + \beta t)^{1/\gamma}$ for a particular choice of parameters α and β. Such functions have many good properties that one can exploit. First, the moments can be represented in closed-form by the beta function. This enables us to figure out the relations among moments of various orders explicitly and obtain a recursive inequality, which is critical for deducing the bounds of one-dimensional geometric properties. Second, $h(t)$ is at the "boundary" of γ-concave class: $h(t)^\eta$ is not a concave function for any $\eta < \gamma$. Therefore, this enables us to analyze the whole class of s-concavity by focusing on $h(t)$. Below, we summarize our high-level proof ideas briefly.



B Additional Related Work

Active Learning of Halfspace under Uniform Distribution: Learning halfspace has been extensively studied in the past decades [16, 45, 29, 36, 52, 41, 40, 39]. Probably one of the most famous results is the VC argument. Vapnik [54] and Blumer et al. [17] showed that any hypothesis that is consistent with $\widetilde{O}(n/\epsilon)$ labeled examples has error at most ϵ, if the VC dimension of the hypothesis class is n. The algorithm works under any data distribution and runs in polynomial time when the consistent hypothesis can be found efficiently, e.g., by linear programming in the realizable case. Other algorithms such as Perception [50], Winnow [47], and Support Vector Machine [55] provide better guarantees if the target vector has low ℓ_1 or ℓ_2 norm. All these results form the basis of passive learning.

To explore the possibility of further improving the label complexity, several algorithms were later proposed in the active learning literature [15, 14] under the uniform distributions [28, 30], among which disagreement-based active learning and margin-based active learning are two typical approaches. In the disagreement-based active learning, the algorithm proceeds in rounds, requesting the labels of instances in the disagreement region among the current candidate hypothesises. Cohn et al. [24] provided the first disagreement-based active learning algorithm in the realizable case. Balcan et al. [7] later extended such an algorithm to the agnostic setting by estimating the confidence interval of disagreement region. The analysis technique was further generalized thanks to Hanneke [34] by introducing the concept of disagreement coefficient, which is a new measure of complexity for active learning problems and serves as an important element for bounding the label complexity. However, this seminal work only focused on the disagreement coefficient under the uniform distribution.

Margin-based active learning is another line of research in the active learning literature. The algorithm proceeds in rounds, requesting labels of examples aggressively in the margin area around the current guess of hypothesis. Balcan et al. [8] first proposed an algorithm for margin-based active learning under the uniform distribution in the realizable case. They also provided guarantees under the *Tsybakov noise* model [53], but the algorithm is inefficient. To mitigate the issue, Awasthi et al. [3] considered a subclass of Tsybakov noise — *Massart noise* [19]. The algorithm runs in polynomial time by doing a sequence of hinge loss minimizations on the labeled instances. However, it was not clear then whether the analysis works for other distributions instead of the uniform one.

Geometry of Log-Concave Distribution: Log-concave distribution, a class of probability distributions such that the logarithm of density function is concave, is a common generalization of uniform distribution over the convex set [49]. Bertsimas and Vempala [12] and Kalai and Vempala [37] noticed that efficient sampling, integration, and optimization algorithms for this distribution class rely heavily on the good isoperimetry of density functions. Informally, a function has good isoperimetry if one cannot remove a small-measure set from its domain and partition the domain into two disjoint large-measure sets. The isoperimetry is commonly believed as a characteristic of good geometric properties. To see this, Lovász and Vempala [49] proved the isoperimetric inequality for the log-concave distribution, and provided a bunch of refined geometric properties for this distribution class. Going slightly beyond the log-concave distribution, Caramanis and Mannor [22] showed good isoperimetry for *nearly log-concave* distributions, but more refined geometry was not provided there.

Active learning of halfspace under (nearly) log-concave distribution has a natural connection to the geometry of that distribution (a.k.a. admissible distribution). The connection was first introduced by [9], and is sufficient for the success of disagreement-based and margin-based active learning under log-concave distribution [9]. To resolve the computational issue, Awasthi et al. [5] studied the probability of disagreement outside the margin under the log-concave distribution, and proposed an efficient algorithm for the challenging adversarial noise. More recently, Awasthi et al. [4] provided stronger guarantees for efficient learning of halfspace in the Massart noise model under log-concave distribution.

S-Concave Distribution: The problem of extending the log-concave distribution to the broader one for provable learning algorithms has received significant attention in recent years. Although some efforts have been devoted to generalizing the probability distribution, e.g., to the nearly log-concave distribution [9], the analysis is intrinsically built upon the geometry of log-concave distribution. Moreover, to the best of our knowledge, there is no *efficient, noise-tolerant* active learning algorithm that goes beyond the log-concave distribution. As a candidate extension, the class of s-concave distributions has many appealing properties that one can exploit [23, 33]: (a) The distribution class is much broader than the log-concave distributions as $s = 0$ implies the log-concavity; (b) The s-concave function mapping from $\mathbb{R}^n$ to $\mathbb{R}_+$ has good isoperimetry if $s \geq -1/(n-1)$; (c) Efficient sampling, integration, and optimization algorithms are available for such distribution class. All these properties inspire our work.

C Proof of Theorem 3

Theorem 3 (restated) *Let $f(x, y)$ be an s-concave density on a convex set $K \subseteq \mathbb{R}^{n+m}$ with $s \geq -\frac{1}{m}$. Denote by $K|_{\mathbb{R}^n} = \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m \text{ s.t. } (x, y) \in K\}$. For every x in $K|_{\mathbb{R}^n}$, consider the section $K(x) \triangleq \{y \in \mathbb{R}^m : (x, y) \in K\}$. Then the marginal density $g(x) \triangleq \int_{K(x)} f(x, y)dy$ is γ-concave on $K|_{\mathbb{R}^n}$, where $\gamma = \frac{s}{1+ms}$. Moreover, if $f(x, y)$ is isotropic, then $g(x)$ is isotropic.*

Proof. The proof that $g(x)$ is isotropic is standard [49]. We now prove the first part. Let x_1, x_2 be any two points. Define $g_i(y) = f(x_i, y)$ for $i = 1, 2$. So the functions $g_i(y)$ is defined on $K(x_i)$, $i = 1, 2$. Now let $x = \lambda x_1 + (1-\lambda)x_2$ for $\lambda \in (0, 1)$ and define $h_s(y) = f(x, y)$ on $K(x)$. Notice that for any $y_i \in K(x_i)$, $i = 1, 2$, $y = \lambda y_1 + (1-\lambda)y_2 \in K(x)$. To see this, by the convexity of the set K, the point $(x, y) = \lambda(x_1, y_1) + (1-\lambda)(x_2, y_2)$ belongs to K. So $y \in K(x)$, i.e., $\lambda K(x_1) + (1-\lambda)K(x_2) \subseteq K(x)$. Using the s-concavity of $f(x, y)$, we have $f(x, y) = f(\lambda(x_1, y_1)+(1-\lambda)(x_2, y_2)) \geq [\lambda f(x_1, y_1)^s + (1-\lambda)f(x_2, y_2)^s]^{1/s}$, which implies that $h_s(y) = h_s(\lambda y_1 + (1-\lambda)y_2) \geq [\lambda g_1(y_1)^s + (1-\lambda)g_2(y_2)^s]^{1/s}$. Denote by $I_{(\cdot)}$ the indicator function. So $h_s(\lambda y_1 + (1-\lambda)y_2)I_{K(x)}(y) \geq \left[\lambda(g_1(y_1)I_{K(x_1)}(y_1))^s \oplus (1-\lambda)(g_2(y_2)I_{K(x_2)}(y_2))^s\right]^{1/s}$. Set $H_s(y)=h_s(\lambda y_1+(1-\lambda)y_2)I_{K(x)}$, $G_1(y_1)=g_1(y_1)I_{K(x_1)}$, $G_2(y_1) = g_2(y_1)I_{K(x_2)}$. By Theorem 2,

$$
\begin{aligned}
g(x) =& \int_{\mathbb{R}^m} H_s(y)dy = \int_{\mathbb{R}^m} h_s(y)I_{K(x)}(y)dy \geq \left[(1-\lambda)\left(\int_{\mathbb{R}^n} G_1(y)dy\right)^\gamma + \lambda\left(\int_{\mathbb{R}^n} G_2(y)dy\right)^\gamma\right]^{1/\gamma} \\
=& \left[(1-\lambda)\left(\int_{\mathbb{R}^n} f(x_1, y_1)I_{K(x_1)}(y_1)dy_1\right)^\gamma + \lambda\left(\int_{\mathbb{R}^n} f(x_2, y_2)I_{K(x_2)}(y_2)dy_2\right)^\gamma\right]^{1/\gamma} \\
=& \left[(1-\lambda)g(x_1)^\gamma + \lambda g(x_2)^\gamma\right]^{1/\gamma},
\end{aligned}
$$

where $\gamma = s/(1+ms)$. Namely, the marginal function $g(x)$ is γ-concave. □

D Proof of Theorem 4

Similar to the marginal, the CDF of an s-concave distribution might not remain in the same class. This is in sharp contrast with the log-concave distributions. The following lemma from [20] provides a useful tool for our analysis of CDF, which basically claims that the measure of any s-concave distribution is γ-concave with a closed-form γ.

Lemma 20 ([20], Cor 3.4)**.** *The density function $f(x)$ is s-concave for $s \geq -1/n$ where $x \in \mathbb{R}^n$, if and only if the corresponding probability measure μ is γ-concave for $\gamma = \frac{s}{1+ns}$, namely, $\mu(\lambda A + (1-\lambda)B) \geq [\lambda\mu(A)^\gamma + (1-\lambda)\mu(B)^\gamma]^{1/\gamma}$, for any $A, B \subseteq \mathbb{R}^n$ and $\lambda \in [0, 1]$, where $\mu(A) = \int_A f(x)dx$.*

Lemma 20 is an extension of celebrated Brunn-Minkowski theorem. The following theorem concerning the CDF of an s-concave distribution is a straightforward result from Lemma 20.

Theorem 4 (restated) *The CDF of s-concave distribution in $\mathbb{R}^n$ is γ-concave, where $\gamma = \frac{s}{1+ns}$ and $s \geq -\frac{1}{n}$.*

Proof. Denote by $F(x)$ the CDF. Applying Lemma 20 to the set $A = \{x : x \le x_1\}$ and $B = \{x : x \le x_2\}$ and taking into account that $F(x_1) = \mu(A)$, $F(x_2) = \mu(B)$, and $F(\lambda x_1 + (1-\lambda)x_2) = \mu(\lambda A + (1-\lambda)B)$, we have the result. □

E Proof of Theorem 5

Tail probability is one of the most distinct characteristic of s-concave distributions compared to the nearly log-concavity. To study this, we first require a concentration result from [18].

Lemma 21 ([18], Thm 5.2)**.** *Let f be a Borel function on $\mathbb{R}^n$ and let m be a median for $|f|$ w.r.t. a κ-concave measure μ, where $\kappa < 0$. Then for every $t > 1$ such that $4\delta_f(\frac{1}{t}) \le 1$, we have $\mu[|f| > mt] \le \left[1 - \frac{c\kappa}{\delta_f(\frac{1}{t})}\right]^{1/\kappa}$, where c is a constant, $\delta_f(\epsilon) = \sup_{x,y} mes\{t \in (0,1) : |f(tx + (1-t)y)| \le \epsilon |f(x)|\}$, $0 \le \epsilon \le 1$, and mes stands for the Lebesgue measure.*

Now we are ready to bound the tail probability of s-concave density. While it can be shown that the nearly log-concave distribution has an exponentially small tail (Theorem 11, [9]), the tail of s-concave distribution is fat, as proved by the following theorem.

Theorem 5 (restated) *Let x come from an isotropic distribution over $\mathbb{R}^n$ with an s-concave density. Then for every $t \ge 16$, we have $\Pr[\|x\| > \sqrt{n}t] \le \left[1 - \frac{cst}{1+ns}\right]^{(1+ns)/s}$, where c is an absolute constant.*

Proof. Set function $f(x)$ in Lemma 21 as $\|x\|$. Bobkov [18] claimed that $\delta_f(\epsilon) \le 2\epsilon$. Also, Lemma 20 implies that the probability measure is $\kappa = \frac{s}{1+ns}$-concave.

By the definition of the median m, the Markov inequality, and the Jensen inequality, we have $\frac{1}{2} = \Pr[\|x\| \ge m] \le \frac{\mathbb{E}\|x\|}{m} \le \frac{\sqrt{\mathbb{E}\|x\|^2}}{m} = \frac{\sqrt{n}}{m}$, where the last equality is due to the isotropicity assumption. So by Lemma 21, we have that for every $t \ge 8$, $\Pr[\|x\| > 2\sqrt{n}t] \le \Pr[\|x\| > mt] \le [1 - cst/(1+ns)]^{(1+ns)/s}$. Replacing t with $t/2$, the proof is completed. □

F Proof of Lemma 6

Lemma 6 (restated) *Let X be a random variable drawn from a one-dimensional distribution with s-concave density for $-1/2 \le s \le 0$. Then*

$$\Pr(X \ge \mathbb{E}X) \ge (1+\gamma)^{-1/\gamma},$$

for $\gamma = s/(1+s)$.

Proof. Without loss of generality, we assume that $\mathbb{E}X = 0$ and $|X| \le K$. The general case then follows by translation transformation and approximating a general distribution with s-concave density by such bounded distributions.

Let $G(x) = \Pr(X \le x)$ be the CDF of the s-concave density. We first prove that $\Pr(X \le \mathbb{E}X) \ge (1+\gamma)^{-1/\gamma}$. By Theorem 4, $G(x)$ is γ-concave, monotone increasing such that $G(x) = 0$ for $x \le -K$ and $G(x) = 1$ for $x \ge K$, where $-1 \le \gamma = \frac{s}{1+s} \le 0$. Notice that the assumption of centroid 0 implies that $\int_{-K}^{K} xG'(x)dx = 0$, which equivalently means $\int_{-K}^{K} G(x)dx = K$ by integration by parts. Our goal is to prove that $G(0) \ge (1+\gamma)^{-1/\gamma}$.

The function G^γ is concave for $\gamma < 0$. Thus it lies above its tangent at 0. This means that $G(x) \le G(0)(1+\gamma cx)^{\frac{1}{\gamma}}$, where $c = G'(0)/G(0) > 0$. We now set K large enough so that $1/c < K$. Then

$$G(x) \le \begin{cases} G(0)(1+\gamma cx)^{\frac{1}{\gamma}}, & \text{if } x \le 1/c, \\ 1, & \text{if } x > 1/c. \end{cases}$$

So

$$\begin{aligned} K &= \int_{-K}^{K} G(x)dx \\ &\le \int_{-K}^{1/c} G(0)(1+\gamma cx)^{\frac{1}{\gamma}}dx + \int_{1/c}^{K} 1dx \\ &= \frac{G(0)}{c(\gamma+1)}[(1+\gamma)^{\frac{1}{\gamma}+1} - (1-\gamma cK)^{\frac{1}{\gamma}+1}] + K - \frac{1}{c} \\ &\le \frac{G(0)(1+\gamma)^{\frac{1}{\gamma}}}{c} + K - \frac{1}{c}, \end{aligned}$$

which implies that $G(0) \ge (1+\gamma)^{-1/\gamma}$ as claimed. Replacing X with $Y = -X$, we obtain the result. □

G Proof of Lemma 7

As a preliminary, we first prove the following lemma concerning the moments of s-concave distribution.

Lemma 22. *Let $g : \mathbb{R}_+ \to \mathbb{R}_+$ be an integrable function. Define $M_n(g) = \int_0^\infty t^n g(t)dt$, and suppose it exists. Then*

(a) The sequence $\{M_n(g) : n = 0, 1, ...\}$ is log-convex, which means $\log M_n(g)$ is convex w.r.t. variable n, or equivalently $M_n(g)M_{n+2}(g) \ge M_{n+1}(g)^2$ for any $n \in N$.

(b) If g is monotone decreasing, then the sequence defined by

$$M_n'(g) = \begin{cases} nM_{n-1}(g), & \text{if } n > 0, \\ g(0), & \text{if } n = 0, \end{cases}$$

is log-convex.

(c) If g is s-concave ($s > -1/(n+1)$), then the sequence $T_n(g) \triangleq M_n(g)/B(-1/s-n-1, n+1)$ is log-concave, which means $\log T_n(g)$ is concave w.r.t. n, or equivalently $T_n(g)T_{n+2}(g) \le T_{n+1}(g)^2$ for any $n \in N$.

(d) If g is s-concave, then $g(0)M_1(g) \le M_0(g)^2 \frac{1+s}{1+2s}$.

Proof. The proofs of Parts (a) and (b) are from [49].

(c) The intuition behind the proof is to choose a baseline s-concave function h which is at the "boundary" between the family of s-concave function and that of the non s-concave function. We show that h satisfies the equation

$$T_n(h)T_{n+2}(h) = T_{n+1}^2(h). \tag{1}$$

Then by the facts that h is at the "boundary" and g is any s-concave function, we have

$$T_{n+1}(h) \le T_{n+1}(g). \tag{2}$$

The conclusion follows from (1) and (2), and from our choice of h such that $T_n(h) = T_n(g)$ and $T_{n+2}(h) = T_{n+2}(g)$, by adjusting the slope and intercept of the linear function.

Formally, let $h(t) = \beta(1+\gamma t)^{1/s}$ be the above-mentioned baseline s-concave function ($\beta, \gamma > 0$) such that

$$M_n(h) = M_n(g) \quad \text{and} \quad M_{n+2}(h) = M_{n+2}(g)$$

(This holds because there are two parameters β, γ and two equations). That means

$$\int_0^\infty t^n(h(t) - g(t))dt = 0 \quad \text{and} \quad \int_0^\infty t^{n+2}(h(t) - g(t))dt = 0.$$

Then it follows that the graph of h must intersect the graph of g at least twice. Since g is s-concave, which implies the uni-modality, the graphs of h and g intersect exactly at two points $0 \le a < b$.

Moreover, $h \le g$ in the interval $[a,b]$ and $h \ge g$ outside the interval. That is to say, $(t-a)(t-b)$ has the same sign as $h-g$. Thus

$$\int_0^\infty (t-a)(t-b)t^n(h(t)-g(t))dt \ge 0.$$

Namely,

$$0 = \int_0^\infty t^{n+2}(h(t)-g(t))dt + ab\int_0^\infty t^n(h(t)-g(t))dt \ge (a+b)\int_0^\infty t^{n+1}(h(t)-g(t))dt.$$

This implies that

$$M_{n+1}(h) = \int_0^\infty t^{n+1}h(t)dt \le \int_0^\infty t^{n+1}g(t)dt = M_{n+1}(g).$$

Since

$$M_n(h) = \int_0^\infty t^n\beta(1+\gamma t)^{1/s}dt = B(-1/s-n-1,n+1)\frac{\beta}{\gamma^{n+1}}$$

for $s > -1/(n+1)$, we have

$$\begin{aligned}
\frac{M_n(g)}{B(-1/s-n-1,n+1)}\frac{M_{n+2}(g)}{B(-1/s-n-3,n+3)} &= \frac{M_n(h)}{B(-1/s-n-1,n+1)}\frac{M_{n+2}(h)}{B(-1/s-n-3,n+3)} \\
&= \frac{\beta}{\gamma^{n+1}}\cdot\frac{\beta}{\gamma^{n+3}} \\
&= \left(\frac{M_{n+1}(h)}{B(-1/s-n-2,n+2)}\right)^2 \\
&\le \left(\frac{M_{n+1}(g)}{B(-1/s-n-2,n+2)}\right)^2,
\end{aligned}$$

as desired.

(d) The proof is almost the same as that of Part (c). Let $h(t) = \beta(1+\gamma t)^{1/s}$ be an s-concave function ($\beta,\gamma > 0$) such that

$$h(0) = g(0) \quad \text{and} \quad M_1(h) = M_1(g).$$

So the graphs of h and g intersect exactly at two points 0 and $a > 0$, and hence

$$\int_0^\infty t(t-a)t^{-1}(h(t)-g(t))dt \ge 0.$$

That means

$$0 = \int_0^\infty t(h(t)-g(t))dt \ge a\int_0^\infty (h(t)-g(t))dt,$$

or equivalently,

$$M_0(h) \le M_0(g).$$

Note that $h(0)M_1(h) = M_0(h)^2\frac{1+s}{1+2s}$ by (G). Then the conclusion follows by the fact

$$g(0)M_1(g) = h(0)M_1(h) = M_0(h)^2\frac{1+s}{1+2s} \le M_0(g)^2\frac{1+s}{1+2s}.$$

□

Now we are ready to prove Lemma 7.

Lemma 7 (restated) *Let $g:\mathbb{R}\to\mathbb{R}_+$ be an isotropic s-concave density function and $s > -1/3$.*

(a) For all x, $g(x) \le \frac{1+s}{1+3s}$.

(b) We have $g(0) \ge \sqrt{\frac{1}{3(1+\gamma)^{3/\gamma}}}$, where $\gamma = \frac{s}{s+1}$.

Proof. (a) Let z be the maximum point of function g. Intuitively, if the value of the function g evaluated at z is too large, the corresponding distribution has a small deviation from z (Second part of the proof below). However, the moment property of Lemma 22 restricts that the second moment cannot be too small (First part of the proof below), which leads to a contradiction.

Formally, suppose that $g(z) > \frac{1+s}{1+3s}$. Define

$$M_i = \int_z^\infty (x-z)^i g(x)dx, \quad \text{and} \quad N_i = \int_{-\infty}^z (z-x)^i g(x)dx.$$

By the isotropicity of function g, we have

$$M_0 + N_0 = 1, \quad N_1 - M_1 = z, \quad M_2 + N_2 = 1 + z^2.$$

Thus

$$\begin{aligned} M_2 + N_2 &= (M_0 + N_0)^2 + (M_1 - N_1)^2 \\ &= (M_0 - M_1)^2 + (N_0 - N_1)^2 + 2(M_0N_0 - M_1N_1) + 2(M_0M_1 + N_0N_1) \\ &\geq 2(M_0M_1 + N_0N_1), \end{aligned}$$

where the last inequality holds since, by Lemma 22 (d), we have $M_1 \leq \frac{M_0^2}{g(z)}\frac{1+s}{1+2s} \leq M_0^2 \leq M_0$ and $N_1 \leq \frac{N_0^2}{g(z)}\frac{1+s}{1+2s} \leq N_0^2 \leq N_0$.

On the other hand, by Lemma 22 (c) (d),

$$M_2 \leq \frac{2M_1^2}{M_0}\frac{1+2s}{1+3s} \leq \frac{2M_1M_0}{g(z)}\frac{1+s}{1+3s} < 2M_1M_0,$$

and similarly, $N_2 < 2N_1N_0$. That means

$$M_2 + N_2 < 2(M_0M_1 + N_0N_1),$$

and we obtain a contradiction.

(b) The proof is by Lemma 22 (b) which lower bounds $g(0)$ by the second order moment of g, which is 1 according to isotropicity.

Specifically, without lose of generality, assume that $g(x)$ is monotone decreasing for $x \geq 0$ (otherwise consider $g(-x)$, since function g is uni-modal). Define g_0 as the restriction of g to the non-negative semi-line. Then by Lemma 22 (b), we have

$$M_1'(g_0)^3 \leq M_0'(g_0)^2 M_3'(g_0),$$

which implies

$$g(0) \geq \sqrt{\frac{M_0(g_0)^3}{3M_2(g_0)}}.$$

Note that $M_2(g_0) \leq M_2(g) = 1$, and by Lemma 6,

$$M_0(g_0) = \int_0^\infty g(t)dt = \Pr[X \geq \mathbb{E}X] \geq (1+\gamma)^{-1/\gamma}.$$

Thus we have

$$g(0) \geq \sqrt{\frac{1}{3(1+\gamma)^{3/\gamma}}},$$

where $\gamma = s/(1+s)$. □

H Proof of Theorem 9

Theorem 9 (restated) *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an isotropic s-concave density.*

(a) Let $d = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$, where $\beta = \frac{s}{1+(n-1)s}$ and $\gamma = \frac{\beta}{1+\beta}$. For any $u \in \mathbb{R}^n$ such that $\|u\| \leq d$, we have $f(u) \geq \left(\frac{\|u\|}{d}((2-2^{-(n+1)s})^{-1} - 1) + 1\right)^{1/s} f(0)$.

(b) $f(x) \le f(0)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s}$ *for every* x *(*$s \ge -\frac{1}{2n+3}$*).*

(c) There exists an $x \in \mathbb{R}^n$ *such that* $f(x) > (4e\pi)^{-n/2}$*.*

(d) We have $(4e\pi)^{-n/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1/s} < f(0) \le (2-2^{-(n+1)s})^{1/s}\frac{n\Gamma(n/2)}{2\pi^{n/2}d^n}$.

(e) $f(x) \le (2-2^{-(n+1)s})^{1/s}\frac{n\Gamma(n/2)}{2\pi^{n/2}d^n}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s}$ *for every* x*.*

(f) For any line ℓ *through the origin,* $\int_\ell f \le (2-2^{-ns})^{1/s}\frac{(n-1)\Gamma((n-1)/2)}{2\pi^{(n-1)/2}d^{n-1}}$.

Proof. (a) Formally, suppose that the conclusion does not hold true, i.e., there is a point u such that $\|u\| = t \le d$ and $f(u) < \left(\frac{t}{d}((2-2^{-(n+1)s})^{-1}-1)+1\right)^{1/s} f(0)$. Define $v = \frac{d}{t}u$ and note that $0 \le \frac{t}{d} \le 1$. Therefore, by the s-concavity of f, we have

$$f(u) = f\left(\frac{t}{d}v + \left(1-\frac{t}{d}\right)0\right) \ge \left[\frac{t}{d}f(v)^s + \left(1-\frac{t}{d}\right)f(0)^s\right]^{1/s},$$

which together with $f(u) < \left(\frac{t}{d}((2-2^{-(n+1)s})^{-1}-1)+1\right)^{1/s} f(0)$ implies $f(v) < (2-2^{-(n+1)s})^{-1/s}f(0)$. Let H be a hyperplane supporting the convex set $\{x \in \mathbb{R}^n : f(x) \ge f(v)\}$ through the point v (the convexity follows from the s-concavity of f). Define an orthogonal coordinate system in which the hyperplane is parallel to coordinate plane so that it can be represented as $x_1 = a$ for some $0 < a \le d$. Thus $f(x) < (2-2^{-(n+1)s})^{-1/s}f(0)$ for any x such that $x_1 \ge a$. We will prove that this implies that the 1-dimensional marginal is not flat.

Denote by g the first marginal of the n-dimensional function f. Then g is isotropic and $\beta = \frac{s}{1+(n-1)s}$-concave by Theorem 3, and $g(x) \le \frac{1+\beta}{1+3\beta}$ for all x by Lemma 7 (a). We prove that

$$g(2b) < \frac{g(b)}{4}$$

for any $b \ge a$, which means that the 1-dimensional function is not flat. To see this, by the s-concavity of function f, we have that, for every x such that $x_1 \ge a$,

$$f(2x)^s \ge 2f(x)^s - f(0)^s > 2^{-(n+1)s}f(x)^s.$$

Namely, $f(2x) < 2^{-(n+1)}f(x)$. Hence

$$g(2b) = \int_{(x_1=2b)} f(x)dx_2...dx_n < 2^{-(n+1)}2^{n-1}\int_{(x_1=b)} f(x)dx_2...dx_n = \frac{g(b)}{4}.$$

So

$$\int_a^\infty g(y)dy = \int_a^{2a} g(y)dy + \int_{2a}^\infty g(y)dy < \int_a^{2a} g(y)dy + \frac{1}{2}\int_a^\infty g(y)dy.$$

Namely, by Lemma 7 (a),

$$\int_a^\infty g(y)dy < 2\int_a^{2a} g(y)dy \le 2a\frac{1+\beta}{1+3\beta}.$$

So

$$\int_0^\infty g(y)dy = \int_0^a g(y)dy + \int_a^\infty g(y)dy < 3a\frac{1+\beta}{1+3\beta} \le 3d\frac{1+\beta}{1+3\beta} = (1+\gamma)^{-1/\gamma},$$

which leads to a contradiction with Lemma 6.

(b) If $f(w) \le f(0)$ for every w, then the conclusion holds true. Otherwise, let w be the point such that $f(w) > f(0)$. Let H_0 be the hyperplane through 0 which supports the convex set $\{x \in \mathbb{R}^n : f(x) \ge f(0)\}$. By defining an orthogonal system, we may set H_0 as the hyperplane

$x_1 = 0$, and so $f(x) \le f(0)$ for any x such that $x_1 = 0$. Define g, which is a $\beta = \frac{s}{1+(n-1)s}$-concave function, as the first marginal of function f. Denote by H_t the hyperplane $x_1 = t$. Without loss of generality, we assume that $w \in H_b$ with $b > 0$.

Let x be any point on H_0 and x' be the intersection between line segment $[x, w]$ and $H_{b/2}$. Then by the s-concavity of f and $f(x) \le f(0)$ for $x \in H_0$, we have

$$f(x') \ge \left[\frac{1}{2}f(x)^s + \frac{1}{2}f(w)^s\right]^{1/s} \ge \left(\frac{1}{2}\right)^{1/s} f(x)\left[1 + \left(\frac{f(w)}{f(0)}\right)^s\right]^{1/s}.$$

Thus

$$g(b/2) = \int_{(x_1 = b/2)} f(x)dx_2...dx_n \ge \frac{1}{2^{n-1+1/s}}\left[1 + \left(\frac{f(w)}{f(0)}\right)^s\right]^{1/s} g(0).$$

By Lemma 7 (a) (b), we have

$$\frac{1+\beta}{1+3\beta} \ge g(b/2) \ge \frac{1}{2^{n-1+1/s}}\left[1 + \left(\frac{f(w)}{f(0)}\right)^s\right]^{1/s} \sqrt{\frac{1}{3(1+\gamma)^{3/\gamma}}},$$

where $\gamma = \beta/(1+\beta)$. Note that $s \ge -\frac{1}{2n+3}$ implies $\frac{1+\beta}{1+3\beta}2^{n-1+1/s}\sqrt{3(1+\gamma)^{3/\gamma}} < 1$. So

$$f(w) \le f(0)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s}.$$

(c) The proof of Part (c) follows from [49].

(d) The proof of lower bound follows from Parts (b) and (c).

For the upper bound, by Part (a), we have

$$1 = \int_{\mathbb{R}^n} f(x)dx \ge \int_{\|x\| \le d} f(x)dx \ge d^n \text{vol}(B_{n-1})(2 - 2^{-(n+1)s})^{-1/s} f(0),$$

where $\text{vol}(B_{n-1})$ represents the volume of $n-1$-dimensional unit ball. So

$$f(0) \le \frac{(2 - 2^{-(n+1)s})^{1/s}}{d^n \text{vol}(B_{n-1})} = (2 - 2^{-(n+1)s})^{1/s}\frac{n\Gamma(n/2)}{2\pi^{n/2}d^n}.$$

(e) The proof of (e) follows from Parts (b) and (d).

(f) Define an orthogonal coordinate system in which ℓ is the x_n-axis. Let h be the marginal of function f over first $n-1$ variables, namely,

$$h(x_1, ..., x_{n-1}) = \int f(x_1, ..., x_{n-1}, x_n)dx_n.$$

Then

$$\int_\ell f = h(0) \le (2 - 2^{-ns})^{1/s}\frac{(n-1)\Gamma((n-1)/2)}{2\pi^{(n-1)/2}d^{n-1}}.$$

□

I Proof of Lemma 10

Lemma 10 (restated) *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be an isotropic s-concave density. Then $f(x) \le \beta_1(n,s)(1 - s\beta_2(n,s)\|x\|)^{1/s}$ for every $x \in \mathbb{R}^n$, where*

$$\beta_1(n,s) = (2 - 2^{-(n+1)s})^{\frac{1}{s}}\frac{1}{2\pi^{n/2}d^n}(1-s)^{-\frac{1}{s}}n\Gamma(n/2)\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{1/s},$$

and

$$\beta_2(n,s) = \frac{2\pi^{(n-1)/2}d^{n-1}}{(n-1)\Gamma((n-1)/2)}(2 - 2^{-ns})^{-1/s}\frac{[(a + (1-s)\beta_1(n,s)^s)^{1+1/s} - a^{1+1/s}]s}{\beta_1(n,s)^s(1+s)(1-s)},$$

$d = (1+\gamma)^{-\frac{1}{\gamma}}\frac{1+3\beta}{3+3\beta}$, $\beta = \frac{s}{1+(n-1)s}$, $\gamma = \frac{\beta}{1+\beta}$, $a = (4e\pi)^{-\frac{ns}{2}}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1}$.

Proof. We first note that when $\|x\| \le 1/\beta_2$, $\beta_1(1-s\beta_2\|x\|)^{1/s} \ge \beta_1(1-s)^{1/s} \ge f(x)$ by Theorem 9 (e). So the conclusion holds.

We now assume that there is a point v such that $\|v\| > 1/\beta_2$ but $f(v) > \beta_1(1-s\beta_2\|v\|)^{1/s}$. Denote by $[0,v]$ the line segment between the origin 0 and the point v, and let ℓ be the line through v and 0. We will prove that

$$\int_\ell f > (2-2^{-ns})^{1/s}\frac{(n-1)\Gamma((n-1)/2)}{2\pi^{(n-1)/2}d^{n-1}},$$

which leads to a contradiction with Theorem 9 (f). Let x be the convex combination of points 0 and v, i.e., $x = (1-\|x\|/\|v\|)0 + (\|x\|/\|v\|)v$, where $0 \le \|x\| \le \|v\|$. Then by the s-concavity of f and Theorem 9 (d),

$$\begin{aligned}
f(x) &\ge \left[\left(1-\frac{\|x\|}{\|v\|}\right)f(0)^s + \frac{\|x\|}{\|v\|}f(v)^s\right]^{1/s} \\
&> \left[f(0)^s + \frac{\|x\|}{\|v\|}f(v)^s\right]^{1/s} \\
&> \left[f(0)^s + \frac{\|x\|}{\|v\|}\beta_1^s - s\beta_1^s\beta_2\|x\|\right]^{1/s} \\
&> \left[f(0)^s + (1-s)\beta_1^s\beta_2\|x\|\right]^{1/s} \\
&> \left\{(4e\pi)^{-ns/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1} + (1-s)\beta_1^s\beta_2\|x\|\right\}^{1/s}.
\end{aligned}$$

Thus

$$\begin{aligned}
\int_\ell f \ge \int_{[0,v]} f &= \int_0^{\|v\|}\left\{(4e\pi)^{-ns/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1} + (1-s)\beta_1^s\beta_2 r\right\}^{1/s} dr \\
&> \int_0^{1/\beta_2}\left\{(4e\pi)^{-ns/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{n-1+1/s}\right)^s - 1\right]^{-1} + (1-s)\beta_1^s\beta_2 r\right\}^{1/s} dr \\
&\ge (2-2^{-ns})^{1/s}\frac{(n-1)\Gamma((n-1)/2)}{2\pi^{(n-1)/2}d^{n-1}}.
\end{aligned}$$

So the proof is completed. □

J Proof of Theorem 11

Theorem 11 (restated) *Let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$. Denote by $f_3(s,n) = 2(1+ns)/(1+(n+2)s)$. Then for any unit vector w,*

$$\Pr_{x\sim\mathcal{D}}[|w\cdot x| \le t] \le f_3(s,n)t. \tag{3}$$

Moreover, if $t \le d = \left(\frac{1+2\gamma}{1+\gamma}\right)^{-\frac{1+\gamma}{\gamma}}\frac{1+3\gamma}{3+3\gamma}$ where $\gamma = \frac{s}{1+(n-1)s}$, then

$$\Pr_{x\sim\mathcal{D}}[|w\cdot x| \le t] > f_2(s,n)t, \tag{4}$$

where $f_2(s,n) = 2(2-2^{-2\gamma})^{-1/\gamma}(4e\pi)^{-1/2}\left(2\left(\frac{1+\gamma}{1+3\gamma}\sqrt{3}\left(\frac{1+2\gamma}{1+\gamma}\right)^{\frac{3+3\gamma}{2\gamma}}\right)^\gamma - 1\right)^{-1/\gamma}$.

Proof. Define an orthogonal coordinate system in which w is an axis. Then the distribution of $w \cdot x$ is equal to the first marginal of the distribution $\mathcal{D}$, with isotropic $\gamma = \frac{s}{1+(n-1)s}$-concave density g by Theorem 3. According to the upper bound given by Lemma 7 (a),

$$\Pr_{x\sim\mathcal{D}}[|w \cdot x| \le t] = \int_{-t}^{t} g(x)dx \le \frac{1+\gamma}{1+3\gamma}\int_{-t}^{t} dx = 2\frac{1+ns}{1+(n+2)s}t.$$

We now prove the later part of the theorem by a similar argument. By Theorem 9 (a) (d), for 1-dimensional γ-concave density $f(u)$ and $\|u\| \le d$, we have

$$\begin{aligned} f(u) &\ge (2-2^{-2\gamma})^{-1/\gamma} f(0) \\ &> (2-2^{-2\gamma})^{-1/\gamma}(4e\pi)^{-1/2}\left(2\left(\frac{1+\gamma}{1+3\gamma}\sqrt{3}\left(\frac{1+2\gamma}{1+\gamma}\right)^{\frac{3+3\gamma}{2\gamma}}\right)^{\gamma} - 1\right)^{-1/\gamma} \\ &\triangleq \frac{f_2(s,n)}{2}. \end{aligned}$$

Therefore,

$$\Pr_{x\sim\mathcal{D}}[|w \cdot x| \le t] = \int_{-t}^{t} g(x)dx > \frac{f_2(s,n)}{2}\int_{-t}^{t} dx = f_2(s,n)t.$$

□

K Proof of Theorem 12

Theorem 12 (restated) *Assume $\mathcal{D}$ is an isotropic s-concave distribution in $\mathbb{R}^n$. Then for any two unit vectors u and v in $\mathbb{R}^n$, we have $d_{\mathcal{D}}(u,v) = \Pr_{x\sim\mathcal{D}}[\text{sign}(u\cdot x) \ne \text{sign}(v\cdot x)] \ge f_1(s,n)\theta(u,v)$, where $f_1(s,n) = c(2-2^{-3\alpha})^{-\frac{1}{\alpha}}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{1+1/\alpha}\right)^{\alpha} - 1\right]^{-\frac{1}{\alpha}}(1+\gamma)^{-2/\gamma}\left(\frac{1+3\beta}{3+3\beta}\right)^2$, c is an absolute constant, $\alpha = \frac{s}{1+(n-2)s}$, $\beta = \frac{s}{1+(n-1)s}$, $\gamma = \frac{s}{1+ns}$.*

Proof. Consider the 2-dimensional space spanned by vectors u and v, and let $\mathcal{D}_2$ be the marginal of distribution $\mathcal{D}$ over that space. Since $d_{\mathcal{D}}(u,v) = d_{\mathcal{D}_2}(u',v')$, where u' and v' are projection of u and v, respectively, we only need to focus on the marginal distribution $\mathcal{D}_2$, which has an α-concave density according to Theorem 3, and is isotropic according to Theorem 3.

Let A be the disagreement region of u and v intersected with the ball of radius $d = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$ in $\mathbb{R}^2$. By Theorem 9 (a) and (d), $d_{\mathcal{D}}(u,v) = d_{\mathcal{D}_2}(u',v') \ge \text{vol}(A)\inf_{x\in A} p(x) \ge f_1(s,n)\theta(u',v') = f_1(s,n)\theta(u,v)$, where $p(x)$ is the density of distribution $\mathcal{D}_2$. □

L Proof of Theorem 13

Theorem 13 (restated) *Let u and v be two vectors in $\mathbb{R}^n$ and assume that $\theta(u,v) = \theta < \pi/2$. Let $\mathcal{D}$ be an isotropic s-concave distribution. Then for any absolute constant $c_1 > 0$ and any function $f_1(s,n) > 0$, there exists a function $f_4(s,n) > 0$ such that*

$$\Pr_{x\sim\mathcal{D}}[\text{sign}(u\cdot x) \ne \text{sign}(v\cdot x) \text{ and } |v\cdot x| \ge f_4(s,n)\theta] \le c_1 f_1(s,n)\theta,$$

where $f_4(s,n) = \frac{4\beta_1(2,\alpha)B(-1/\alpha-3,3)}{-c_1 f_1(s,n)\alpha^3\beta_2(2,\alpha)^3}$, $B(\cdot,\cdot)$ is the beta function, $\alpha = s/(1+(n-2)s)$, $\beta_1(2,\alpha)$ and $\beta_2(2,\alpha)$ are given by Lemma 10.

Proof. Let E be the event that we want to bound. Theorem 3 implies that, without loss of generality, we can focus on the case when $n = 2$. Then the projected distribution $\mathcal{D}_2$ has an α-concave density, where $\alpha = \frac{s}{1+(n-2)s}$.

We first claim that each member x of E satisfies $\|x\| \ge f_4$. To see this, without loss of generality, we assume that $v \cdot x$ is positive. Then for any $x \in E$, $u \cdot x < 0$, i.e., $\theta(u,x) \ge \pi/2$. Hence

$\theta(x,v) \geq \theta(u,x) - \theta(u,v) \geq \pi/2 - \theta$. Since $|v \cdot x| \geq f_4\theta$ implies $\|x\| \cos\theta(v,x) \geq f_4\theta$, we must have $\|x\| \cos(\pi/2-\theta) \geq f_4\theta$, namely, $\|x\| \geq f_4\theta/\sin(\theta) \geq f_4$. Let $\mathsf{ball}(r)$ denote the ball of radius r centered at the origin. This implies that

$$\Pr[E] = \sum_{i=1}^{\infty} \Pr[E \cap (\mathsf{ball}((i+1)f_4) - \mathsf{ball}(if_4))].$$

Denote by $f(x_1,x_2)$ the α-concave density function of $\mathcal{D}_2$. For any term $i \geq 1$, by Lemma 10, we have

$$\begin{aligned}
&\Pr[E \cap (\mathsf{ball}((i+1)f_4) - \mathsf{ball}(if_4))] \\
&= \int_{(x_1,x_2)\in \mathsf{ball}((i+1)f_4)-\mathsf{ball}(if_4)} 1_E(x_1,x_2) f(x_1,x_2) dx_1 dx_2 \\
&\leq \int_{(x_1,x_2)\in \mathsf{ball}((i+1)f_4)-\mathsf{ball}(if_4)} 1_E(x_1,x_2) \beta_1(2,\alpha)(1-\alpha\beta_2(2,\alpha)\|x\|)^{1/\alpha} dx_1 dx_2 \\
&\leq \beta_1(2,\alpha)\left(1-\alpha\beta_2(2,\alpha) i f_4\right)^{1/\alpha} \int_{(x_1,x_2)\in \mathsf{ball}((i+1)f_4)-\mathsf{ball}(if_4)} 1_E(x_1,x_2) dx_1 dx_2 \\
&\leq \beta_1(2,\alpha)\left(1-\alpha\beta_2(2,\alpha) i f_4\right)^{1/\alpha} \int_{(x_1,x_2)\in \mathsf{ball}((i+1)f_4)} 1_E(x_1,x_2) dx_1 dx_2.
\end{aligned}$$

Denote by B_1 the unit ball in $\mathbb{R}^2$. Notice that

$$\int_{(x_1,x_2)\in \mathsf{ball}((i+1)f_4)} 1_E(x_1,x_2) dx_1 dx_2 \leq \mathrm{vol}(B_1)(i+1)^2 f_4^2 \theta/\pi.$$

Thus

$$\begin{aligned}
\Pr[E] &\leq \sum_{i=1}^{\infty} \beta_1(2,\alpha)\left(1-\alpha\beta_2(2,\alpha) i f_4\right)^{1/\alpha} \mathrm{vol}(B_1)(i+1)^2 f_4^2\theta/\pi \\
&\leq \frac{4f_4^2}{\pi}\mathrm{vol}(B_1)\beta_1(2,\alpha)\theta \sum_{i=1}^{\infty}(1-\alpha\beta_2(2,\alpha) i f_4)^{1/\alpha} i^2 \\
&\leq \frac{4f_4^2}{\pi}\mathrm{vol}(B_1)\beta_1(2,\alpha)\theta \int_0^{\infty} x^2(1-\alpha\beta_2(2,\alpha) f_4 x)^{1/\alpha} dx \\
&= \frac{4f_4^2}{\pi}\mathrm{vol}(B_1)\beta_1(2,\alpha)\frac{B(-1/\alpha-3,3)}{(-\alpha\beta_2(2,\alpha)f_4)^3} \times \theta \\
&= 4\beta_1(2,\alpha)\frac{B(-1/\alpha-3,3)}{-\alpha^3\beta_2(2,\alpha)^3 f_4} \times \theta.
\end{aligned}$$

Choosing $f_4(s,n) = \frac{4\beta_1(2,\alpha)B(-1/\alpha-3,3)}{-c_1 f_1(s,n)\alpha^3\beta_2(2,\alpha)^3}$, the proof is completed. □

M Proof of Theorem 14

Before proceeding, we first prove the following lemma which is critical to the proof of Theorem 14.

Lemma 23. *For d given by Theorem 9 (a), there exist such that for any isotropic s-concave distribution $\mathcal{D}$, for any a such that $\|a\| \leq 1$ and $\|u-a\| \leq r$, for any $0 < t \leq d$, and for any $K \geq 4$, we have*

$$\Pr_{X\sim\mathcal{D}_{u,t}}(|a\cdot x| > K\sqrt{r^2+t^2}) \leq \frac{4\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}\left(1 - c\eta\beta_2(2,\eta)K\sqrt{1+\left(\frac{t}{r}\right)^2}\right)^{\frac{\eta+1}{\eta}},$$

where $\beta_1(2,\eta)$, $\beta_2(2,\eta)$, and $Q(\gamma)$, are given by Lemma 10 and Theorem 11, respectively, $\eta = \frac{s}{1+(n-2)s}$, and c is an absolute constant.

Proof. Without loss of generality, we assume that $u = (1, 0, ..., 0)$. Let $a' = (0, a_2, ..., a_d)$ and $X = (x_1, x_2, ..., x_d) \sim \mathcal{D}_{u,t}$. So the probability that we want to bound is

$$\Pr_{X\sim\mathcal{D}_{u,t}}(|a\cdot x| > K\sqrt{r^2+t^2}) = \frac{\Pr_{x\sim\mathcal{D}}(|a\cdot x| > K\sqrt{r^2+t^2},\ |x_1|\le t)}{\Pr_{x\sim\mathcal{D}}(|x_1|\le t)}.$$

According to Theorem 11, there is a function $Q(\gamma)$ such that the denominator obeys the following lower bound when $t \le d$:

$$\Pr_{X\sim\mathcal{D}}(|x_1|\le t) \ge f_2(s,n)t.$$

So the remainder of the proof is to bound the numerator. Note that we have

$$\begin{aligned}\Pr_{x\sim\mathcal{D}}(|a\cdot x| > K\sqrt{r^2+t^2},\ |x_1|\le t) &\le \Pr_{x\sim\mathcal{D}}(|a'\cdot x| > K\sqrt{r^2+t^2} - t,\ |x_1|\le t)\\ &\le \Pr_{x\sim\mathcal{D}}(|a'\cdot x| > (K-1)\sqrt{r^2+t^2},\ |x_1|\le t).\end{aligned}$$

Denote by $a'' = \frac{a'}{\|a'\|}$. Define random variable Y as $a''\cdot x$ and Z as x_1 where $x\sim\mathcal{D}$. Then the joint distribution of Y and Z is isotropic β-concave with $\eta = \frac{s}{1+(n-2)s}$. Let $f(y,z)$ be the density of such a distribution. Then we can bound the numerator by

$$\begin{aligned}4\Pr_{x\sim\mathcal{D}}(a'\cdot x > (K-1)\sqrt{r^2+t^2},\ 0\le x_1\le t) &= 4\Pr_{X\sim\mathcal{D}}(a''\cdot x > \frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|},\ 0\le x_1\le t)\\ &\le 4\int_0^t\int_{\frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|}}^{\infty} f(y,z)dydz.\end{aligned}$$

By Lemma 10, we note that

$$f(y,z) \le \beta_1(2,\eta)(1-\eta\beta_2(2,\eta)\sqrt{y^2+z^2})^{1/\eta}.$$

Therefore, the numerator can be upper bounded by

$$\begin{aligned}&4\beta_1(2,\eta)\int_0^t\int_{\frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|}}^{\infty}(1-\eta\beta_2(2,\eta)\sqrt{y^2+z^2})^{1/\eta}dydz\\ &\le 4\beta_1(2,\eta)\int_0^t\int_{\frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|}}^{\infty}(1-\eta\beta_2(2,\eta)y)^{1/\eta}dydz\\ &= 4t\beta_1(2,\eta)\int_{\frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|}}^{\infty}(1-\eta\beta_2(2,\eta)y)^{1/\eta}dy\\ &= \frac{4t\beta_1(2,\eta)}{\beta_2(2,\eta)}\frac{1}{\eta+1}\left(1-\eta\beta_2(2,\eta)\frac{(K-1)\sqrt{r^2+t^2}}{\|a'\|}\right)^{\frac{\eta+1}{\eta}}.\end{aligned} \tag{5}$$

Note that $\|a'\| \le r$. Finally, we have

$$\begin{aligned}\Pr_{X\sim\mathcal{D}_{u,t}}(|a\cdot x| > K\sqrt{r^2+t^2}) &\le \frac{4\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}\left(1-\eta\beta_2(2,\eta)\frac{(K-1)\sqrt{r^2+t^2}}{r}\right)^{\frac{\eta+1}{\eta}}\\ &\le \frac{4\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}\left(1-c\eta\beta_2(2,\eta)\frac{K\sqrt{r^2+t^2}}{r}\right)^{\frac{\eta+1}{\eta}},\end{aligned}$$

for an absolute constant c. □

Theorem 14 (restated) *Assume that $\mathcal{D}$ is isotropic s-concave. For d given by Theorem 9 (a), there is an absolute C_0 such that for all $0 < t \le d$ and for all a such that $\|u-a\| \le r$ and $\|a\| \le 1$, $\mathbb{E}_{X\sim\mathcal{D}_{u,t}}[(a\cdot x)^2] \le f_5(s,n)(r^2+t^2)$, where $f_5(s,n) = 16 + C_0\frac{8\beta_1(2,\eta)B(-1/\eta-3,2)}{f_2(s,n)\beta_2(2,\eta)^3(\eta+1)\eta^2}$, $(\beta_1(2,\eta),\beta_2(2,\eta))$ and $f_2(s,n)$ are given by Lemma 10 and Theorem 11, respectively, and $\eta = \frac{s}{1+(n-2)s}$.*

Proof. Denote by $z = \sqrt{r^2 + t^2}$. Then we have

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{D}_{u,t}}[(a\cdot x)^2] &= \int_0^\infty \Pr_{x\sim\mathcal{D}_{u,t}}[(a\cdot x)^2 \geq z]dz \\
&\leq 16z^2 + \int_{16z^2}^\infty \Pr_{x\sim\mathcal{D}_{u,t}}[(a\cdot x)^2 \geq z]dz \\
&\leq 16z^2 + \frac{4\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}\int_0^\infty \left(1-\eta\beta_2(2,\eta)\frac{c\sqrt{z}}{r}\right)^{\frac{\eta+1}{\eta}} dz \\
&= 16z^2 + \frac{8\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}\int_0^\infty y\left(1-\eta\beta_2(2,\eta)\frac{cy}{r}\right)^{\frac{\eta+1}{\eta}} dy \qquad (6) \\
&= 16z^2 + \frac{8\beta_1(2,\eta)}{f_2(s,n)\beta_2(2,\eta)}\frac{1}{\eta+1}C_0B(-1/\eta-3,2)\frac{r^2}{\eta^2\beta_2(2,\eta)^2} \\
&= \left(16 + C_0\frac{8\beta_1(2,\eta)B(-1/\eta-3,2)}{f_2(s,n)\beta_2(2,\eta)^3(\eta+1)\eta^2}\right)r^2 + 16t^2 \\
&\leq \left(16 + C_0\frac{8\beta_1(2,\eta)B(-1/\eta-3,2)}{f_2(s,n)\beta_2(2,\eta)^3(\eta+1)\eta^2}\right)(r^2+t^2),
\end{aligned}$$

where c, C_0 are absolute constants. □

N Proof of Theorem 15

Theorem 15 (restated) *In the realizable case, let $\mathcal{D}$ be an isotropic s-concave distribution in $\mathbb{R}^n$. There exist constants C and c such that for any $0 < \epsilon < 1/4$ and $\delta > 0$, Algorithm 2 with $b_k = \min\{\Theta(2^{-k}f_4f_1^{-1}), d\}$ and $m_k = C\left(\frac{f_3b_{k-1}}{2^{-k}}\left(n\log\frac{f_3b_{k-1}}{2^{-k}} + \log\frac{1+s-k}{\delta}\right)\right)$, after $T = \lceil\log\frac{1}{c\epsilon}\rceil$ iterations, outputs a linear separator of error at most ϵ with probability at least $1-\delta$.*

Proof. We will show by induction that for all $k \leq s$, with probability at least $1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$, any w that is consistent with the examples in $W(k)$, e.g. w_k, has error at most $c2^{-k}$.

The case of $k = 1$ follows from the VC theory (Theorem 30). Assume now that the claim is true for $k-1$. We now consider the kth iteration. Denote by $S_{k-1} = \{x : |w_{k-1}\cdot x| \leq b_{k-1}\}$ and $\bar{S}_{k-1} = \{x : |w_{k-1}\cdot x| > b_{k-1}\}$. By the induction hypothesis, with probability at least $1 - \frac{\delta}{2}\sum_{i<k-1}\frac{1}{(1+s-i)^2}$, any w that is consistent with $W(k-1)$, including w_{k-1}, has error at most $c2^{-(k-1)}$. For such a hypothesis w and w_{k-1}, by Theorem 12, we have $\theta(w, w^*) \leq cf_1^{-1}2^{-(k-1)}$ and $\theta(w_{k-1}, w^*) \leq cf_1^{-1}2^{-(k-1)}$. Thus $\theta(w_{k-1}, w) \leq \theta(w_{k-1}, w^*) + \theta(w^*, w) \leq 4cf_1^{-1}2^{-k}$. So by Theorem 13, there is a choice of band width $b_{k-1} = \min\{\Theta(f_4f_1^{-1}2^{-k}), d\}$ such that $\Pr(\text{sign}(w\cdot x) \neq \text{sign}(w_{k-1}\cdot x), x \in \bar{S}_{k-1}) \leq \frac{c2^{-k}}{4}$ and $\Pr[\text{sign}(w_{k-1}\cdot x) \neq \text{sign}(w^*\cdot x), x \in \bar{S}_{k-1}] \leq \frac{c2^{-k}}{4}$. Therefore, $\Pr[\text{sign}(w\cdot x) \neq \text{sign}(w^*\cdot x), x \in \bar{S}_{k-1}] \leq \frac{c2^{-k}}{2}$.

We now consider the case when $x \in S_{k-1}$. By Algorithm 2, we label m_k data points in S_{k-1} at the $(k-1)$th iteration. So according to the VC theory (Theorem 30), with probability at least $1 - \delta/(4(1+s-k)^2)$, for all w that is consistent with the examples in $W(k)$, $\mathsf{err}(w|S_{k-1}) = \Pr[\text{sign}(w\cdot x) \neq \text{sign}(w^*\cdot x)|x \in S_{k-1}] \leq \frac{c2^{-k}}{2b_{k-1}f_3}$. Finally, note that Theorem 11 implies that $\Pr(S_{k-1}) \leq f_3b_{k-1}$. So we have $\mathsf{err}(w) = \Pr[\text{sign}(w\cdot x) \neq \text{sign}(w^*\cdot x), x \in \bar{S}_{k-1}] + \Pr[\text{sign}(w\cdot x) \neq \text{sign}(w^*\cdot x), x \in S_{k-1}] \leq \frac{c2^{-k}}{2} + \frac{c2^{-k}}{2b_{k-1}f_3} \times f_3b_{k-1} = c2^{-k}$. The proof is completed. □

O Proof of Theorem 16

Before proceeding, let $\ell_\tau(w, x, y) = \max\{0, 1 - y(w\cdot x)/\tau\}$, $\ell_\tau(w, T) = \frac{1}{|T|}\sum_{(x,y)\in T}\ell_\tau(w, x, y)$, and $L_\tau(w, \mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}\ell_\tau(w, x, \text{sign}(w^*\cdot x))$. Our analysis will involve the distribution $\mathcal{D}_{w,t}$ obtained by conditioning $\mathcal{D}$ on membership in the band, namely, the set $\{x : |w\cdot x| \leq t\}$.

Lemma 24. *$L_{\tau_k}(w^*, \mathcal{D}_{w_{k-1},b_{k-1}}) \le \kappa/6$, if $\kappa \ge \frac{6 f_3 \tau_k}{f_2 b_{k-1}}$ and $b_{k-1} \le d$.*

Proof. Note that $y(w^* \cdot x)$ cannot be negative on any clean example (x, y). So we have $\ell(w^*, x, y) = \max\{0, 1 - y(w^* \cdot x)/\tau_k\} \le 1$ and w^* pays a non-zero hinge loss only inside the margin $\{x : |w^* \cdot x| \le \tau_k\}$. Thus $L_{\tau_k}(w^*, \mathcal{D}_{w_{k-1},b_{k-1}}) \le \Pr_{\mathcal{D}_{w_{k-1},b_{k-1}}}(|w^* \cdot x| \le \tau_k) = \Pr_{\mathcal{D}}(|w^* \cdot x| \le \tau_k, |w_{k-1} \cdot x| \le b_{k-1}) / \Pr_{\mathcal{D}}(|w_{k-1} \cdot x| \le b_{k-1})$. Notice that the numerator can be bounded by $\Pr_{\mathcal{D}}(|w^* \cdot x| \le \tau_k) \le f_3 \tau_k$ according to Theorem 11. As for the denominator, by Theorem 11 we have $\Pr_{\mathcal{D}}(|w_{k-1} \cdot x| \le b_{k-1}) \ge f_2 \min\{b_{k-1}, d\}$. So we have $L_{\tau_k}(w^*, \mathcal{D}_{w_{k-1},b_{k-1}}) \le f_3 \tau_k / (f_2 \min\{b_{k-1}, d\}) \le \kappa/6$. □

Let $\widetilde{\mathcal{P}}_k$ be the noisy distribution of (x, y) where $x \sim \mathcal{D}_{w_{k-1},b_{k-1}}$ and y obeys the adversarial noise model, and denote by $\mathcal{P}_k$ the clean distribution of (x, y) where $x \sim \mathcal{D}_{w_{k-1},b_{k-1}}$ and $y = \text{sign}(w^* \cdot x)$. The following key lemma bounds the distance of expected loss w.r.t. the distributions $\widetilde{\mathcal{P}}_k$ and $\mathcal{P}_k$.

Lemma 25. *There exists an absolute constant c such that for any $w \in \mathsf{ball}(w_{k-1}, r_k)$, we have that $\left|\mathbb{E}_{(x,y)\sim\mathcal{P}_k} \ell(w, x, y) - \mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k} \ell(w, x, y)\right| \le \frac{2}{\tau_k} \sqrt{\frac{\eta f_5 (r_k^2 + b_{k-1}^2)}{f_2 b_{k-1}}}$.*

Proof. Denote by N the set of noisy examples. Let $\widetilde{\mathcal{P}}$ be the noisy distribution of (x, y) where $x \sim \mathcal{D}$ and y obeys the adversarial noise model. We have

$$
\begin{aligned}
&\left|\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}[\ell_{\tau_k}(w^*, x, y)] - \mathbb{E}_{(x,y)\sim\mathcal{P}_k}[\ell_{\tau_k}(w^*, x, y)]\right| \\
&\quad\le \left|\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}\left[\mathbf{1}_{x\in N}\left(\ell_{\tau_k}(w^*, x, y) - \ell_{\tau_k}(w^*, x, \text{sign}(w^* \cdot x))\right)\right]\right| \\
&\quad\le 2\,\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}\left[\mathbf{1}_{x\in N}\left(\frac{|w^* \cdot x|}{\tau_k}\right)\right] \\
&\quad\le \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim\widetilde{\mathcal{P}}_k}[x \in N]} \times \sqrt{\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}[(w^* \cdot x)^2]} \qquad \text{(By Cauchy Schwarz)} \\
&\quad\le \frac{2}{\tau_k}\sqrt{\frac{\eta}{\Pr_{\widetilde{\mathcal{P}}}(|w_{k-1} \cdot x| \le b_{k-1})}} \times \sqrt{f_5(r_k^2 + b_{k-1}^2)} \qquad \text{(By Theorem 14)} \\
&\quad\le \frac{2}{\tau_k}\sqrt{\frac{\eta f_5 (r_k^2 + b_{k-1}^2)}{f_2 b_{k-1}}}. \qquad \text{(By Theorem 11)}
\end{aligned}
$$

□

Lemma 26. *Denote by W the samples drawn from the noisy distribution $\widetilde{\mathcal{P}}_k$ and suppose that $|W| = O\left(\frac{[b_{k-1}s + \tau_k(1+ns)\sqrt{n}[1-(\delta/(\sqrt{n}(k+k^2)))^{s/(1+ns)}] + \tau_k s]^2}{\kappa^2 \tau_k^2 s^2}\left(n + \log\frac{k+k^2}{\delta}\right)\right)$. Then with probability at least $1 - \frac{\delta}{k+k^2}$, for all $w \in \mathsf{ball}(w_{k-1}, r_k)$, we have*

$$\left|\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k} \ell(w, x, y) - \ell(w, W)\right| \le \kappa/16.$$

Proof. To establish the lemma, we apply some standard VC tools (Theorem 31). Note that the pseudo-dimension of $\{\ell(w, \cdot) : w \in \mathbb{R}^n\}$ is at most n [5]. To use Theorem 31, we first provide the upper bound on the loss. On one hand, note that

$$
\begin{aligned}
\ell(w, x, y) &\le 1 + \frac{|w \cdot x|}{\tau_k} \le 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|\|x\|}{\tau_k} \\
&\le 1 + \frac{b_{k-1} + \tau_k \|x\|}{\tau_k}.
\end{aligned}
$$

On the other hand, by Theorem 5 and the union bound, with probability at least $1 - \frac{\delta}{k+k^2}$, we have that $\max_{x\in W} \|x\| \le C\frac{(1+ns)\sqrt{n}}{s}\left\{1 - \left[\frac{\delta}{6(k+k^2)|W|}\right]^{s/(1+ns)}\right\}$, for an absolute constant C. The conclusion then follows from Theorem 31. □

Lemma 27. *Let* $k \leq \lceil \log(1/(c\epsilon)) \rceil$ *where* c *is an absolute constant. If* $\kappa = \max\left\{\frac{f_3\tau_k}{f_2 \min\{b_{k-1},d\}}, \frac{b_{k-1}\sqrt{f_5}}{\tau_k\sqrt{f_2}}\right\}$, $r_k \leq O(b_{k-1})$, $\eta \leq O(b_{k-1})$, $m_k = O\left(\frac{[b_{k-1}s+\tau_k(1+ns)\sqrt{n}[1-(\delta/(k+k^2))^{s/(1+ns)}]+\tau_k s]^2}{\kappa^2\tau_k^2 s^2}\left(n+\log\frac{k+k^2}{\delta}\right)\right)$, *and* $b_{k-1} \leq d$, *then with probability at least* $1-\frac{\delta}{k+k^2}$, *we have* $\mathsf{err}_{\mathcal{D}_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$.

Proof. With probability at least $1-\frac{\delta}{k+k^2}$, we have

$$\begin{aligned}
\mathsf{err}_{\mathcal{D}_{w_{k-1},b_{k-1}}}(w_k) &= \mathsf{err}_{\mathcal{D}_{w_{k-1},b_{k-1}}}(v_k) \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}_k}\ell(v_k,x,y) \\
&\leq \mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}\ell(v_k,x,y) + \frac{2}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} \qquad \text{(By Lemma 25)} \\
&\leq \ell(v_k,W) + \frac{2}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} + \frac{\kappa}{16} \qquad \text{(By Lemma 26)} \\
&\leq \ell(w^*,W) + \frac{4}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} + \frac{\kappa}{8} \qquad \text{(Since } \|v_k\| \geq 1/2\text{)} \\
&\leq \mathbb{E}_{(x,y)\sim\widetilde{\mathcal{P}}_k}\ell(w^*,x,y) + \frac{4}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} + \frac{\kappa}{4} \qquad \text{(By Lemma 26)} \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}_k}\ell(w^*,x,y) + \frac{6}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} + \frac{\kappa}{4} \qquad \text{(By Lemma 25)} \\
&\leq \frac{6}{\tau_k}\sqrt{\frac{\eta f_5(r_k^2+b_{k-1}^2)}{f_2 b_{k-1}}} + \frac{\kappa}{2} \qquad \text{(By Lemma 24)} \\
&\leq \kappa,
\end{aligned}$$

where the last inequality holds because $\kappa\tau_k\sqrt{\frac{f_2}{f_5}} \geq \Theta(b_{k-1})$, $r_k \leq O(b_{k-1})$, and $\eta \leq O(b_{k-1})$. □

Now we are ready to prove Theorem 16.

Theorem 16 (restated) *Let* $\mathcal{D}$ *be an isotropic* s*-concave distribution in* $\mathbb{R}^n$ *and the label* y *obeys the adversarial noise model. If the rate* η *of adversarial noise satisfies* $\eta < c_0\epsilon$ *for some absolute constant* c_0*, then there exists an absolute constant* c *such that for any* $0 < \epsilon < 1/4$ *and* $\delta > 0$*, Algorithm 1 with* $b_k = \min\{\Theta(2^{-k}f_4f_1^{-1}), d\}$, $\tau_k = \Theta\left(f_1^{-2}f_2^{-1/2}f_3f_4^2f_5^{1/2}2^{-(k-1)}\right)$, $r_k = \Theta(2^{-k}f_1^{-1})$, $m_k = O\left(\frac{[b_{k-1}s+\tau_k(1+ns)\sqrt{n}[1-(\delta/(k+k^2))^{s/(1+ns)}]+\tau_k s]^2}{\kappa^2\tau_k^2 s^2}\left(n+\log\frac{k+k^2}{\delta}\right)\right)$, *and* $\kappa = \max\left\{\frac{f_3\tau_k}{f_2\min\{b_{k-1},d\}}, \frac{b_{k-1}\sqrt{f_5}}{\tau_k\sqrt{f_2}}\right\}$, *after* $T = \lceil \log\frac{1}{c\epsilon} \rceil$ *iterations, outputs a linear separator* w_T *such that* $\Pr_{x\sim\mathcal{D}}[\text{sign}(w_T\cdot x) \neq \text{sign}(w^*\cdot x)] \leq \epsilon$ *with probability at least* $1-\delta$.

Proof. The case of $k=1$ is obvious. Assume now that the claim is true for $k-1$. We now consider the kth iteration. Denote by $S_{k-1} = \{x : |w_{k-1}\cdot x| \leq b_{k-1}\}$ and $\bar{S}_{k-1} = \{x : |w_{k-1}\cdot x| > b_{k-1}\}$. By the induction hypothesis, with probability at least $1-\frac{\delta}{2}\sum_{i<k-1}\frac{1}{(1+s-i)^2}$, w_{k-1} has error at most $c2^{-(k-1)}$. Then by Theorem 12, we have $\theta(w_{k-1},w^*) \leq cf_1^{-1}2^{-(k-1)}$. On the other hand, since $\|w_{k-1}\| = 1$ and $v_k \in B(w_{k-1},r_k)$, we have $\theta(w_{k-1},v_k) \leq r_k$. This in turn implies $\theta(w_{k-1},w_k) \leq 2^{-k}f_1^{-1}$. So by Theorem 13, there is a choice of band width $2b_{k-1} = O(f_4f_1^{-1}2^{-k})$ such that $\Pr(\text{sign}(w_k\cdot x) \neq \text{sign}(w_{k-1}\cdot x), x \in \bar{S}_{k-1}) \leq \frac{c2^{-k}}{4}$ and $\Pr[\text{sign}(w_{k-1}\cdot x) \neq \text{sign}(w^*\cdot x), x \in \bar{S}_{k-1}] \leq \frac{c2^{-k}}{4}$. Therefore, $\Pr[\text{sign}(w_k\cdot x) \neq \text{sign}(w^*\cdot x), x \in \bar{S}_{k-1}] \leq \frac{c2^{-k}}{2}$. Finally, note that Theorem 11 implies that $\Pr(S_{k-1}) \leq f_3b_{k-1}$. So we have $\mathsf{err}_{\mathcal{D}}(w_k) = \Pr[\text{sign}(w_k\cdot x) \neq$

$\text{sign}(w^* \cdot x), x \in \bar{S}_{k-1}] + \Pr[\text{sign}(w_k \cdot x) \neq \text{sign}(w^* \cdot x), x \in S_{k-1}] = \Pr[\text{sign}(w_k \cdot x) \neq \text{sign}(w^* \cdot x), x \in \bar{S}_{k-1}] + \mathsf{err}_{\mathcal{D}_{w_{k-1},b_{k-1}}}(w_k) \Pr(x \in S_{k-1}) \leq \frac{c2^{-k}}{2} + \kappa \times f_3 b_{k-1} \leq c2^{-k} = \epsilon$, where the penultimate inequality follows from Lemma 27. The proof is completed. □

P Proof of Theorem 17

Theorem 17 (restated) *Let $\mathcal{D}$ be an isotropic s-concave distribution over $\mathbb{R}^n$. Then for any $w^* \in \mathbb{R}^n$ and $r > 0$, the disagreement coefficient is $\Theta_{w^*,\mathcal{D}}(\epsilon) = O\left(\sqrt{n}\frac{(1+ns)^2}{s(1+(n+2)s)f_1(s,n)}(1-\epsilon^{s/(1+ns)})\right)$, where $f_1(s,n)$ is given by Theorem 12. In particular, when $s \to 0$ (a.k.a. log-concave), $\Theta_{w^*,\mathcal{D}}(\epsilon) = O(\sqrt{n}\log(1/\epsilon))$.*

Proof. Consider any unit w such that $d_{\mathcal{D}}(w, w^*) \leq r$. According to Theorem 12, we have $\|w - w^*\| < \theta(w, w^*) \leq d_{\mathcal{D}}(w, w^*)/f(s) \leq r/f(s)$. Thus for any x such that $\|x\| \leq O(\sqrt{n}\frac{1+ns}{s}(1-r^{s/(1+ns)}))$, we have $w \cdot x - w^* \cdot x \leq \|w - w^*\| \, \|x\| < r\sqrt{n}\frac{1+ns}{sf(s)}(1-r^{s/(1+ns)})$. So as soon as $|w^* \cdot x| \geq r\sqrt{n}\frac{1+ns}{sf(s)}(1-r^{s/(1+ns)})$, we will have $\text{sign}(w \cdot x) = \text{sign}(w^* \cdot x)$, i.e., w and w^* agree with each other. We now evaluate the probability. By Theorem 11, $\Pr_{x\sim\mathcal{D}}\left[|w^* \cdot x| \leq r\sqrt{n}\frac{1+ns}{sf(s)}(1-r^{s/(1+ns)})\right] \leq 2\frac{1+ns}{1+(n+2)s} r\sqrt{n}\frac{1+ns}{sf(s)}(1-r^{s/(1+ns)})$. Moreover, $\Pr_{x\sim\mathcal{D}}\left[\|x\| \geq c\sqrt{n}\frac{1+ns}{s}(1-r^{s/(1+ns)})\right] \leq Cr$ by Theorem 5. Thus

$$\begin{aligned}\text{cap}_{w^*,\mathcal{D}}(r) &\leq \frac{\Pr_{x\sim\mathcal{D}}\left[|w^* \cdot x| \leq r\sqrt{n}\frac{1+ns}{sf(s)}(1-r^{s/(1+ns)})\right]}{r} + \frac{\Pr_{x\sim\mathcal{D}}\left[\|x\| \geq c\sqrt{n}\frac{1+ns}{s}(1-r^{s/(1+ns)})\right]}{r} \\ &= O\left(\sqrt{n}\frac{(1+ns)^2}{s(1+(n+2)s)f(s)}(1-r^{s/(1+ns)})\right).\end{aligned}$$

Therefore, $\Theta_{w^*,\mathcal{D}}(\epsilon) = \sup_{r\geq\epsilon}[\text{cap}_{w^*,\mathcal{D}}(r)] = O\left(\sqrt{n}\frac{(1+ns)^2}{s(1+(n+2)s)f(s)}(1-\epsilon^{s/(1+ns)})\right)$. □

Q Proof of Theorem 18

Lemma 28. *Denote by R the intersections of three origin-centered halfspaces in $\mathbb{R}^n$. Suppose that the instance x in $\mathbb{R}^n$ is drawn from an isotropic s-concave distribution. Then $\Pr[x \in -R] \leq K \Pr[x \in R]$, where $K = \beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3}\frac{3+1/\kappa}{h(\kappa)d^{3+1/\kappa}}$, $\beta_1(3,\kappa)$, $\beta_2(3,\kappa)$, and $a(3,\kappa)$ are as in Lemma 10, $h(\kappa) = \left(\frac{1}{d}((2-2^{-4\kappa})^{-1}-1)+1\right)^{1/\kappa}(4e\pi)^{-3/2}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{2+1/\kappa}\right)^{\kappa}-1\right]^{-1/\kappa}$, $d = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$, $\beta = \frac{\kappa}{1+2\kappa}$, $\gamma = \frac{\kappa}{1+\kappa}$, and $\kappa = s/(1+(n-3)s)$.*

Proof. Let u_1, u_2, and u_3 be normals to the hyperplanes bounding the region R, namely $R = \{x \in \mathbb{R}^n : u_1 \cdot x \geq 0 \text{ and } u_2 \cdot x \geq 0 \text{ and } u_3 \cdot x \geq 0\}$. Denote by U the linear span of vectors u_1, u_2, and u_3, and let (e_1, e_2, e_3) be an orthogonal basis of U and $(e_1, e_2, e_3, ..., e_n)$ be an extension of basis (e_1, e_2, e_3) to $\mathbb{R}^n$. Represent the components of x, u_1, u_2, and u_3 in term of basis $(e_1, e_2, e_3, ..., e_n)$ as

$$\begin{aligned}x &= (x_1, x_2, x_3, x_4, ..., x_n), \\ u_1 &= (u_{1,1}, u_{1,2}, u_{1,3}, 0, ..., 0), \\ u_2 &= (u_{2,1}, u_{2,2}, u_{2,3}, 0, ..., 0), \\ u_3 &= (u_{3,1}, u_{3,2}, u_{3,3}, 0, ..., 0).\end{aligned}$$

Denote by $\mathsf{proj}_U(x) \triangleq (x_1, x_2, x_3)$ the projection of x onto subspace U, and let $\mathsf{proj}_U(R)$ be the projection of R onto U. Because the dot products of a point with normal vectors of R are all that is needed to determine the membership in R, we have

$$\begin{aligned}x \in R &\Leftrightarrow u_{j,1}x_1 + u_{j,2}x_2 + u_{j,3}x_3 \geq 0 \text{ for all } j \in \{1,2,3\} \\ &\Leftrightarrow \mathsf{proj}_U(x) \in \mathsf{proj}_U(R).\end{aligned} \tag{7}$$

Let f be the density of the isotropic s-concave distribution and g be the marginal density of f w.r.t. (x_1, x_2, x_3). Thus by (7),

$$\begin{aligned}\Pr[x \in R] &= \int \cdots \int_R f(x_1, x_2, x_3, x_4, ..., x_n) dx_1 ... dx_n \\ &= \int \int \int_{\mathsf{proj}_U(R)} g(x_1, x_2, x_3) dx_1 dx_2 dx_3.\end{aligned}$$

Note that f is isotropic and s-concave. So according to Theorem 3, g is isotropic and κ-concave with $\kappa = s/(1 + (n-3)s)$. We now use Theorem 9 and Lemma 10 to bound g. Specifically, let $u \triangleq (x_1, x_2, x_3)$. On one hand, according to Theorem 9 (a) and (d), for any $u \in \mathbb{R}^3$ such that $\|u\| \le d$,

$$\begin{aligned}g(u) &\ge \left(\frac{\|u\|}{d}((2 - 2^{-4\kappa})^{-1} - 1) + 1\right)^{1/\kappa} f(0) \\ &> \left(\frac{\|u\|}{d}((2 - 2^{-4\kappa})^{-1} - 1) + 1\right)^{1/\kappa} (4e\pi)^{-3/2} \left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}} 2^{2+1/\kappa}\right)^{\kappa} - 1\right]^{-1/\kappa} \\ &\triangleq \|u\|^{1/\kappa} h(\kappa),\end{aligned}$$

where $d = (1+\gamma)^{-1/\gamma} \frac{1+3\beta}{3+3\beta}$, $\beta = \frac{\kappa}{1+2\kappa}$, $\gamma = \frac{\kappa}{1+\kappa}$, and

$$h(\kappa) = \left(\frac{1}{d}((2 - 2^{-4\kappa})^{-1} - 1) + 1\right)^{1/\kappa} (4e\pi)^{-3/2} \left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}} 2^{2+1/\kappa}\right)^{\kappa} - 1\right]^{-1/\kappa}.$$

On the other hand, by Lemma 10, for every $u \in \mathbb{R}^3$,

$$g(u) \le \beta_1(3, \kappa)(1 - \kappa\beta_2(3, \kappa)\|u\|)^{1/\kappa},$$

where

$$\beta_1(3, \kappa) = (2 - 2^{-4\kappa})^{1/\kappa} \frac{1}{2\pi^{3/2} d^3} (1-\kappa)^{-1/\kappa} 3\Gamma(3/2) \left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}} 2^{2+1/\kappa}\right)^{\kappa} - 1\right]^{1/\kappa},$$

$$\beta_2(3, \kappa) = \frac{2\pi d^2}{2} (2 - 2^{-3s})^{-1/s} \frac{[(a + (1-s)\beta_1(3,\kappa)^{\kappa})^{1+1/\kappa} - a^{1+1/\kappa}]\kappa}{\beta_1(3,\kappa)^s (1+\kappa)(1-\kappa)},$$

and

$$a = (4e\pi)^{-3\kappa/2} \left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}} 2^{2+1/\kappa}\right)^{\kappa} - 1\right]^{-1}.$$

Denote by $R' = \mathsf{proj}_U(R) \cap \mathsf{ball}(0, d)$, and $\mathsf{ball}(0, d)$ is the origin-centered ball of radius d in $\mathbb{R}^3$. Thus we have

$$\begin{aligned}\int \int \int_{R'} \|u\|^{1/\kappa} h(\kappa) du_1 du_2 du_3 &\le \Pr[x \in R] \\ &\le \int \int \int_{\mathsf{proj}_U(R)} \beta_1(3, \kappa)(1 - \kappa\beta_2(3, \kappa)\|u\|)^{1/\kappa} du_1 du_2 du_3.\end{aligned}$$

Let $A \triangleq \int \int_{\mathsf{proj}_U(R) \cap \mathbb{S}^2} sin\theta d\varphi d\theta$. Note that

$$\int \int \int_{R'} \|u\|^{1/\kappa} h(\kappa) du_1 du_2 du_3 = A \int_0^d r^2 r^{1/\kappa} h(\kappa) dr = A h(\kappa) \frac{1}{3 + 1/\kappa} d^{3+1/\kappa},$$

and

$$\begin{aligned}&\int \int \int_{\mathsf{proj}_U(R)} \beta_1(3, \kappa)(1 - \kappa\beta_2(3, \kappa)\|u\|)^{1/\kappa} du_1 du_2 du_3 \\ &= A\beta_1(3, \kappa) \int_0^{\infty} r^2 (1 - \kappa\beta_2(3, \kappa) r)^{1/\kappa} dr \\ &= A\beta_1(3, \kappa) \frac{B(-1/\kappa - 3, 3)}{(-\kappa\beta_2(3, \kappa))^3}.\end{aligned}$$

So we have

$$Ah(\kappa)\frac{1}{3+1/\kappa}d^{3+1/\kappa} \le \Pr[x \in R] \le A\beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3},$$

and by symmetry,

$$Ah(\kappa)\frac{1}{3+1/\kappa}d^{3+1/\kappa} \le \Pr[x \in -R] \le A\beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3}.$$

Therefore,

$$\Pr[x \in -R] \le \Pr[x \in R]\beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3}\frac{3+1/\kappa}{h(\kappa)d^{3+1/\kappa}}.$$

□

Theorem 18 (restated) *In the PAC realizable case, Algorithm 4 outputs a hypothesis h of error at most ϵ with probability at least $1-\delta$ under isotropic s-concave distribution. The label complexity is $M(\epsilon/2,\delta/4,n^2) + \max\{2m_2/\epsilon, (2/\epsilon^2)\log(4/\delta)\}$, where $M(\epsilon,\delta,m)$ is defined by $M(\epsilon,\delta,n) = O\left(\frac{n}{\epsilon}\log\frac{1}{\epsilon}+\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, $m_2 = M(\max\{\delta/(4eKm_1),\epsilon/2\},\delta/4,n)$, $K = \beta_1(3,\kappa)\frac{B(-1/\kappa-3,3)}{(-\kappa\beta_2(3,\kappa))^3}\frac{3+1/\kappa}{h(\kappa)d^{3+1/\kappa}}$, $d = (1+\gamma)^{-1/\gamma}\frac{1+3\beta}{3+3\beta}$, $h(\kappa) = \left(\frac{1}{d}((2-2^{-4\kappa})^{-1}-1)+1\right)^{\frac{1}{\kappa}}(4e\pi)^{-\frac{3}{2}}\left[\left(\frac{1+\beta}{1+3\beta}\sqrt{3(1+\gamma)^{3/\gamma}}2^{2+\frac{1}{\kappa}}\right)^{\kappa}-1\right]^{-1/\kappa}$, $\beta = \frac{\kappa}{1+2\kappa}$, $\gamma = \frac{\kappa}{1+\kappa}$, and $\kappa = \frac{s}{1+(n-3)s}$. In particular, when $s \to 0$ (a.k.a. log-concave), K is an absolute constant.*

Proof. Denote by p the probability of observing a positive example. We discuss the following three cases.

1. $r < m_2$ and $p < \epsilon$**.**

In this case, the hypothesis that labels every examples as negative has error less than ϵ. Therefore, the algorithm behaves with error at most ϵ when $r < m_2$.

2. $r < m_2$ and $p \ge \epsilon$**.**

By the Hoeffding inequality,

$$\Pr(r < m_2) \le \Pr\left(\frac{r}{m_3} < \frac{\epsilon}{2}\right) \le \Pr\left(\frac{r}{m_3} < p - \frac{\epsilon}{2}\right) \le e^{-m_3\epsilon^2/2} \le \delta/4.$$

So the probability that this case happens is at most $\delta/4$.

3. $r \ge m_2$**.**

We note that

$$\mathsf{err}(h) = \Pr(-H')\Pr(H_u \cap H_v| - H') + \Pr(H')\Pr(h_{xor}(x) \ne c(x)|x \in H'), \tag{8}$$

where $c : \mathbb{R}^n \to \{-1,1\}$ is the hypothesis w.r.t. $H_u \cap H_v$. Observe that

$$\Pr(-H')\Pr(H_u \cap H_v| - H') = \Pr(H_u \cap H_v)\Pr(-H'|H_u \cap H_v),$$

where $\Pr(-H'|H_u \cap H_v)$ is the error of H' over the distribution conditioned on $H_u \cap H_v$. Since the VC argument works for any distribution, and H' contains all $r \ge m_2$ positive examples according to Step 5 in Algorithm 4, by the VC argument, with probability at least $1-\delta/4$,

$$\Pr(-H'|H_u \cap H_v) \le \max\left\{\frac{\delta}{4(1+\gamma)^{1/\gamma}Km_1}, \frac{\epsilon}{2}\right\}.$$

So $\Pr(-H')\Pr(H_u \cap H_v| - H') = \Pr(H_u \cap H_v \cap (-H')) \le \Pr(-H'|H_u \cap H_v) \le \max\left\{\frac{\delta}{4(1+\gamma)^{1/\gamma}Km_1}, \frac{\epsilon}{2}\right\} \le \frac{\epsilon}{2}$.

We now bound the second term in (8). According to Lemma 28,

$$\Pr((-H_u) \cap (-H_v) \cap H') \le K\Pr(H_u \cap H_v \cap (-H')) \le \frac{\delta}{4(1+\gamma)^{1/\gamma}m_1}.$$

On the other hand, by Lemma 8, $\Pr(H') \geq (1+\gamma)^{-1/\gamma}$ with $\gamma = s/(1+ns)$. Thus

$$\Pr((-H_u) \cap (-H_v)|H') = \frac{\Pr((-H_u) \cap (-H_v) \cap (H'))}{\Pr(H')} \leq \frac{\delta}{4m_1}.$$

That is to say, each point in S has probability at most $\delta/(4m_1)$ of being in $(-H_u) \cap (-H_v)$. So by the union bound, with probability at least $1 - \delta/4$, none of points in S is in $(-H_u) \cap (-H_v)$. Therefore, Step 6 in Algorithm 4 is able to find h_{xor} that is consistent with all the instances in S. Then by the VC argument, we have

$$\Pr(h_{xor(x)} \neq c(x)|x \in H') \leq \frac{\epsilon}{2},$$

with probability at least $1 - \delta/4$. In summary, we have

$$\begin{aligned}\mathsf{err}(h) &= \Pr(-H') \Pr(H_u \cap H_v| - H') + \Pr(H') \Pr(h_{xor}(x) \neq c(x)|x \in H') \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,\end{aligned}$$

with failure probability at most $\delta/4 + \delta/4 + \delta = \delta$ by the union bound. Therefore, the proof is completed. □

R Proof of Lower Bounds

The proof of our lower bounds essentially depends on a lower bound on the packing number of all homogeneous linear separators $\mathbb{C}$ under distribution $\mathcal{D}$. Remind that the ϵ-packing number, denoted by $M_{\mathcal{D}}(\mathbb{C}, \epsilon)$, is the maximal cardinality of an ϵ-separated set with classifiers from $\mathbb{C}$, where we say N classifiers $w_1, ..., w_N$ are ϵ-separated w.r.t. $\mathcal{D}$ if $d_{\mathcal{D}}(w_i, w_j) \triangleq \Pr_{x \sim \mathcal{D}}[\text{sign}(w_i \cdot x) \neq \text{sign}(w_j \cdot x)] > \epsilon$ for any $i \neq j$.

Lemma 29. *Suppose that $\mathcal{D}$ is s-concave in $\mathbb{R}^n$, and that its covariance matrix is of full rank. Then for all sufficiently small ϵ, we have $M_{\mathcal{D}}(\mathbb{C}, \epsilon) \geq \frac{\sqrt{n}}{2}\left(\frac{f_1(s,n)}{2\epsilon}\right)^{n-1} - 1$.*

Proof. We begin with proving the lemma in the case of isotropic $\mathcal{D}$. Our proof inspires from proofs for the special case of uniform and log-concave distributions by [48] and [9], respectively.

Denote by UBALL_n the uniform distribution on the sphere in $\mathbb{R}^n$. According to Theorem 12, for any two unit vectors u and v in $\mathbb{R}^n$ we have $f_1(s,n)\theta(u,v) \leq d_{\mathcal{D}}(u,v)$. Thus for a fixed u the probability that a uniformly chosen v obeys $d_{\mathcal{D}}(u,v) \leq \epsilon$ is upper bounded by the volume of those points in the interior of unit ball whose angle is at most $\epsilon/f_1(s,n)$ divided by the volume of unit ball in $\mathbb{R}^n$. By known bound on this ratio [48], we have $\Pr_{v \in \mathsf{UBALL}_n}[d_{\mathcal{D}}(u,v) \leq \epsilon] \leq \frac{1}{\sqrt{n}}\left(\frac{2\epsilon}{f_1(s,n)}\right)^{n-1}$. So $\Pr_{u,v \in \mathsf{UBALL}_n}[d_{\mathcal{D}}(u,v) \leq \epsilon] \leq \frac{1}{\sqrt{n}}\left(\frac{2\epsilon}{f_1(s,n)}\right)^{n-1}$, meaning that if we select a set S of s normalized vectors uniformly from the unit sphere, the expected number of pairs of vectors that are ϵ-close in the sense of $d_{\mathcal{D}}$ is at most $\frac{s^2}{\sqrt{n}}\left(\frac{2\epsilon}{f_1(s,n)}\right)^{n-1}$. Removing one vector from each pair of S yields a set of $s - \frac{s^2}{\sqrt{n}}\left(\frac{2\epsilon}{f_1(s,n)}\right)^{n-1}$ homogeneous linear separators that are ϵ-separated. The proof for isotropic $\mathcal{D}$ is completed when we set $s = \frac{\sqrt{n}}{(2\epsilon/f_1(s,n))^{n-1}}$.

We now discuss the case when $\mathcal{D}$ is non-isotropic. Denote by Σ the covariance matrix of $\mathcal{D}$ and let isotropic $\mathcal{D}'$ be the whitened version of $\mathcal{D}$, namely, the distribution obtained by first sampling x from $\mathcal{D}$ and then computing $\Sigma^{-1/2}x$. Notice that $d_{\mathcal{D}}(u,v) = d_{\mathcal{D}'}(u\Sigma^{1/2}, v\Sigma^{1/2})$. Therefore, we can apply an ϵ-packing w.r.t. $\mathcal{D}'$ to construct an ϵ-packing w.r.t. $\mathcal{D}'$ of the same size. □

Now we are ready to prove Theorem 19.

Theorem 19 (restated) *For a fixed value $-\frac{1}{2n+3} \leq s \leq 0$ we have: (a) For any s-concave distribution $\mathcal{D}$ in $\mathbb{R}^n$ whose covariance matrix is of full rank, the sample complexity of learning origin-centered*

linear separators under $\mathcal{D}$ in the passive learning scenario is $\Omega\left(\frac{nf_1(s,n)}{\epsilon}\right)$; (b) The label complexity of active learning of linear separators under s-concave distribution is $\Omega\left(n\log\left(\frac{f_1(s,n)}{\epsilon}\right)\right)$.

Proof. It is known that for any distribution $\mathcal{D}$ in $\mathbb{R}^n$, the sample complexity of (passive) PAC learning of homogeneous linear separators under $\mathcal{D}$ is at least $\frac{n-1}{e}\left(\frac{M_{\mathcal{D}}(\mathbb{C},2\epsilon)}{4}\right)^{1/(n-1)}$ [48]. By Lemma 29, we have an $\Omega\left(\frac{nf_1(s,n)}{\epsilon}\right)$ lower bound of sample complexity for passive learning homogeneous halfspace.

We now discuss the label complexity lower bound in the active learning scenario. By [46], any active learning algorithm that is allowed to make arbitrary binary queries must take at least $\Omega(\log M_{\mathcal{D}}(\mathbb{C},\epsilon))$ so as to output a hypothesis of error at most ϵ with high probability. Applying Lemma 29, we obtain the desired result. □

S Related Algorithms

S.1 Margin Based Active Learning (Realizable Case)

Algorithm 2 Margin Based Active Learning under S-Concave Distributions (Realizable Case)

Input: $b_k = \min\{\Theta(2^{-k}f_4f_1^{-1}), d\}$, $m_k = C\left(\frac{f_3b_{k-1}}{2^{-k}}\left(n\log\frac{f_3b_{k-1}}{2^{-k}} + \log\frac{1+s-k}{\delta}\right)\right)$, and $T = \lceil\log\frac{1}{c\epsilon}\rceil$.
1: Draw m_1 examples from $\mathcal{D}$, label them and put them into $W(1)$.
2: For $k = 1, 2, ..., T$
3: Find a hypothesis w_k with $\|w_k\| = 1$ that is consistent with $W(k)$.
4: $W(k+1) \leftarrow W(k)$.
5: **While** m_{k+1} additional data points are not labeled
6: Draw sample x from $\mathcal{D}$.
7: **If** $|w_k \cdot x| \geq b_k$
8: Reject x.
9: **Else**
10: Ask for label of x and put into $W(k+1)$.
11: **End If**
12: **End While**
13: End For
Output: Hypothesis w_T.

S.2 Margin Based Active Learning (Adversarial Noise)

Procedure 3 Margin Based Active Learning under S-Concave Distributions (Adversarial Noise)

Input: Parameters b_k, τ_k, r_k, m_k, κ, and T as in Theorem 16.
1: Draw m_1 examples from $\mathcal{D}$, label them and put them into W.
2: **For** $k = 1, 2, ..., T$
3: Find $v_k \in \mathsf{ball}(w_{k-1}, r_k)$ to approximately minimize the hinge loss over W s.t. $\|v_k\| \leq 1$: $\ell_{\tau_k} \leq \min_{w\in\mathsf{ball}(w_{k-1},r_k)\cap\mathsf{ball}(0,1)} \ell_{\tau_k}(w, W) + \kappa/8$.
4: Normalize v_k, yielding $w_k = \frac{v_k}{\|v_k\|}$.
5: Clear the working set W.
6: **While** m_{k+1} additional data points are not labeled
7: Draw sample x from $\mathcal{D}$.
8: **If** $|w_k \cdot x| \geq b_k$, reject x; **else** ask for label of x and put into W.
9: **End While**
10: End For
Output: Hypothesis w_T.

S.3 Learning Intersections of Halfspaces

Algorithm 4 Learning Intersections of Halfspaces under S-Concave Distributions

Input: Parameters m_1, m_2, and m_3 as in Theorem 18.
1: Draw m_3 examples. Denote by r the number of observed positive examples.
2: If $r < m_2$, output the hypothesis that labels every point as negative, and end the algorithm.
3: Learn an origin-centered halfspace H' which contains all r positive examples.
4: Draw a set S of m_1 i.i.d. examples in H'. Learn a weight vector $w \in \mathbb{R}^{n \times n}$ such that the hypothesis $h_{xor} = \text{sign}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j\right)$ is consistent with the set S.
Output: $h : \mathbb{R}^n \to \{-1, 1\}$ such that $h(x) = h_{xor}(x)$ if $x \in H'$; Otherwise, $h(x) = -1$.

T A Collection of Concentration Results

Theorem 30 ([54, 17]). *Denote by $\mathcal{C}$ a class of concepts from a set X to $\{-1, 1\}$ with VC dimension n. Let $c \in \mathcal{C}$, and assume that*

$$M(\epsilon, \delta, n) = O\left(\frac{n}{\epsilon} \log \frac{1}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$$

examples $x_1, ..., x_M$ are sampled from any probability distribution $\mathcal{D}$ over X. Then any hypothesis $h \in \mathcal{C}$ which is consistent with c on $x_1, ..., x_M$ has error at most ϵ, with probability at least $1 - \delta$.

Theorem 31 ([1]). *Let F be a set of functions mapping from domain X to $[a, b]$, and let n be the pseudo-dimension of F. Then for any distribution $\mathcal{D}$ over X and $m = O\left(\frac{(b-a)^2}{\kappa^2}(d + \log(1/\delta))\right)$, if $x_1, ..., x_m$ are drawn independently from $\mathcal{D}$, with probability at least $1 - \delta$, for all $f \in F$,*

$$\left| \mathbb{E}_{x \sim \mathcal{D}} f(x) - \frac{1}{m} \sum_{i=1}^{m} f(x_i) \right| \leq \kappa.$$