4-2013

Active and passive learning of linear separators under log-concave distributions

Maria-Florina Balcan
Carnegie Mellon University, ninamf@cs.cmu.edu

Philip M. Long
Microsoft

Published In
Journal of Machine Learning Research, 30, 1-29.

JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–29

Active and passive learning of linear separators under log-concave distributions

Maria Florina Balcan NINAMF@CC.GATECH.EDU
Georgia Institute of Technology

Philip M. Long PLONG@MICROSOFT.COM
Microsoft

Abstract

We provide new results concerning label efficient, polynomial time, passive and active learning of linear separators. We prove that active learning provides an exponential improvement over PAC (passive) learning of homogeneous linear separators under nearly log-concave distributions. Building on this, we provide a computationally efficient PAC algorithm with optimal (up to a constant factor) sample complexity for such problems. This resolves an open question of (Long, 1995, 2003; Bshouty et al., 2009) concerning the sample complexity of efficient PAC algorithms under the uniform distribution in the unit ball. Moreover, it provides the first bound for a polynomial-time PAC algorithm that is tight for an interesting infinite class of hypothesis functions under a general and natural class of data-distributions, providing significant progress towards a longstanding open question of (Ehrenfeucht et al., 1989; Blumer et al., 1989).

We also provide new bounds for active and passive learning in the case that the data might not be linearly separable, both in the agnostic case and and under the Tsybakov low-noise condition. To derive our results, we provide new structural results for (nearly) log-concave distributions, which might be of independent interest as well.

Keywords: Active learning, PAC learning, ERM, nearly log-concave distributions, Tsybakov low-noise condition, agnostic learning.

1. Introduction

Learning linear separators is one of the central challenges in machine learning. They are widely used and have been long studied both in the statistical and computational learning theory. A seminal result of (Blumer et al., 1989), using tools due to (Vapnik and Chervonenkis, 1971), showed that d-dimensional linear separators can be learned to accuracy $1-\epsilon$ with probability $1-\delta$ in the classic PAC model in polynomial time with $O((d/\epsilon)\log(1/\epsilon)+(1/\epsilon)\log(1/\delta))$ examples. The best known lower bound for linear separators is $\Omega(d/\epsilon+(1/\epsilon)\log(1/\delta))$, and this holds even in the case in which the distribution is uniform (Long, 1995). Whether the upper bound can be improved to match the lower bound via a polynomial-time algorithm is been long-standing open question, both for general distributions (Ehrenfeucht et al., 1989; Blumer et al., 1989) and for the case of the uniform distribution in the unit ball (Long, 1995, 2003; Bshouty et al., 2009). In this work we resolve this question in the case where the underlying distribution belongs to the class of log-concave and nearly log-concave distributions, a wide class of distributions that includes the gaussian distribution and uniform distribution over any convex set, and which has played an important role in several areas including sampling, optimization, integration, and learning (Lovasz and Vempala, 2007).

We also consider active learning, a major area of research of modern machine learning, where the algorithm only receives the classifications of examples when it requests them (Dasgupta, 2011).

Our main result here is a polynomial-time active learning algorithm with label complexity that is exponentially better than the label complexity of any passive learning algorithm in these settings. This answers an open question in (Balcan et al., 2007) and it also significantly expands the set of cases for which we can show that active learning provides a clear exponential improvement in $1/\epsilon$ (without increasing the dependence on d) over passive learning. Remarkably, our analysis for passive learning is done via a connection to our analysis for active learning – to our knowledge, this is the first paper using this technique.

We also study active and passive learning in the case that the data might not be linearly separable. We specifically provide new improved bounds for the widely studied Tsybakov low-noise condition (Mammen and Tsybakov, 1999; Bartlett et al., 2005; Massart and Nedelec, 2006), as well as new bounds on the disagreement coefficient, with implications for the agnostic case (i.e., arbitrary forms of noise).

Passive Learning In the classic passive supervised machine learning setting, the learning algorithm is given a set of labeled examples drawn i.i.d. from some fixed but unknown distribution over the instance space and labeled according to some fixed but unknown target function, and the goal is to output a classifier that does well on new examples coming from the same distribution. This setting has been long studied in both computational learning theory (within the PAC model (Valiant, 1984; Kearns and Vazirani, 1994)) and statistical learning theory (Vapnik, 1982, 1998; Boucheron et al., 2005), and has played a crucial role in the developments and successes of machine learning.

However, despite remarkable progress, the basic question of providing polynomial-time algorithms with *tight* bounds on the sample complexity has remained open. Several milestone results along these lines that are especially related to our work include the following. The analysis of (Blumer et al., 1989), proved using tools from (Vapnik and Chervonenkis, 1971), implies that linear separators can be learned in polynomial time with $O((d/\epsilon)\log(1/\epsilon) + (1/\epsilon)\log(1/\delta))$ labeled examples. (Ehrenfeucht et al., 1989) proved a bound that implies an $\Omega(d/\epsilon + (1/\epsilon)\log(1/\delta))$ lower bound for linear separators and explicitly posed the question of providing tight bounds for this class. (Haussler et al., 1994) established an upper bound of $O((d/\epsilon)\log(1/\delta))$, which can be achieved in polynomial-time for linear separators.

(Blumer et al., 1989) achieved polynomial-time learning by finding a consistent hypothesis (i.e., a hypothesis which correctly classifies all training examples); this is a special case of ERM (Vapnik, 1982). An intensive line of research in the empirical process and statistical learning theory literature has taken account of "local complexity" to prove stronger bounds for ERM (van der Vaart and Wellner, 1996; van de Geer, 2000; Bartlett et al., 2005; Long, 2003; Mendelson, 2003; Giné and Koltchinskii, 2006; Hanneke, 2007; Hanneke and Yang, 2012). In the context of learning, local complexity takes account of the fact that really bad classifiers can be easily discarded, and the set of "local" classifiers that are harder to disqualify is sometimes not as rich. A recent landmark result of (Giné and Koltchinskii, 2006) (see also (Raginsky and Rakhlin, 2011; Hanneke and Yang, 2012)) is the bound for consistent algorithms of

$$O((d/\epsilon)\log(\mathrm{cap}(\epsilon)) + (1/\epsilon)\log(1/\delta)) \tag{1}$$

where $\mathrm{cap}(\epsilon)$ is the Alexander capacity, which depends on the distribution (Alexander., 1987) (see Section 8 and Appendix A for further discussion). However, this bound can be suboptimal for linear separators.

In particular, for linear separators in the case in which the underlying distribution is uniform in the unit ball, the sample complexity is known (Long, 1995, 2003) to be $\Theta\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$, when

computational considerations are ignored. (Bshouty et al., 2009), using the doubling dimension (Assouad, 1983), another measure of local complexity, proved a bound of

$$O((d/\epsilon)\sqrt{\log(1/\epsilon)} + (1/\epsilon)\log(1/\delta)) \tag{2}$$

for a polynomial-time algorithm. As a lower bound of $\Omega(\sqrt{d})$ on $\mathrm{cap}(\epsilon)$ for $\epsilon = o(1/\sqrt{d})$ for the case of linear separators and the uniform distribution is implicit in (Hanneke, 2007), the bound of (Giné and Koltchinskii, 2006) given by (1) cannot yield a bound better than

$$O((d/\epsilon)\min\{\log d, \log(1/\epsilon)\} + (1/\epsilon)\log(1/\delta)) \tag{3}$$

in this case.

In this paper we provide a *tight* bound (up to constant factors) on the sample complexity of polynomial-time learning of linear separators with respect to log-concave distributions. Specifically, we prove an upper bound of $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ using a polynomial-time algorithm that holds for any zero-mean log-concave distribution. We also prove an information theoretic lower bound that matches our (computationally efficient) upper bound for *each* log-concave distribution. This provides the first bound for a polynomial-time algorithm that is tight for an interesting non-finite class of hypothesis functions under a general class of data-distributions, and also characterizes (up to a constant factor) the distribution-specific sample complexity for each distribution in the class. In the special case of the uniform distribution, our upper bound closes the existing $\Omega(\min\{\sqrt{\log(1/\epsilon)}, \log(d)\})$ gap between the upper bounds (2) and (3) and the lower bound of (Long, 1995).

Active Learning We also study learning of linear separators in the active learning model; here the learning algorithm can access unlabeled (i.e., unclassified) examples and ask for labels of unlabeled examples of its own choice, and the hope is that a good classifier can be learned with significantly fewer labels by actively directing the queries to informative examples. This has been a major area of machine learning research in the past fifteen years mainly due the availability of large amounts of unannotated or raw data in many modern applications (Dasgupta, 2011), with many exciting developments on understanding its underlying principles as well (Freund et al., 1997; Dasgupta, 2005; Balcan et al., 2006, 2007; Hanneke, 2007; Dasgupta et al., 2007; Castro and Nowak, 2007; Balcan et al., 2008; Koltchinskii, 2010; Beygelzimer et al., 2010). However, with a few exceptions (Balcan et al., 2007; Castro and Nowak, 2007; Dasgupta et al., 2005), most of the theoretical developments have focused on the so called disagreement-based active learning paradigm (Hanneke, 2011; Koltchinskii, 2010); methods and analyses developed in this context are often suboptimal, as they take a conservative approach and consider strategies that query even points on which there is a small amount of uncertainty (or disagreement) among the classifiers still under consideration given the labels queried so far. The results derived in this manner often show an improvement in the $1/\epsilon$ factor in the label complexity of active versus passive learning; however, unfortunately, the dependence on the d term typically gets worse.

By analyzing a more aggressive, margin-based active learning algorithm, we prove that we can efficiently (in polynomial time) learn homogeneous linear separators when the underlying distribution is log-concave by using only $O((d+\log(1/\delta)+\log\log(1/\epsilon))\log(1/\epsilon))$ label requests, answering an open question in (Balcan et al., 2007). This represents an exponential improvement of active learning over passive learning and it significantly broadens the cases for which we can show that the dependence on $1/\epsilon$ in passive learning can be improved to only $\tilde{O}(\log(1/\epsilon))$ in active learning, but without increasing the dependence on the dimension d. We note that an improvement of this type was known to be possible only for the case when the underlying distributions is (nearly) uniform

in the unit ball (Balcan et al., 2007; Dasgupta et al., 2005; Freund et al., 1997); even for this special case, our analysis improves by a multiplicative $\log\log(1/\epsilon)$ factor the results of (Balcan et al., 2007); it also provides better dependence on d than any other previous analyses implementable in a computationally efficient manner (both disagreement-based (Hanneke, 2011, 2007) and more aggressive ones (Dasgupta et al., 2005; Freund et al., 1997)), and over the inefficient splitting index analysis of (Dasgupta, 2005).

Techniques At the core of our results is a novel characterization of the region of disagreement of two linear separators under a log-concave measure. We show that for any two linear separators specified by normal vectors u and v, for any constant $c \in (0,1)$ we can pick a margin as small as $\gamma = \theta(\alpha)$, where α is the angle between u and v, and still ensure that the probability mass of the region of disagreement outside of band of margin γ of one of them is $c\alpha$ (Theorem 4). Using this fact, we then show how we can use a margin-based active learning technique, where in each round we only query points near the hypothesized decision boundary, to get an exponential improvement over passive learning.

We then show that any passive learning algorithm that outputs a hypothesis consistent with $O(d/\epsilon + (1/\epsilon)\log(1/\delta))$ random examples will, with probability at least $1-\delta$, output a hypothesis of error at most ϵ (Theorem 6). Interestingly, our analysis is quite dissimilar to the classic analyses of ERM. It proceeds by conceptually running the algorithm online on progressively larger chunks of examples, and using the intermediate hypotheses to track the progress of the algorithm. We show, using the same tools as in the active learning analysis, that it is always likely that the algorithm will receive informative examples. Our analysis shows that the algorithm would also achieve $1-\epsilon$ accuracy with high probability even if it periodically built preliminary hypotheses using some of the examples, and then only used borderline cases for those preliminary classifiers for further training.[1] To achieve the optimal sample complexity, we have to carefully distribute the confidence parameter, by allowing higher probability of failure in the later stages, to compensate for the fact that, once the hypothesis is already pretty good, it takes longer to get examples that help to further improve it.

Non-separable case We also study label-efficient learning in the presence of noise. We show how our results for the realizable case can be extended to handle (a variant of) the Tsybakov noise, which has received substantial attention in statistical learning theory, both for passive and active learning (Mammen and Tsybakov, 1999; Bartlett et al., 2005; Massart and Nedelec, 2006; Giné and Koltchinskii, 2006; Balcan et al., 2007; Koltchinskii, 2010; Hanneke, 2011); this includes the random classification noise commonly studied in computational learning theory (Kearns and Vazirani, 1994), and the more general bounded (or Massart) noise (Bartlett et al., 2005; Massart and Nedelec, 2006; Giné and Koltchinskii, 2006; Koltchinskii, 2010). Our analysis for Massart noise leads to optimal bounds (up to constant factors) for active and passive learning of linear separators when the marginal distribution on the feature vectors is log-concave, improving the dependence on d over previous best known results. Our analysis for Tsybakov noise leads to bounds on active learning with improved dependence on d over previous known results in this case as well.

We also provide a bound on the Alexander's capacity (Alexander., 1987; Giné and Koltchinskii, 2006) and the closely related disagreement coefficient notion (Hanneke, 2007), which have been widely used to characterize the sample complexity of various (active and passive) algorithms (Hanneke, 2007; Koltchinskii, 2010; Giné and Koltchinskii, 2006; Beygelzimer et al., 2010). This immediately implies concrete bounds on the labeled data complexity of several algorithms in the literature, in-

1. Note that such examples would not be i.i.d from the underlying distribution!

cluding active learning algorithms designed for the purely agnostic case (i.e., arbitrary forms of noise), e.g., the A^2 algorithm (Balcan et al., 2006) and the DHM algorithm (Dasgupta et al., 2007).

Nearly log-concave distributions We also extend our results both for passive and active learning to deal with nearly log-concave distributions; this is a broader class of distributions introduced by (Applegate and Kannan, 1991), which contains mixtures of (not too separated) log-concave distributions. In deriving our results, we provide new tail bounds and structural results for these distributions, which might be of independent interest and utility, both in learning theory and in other areas including sampling and optimization.

We note that our bounds on the disagreement coefficient improve by a factor of $\Omega(d)$ over the bounds of (Friedman, 2009) (matching what was known for the much less general case of nearly uniform distribution over the unit sphere); furthermore, they apply to the nearly log-concave case where we allow an arbitrary number of discontinuities, a case not captured by the (Friedman, 2009) conditions at all. We discuss other related papers in Appendix A.

2. Preliminaries and Notation

We focus on binary classification problems; that is, we consider the problem of predicting a binary label y based on its corresponding input vector x. As in the standard machine learning formulation, we assume that the data points (x, y) are drawn from an unknown underlying distribution D_{XY} over $X \times Y$; X is called the *instance space* and Y is the *label space*. In this paper we assume that $Y = \{\pm 1\}$ and $X = \mathbb{R}^d$; we also denote the marginal distribution over X by D. Let $\mathbb{C}$ be the class of linear separators through the origin, that is $\mathbb{C} = \{\text{sign}(w \cdot x) : w \in \mathbb{R}^d, \|w\| = 1\}$. To keep the notation simple, we sometimes refer to a weight vector and the linear classifier with that weight vector interchangeably. Our goal is to output a hypothesis function $w \in \mathbb{C}$ of small error, where $\text{err}(w) = \text{err}_{D_{XY}}(w) = P_{(x,y)\sim D_{XY}}[\text{sign}(w \cdot x) \neq y]$.

We consider two learning protocols: passive learning and active learning. In the passive learning setting, the learning algorithm is given a set of labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$ drawn i.i.d. from D_{XY} and the goal is output a hypothesis of small error by using only a polynomial number of labeled examples. In the (pool-based) active learning setting, a set of labeled examples $(x_1, y_1) \ldots (x_m, y_m)$ is also drawn i.i.d. from D_{XY}; the learning algorithm is permitted direct access to the sequence of x_i values (unlabeled data points), but has to make a label request to obtain the label y_i of example x_i. The hope is that in the active learning setting we can output a classifier of small error by using many fewer label requests than in the passive learning setting by actively directing the queries to informative examples (while keeping the number of unlabeled examples polynomial). For added generality, we also consider the selective sampling active learning model, where the algorithm visits the unlabeled data points x_i in sequence, and, for each i, makes a decision on whether or not to request the label y_i based only on the previously-observed x_j values ($j \leq i$) and corresponding requested labels, and never changes this decision once made. Both our upper and lower bounds will apply to both selective sampling and pool-based active learning.

In the "realizable case", we assume that the labels are deterministic and generated by a target function that belongs to $\mathbb{C}$. In the non-realizable case (studied in Sections 8 and 9) we do not make this assumption and instead aim to compete with the best function in $\mathbb{C}$.

Given two vectors u and v and any distribution $\tilde{D}$ we denote by $\text{d}_{\tilde{D}}(u, v) = \mathbb{P}_{x\sim\tilde{D}}(\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x))$; we also denote by $\theta(u, v)$ the angle between the vectors u and v.

3. Log-Concave Densities

Throughout this paper we focus on the case where the underlying distribution D is log-concave or nearly log-concave. Such distributions have played a key role in the past two decades in several areas including sampling, optimization, and integration algorithms (Lovasz and Vempala, 2007), and more recently for learning theory as well (Kalai et al., 2005; Klivans et al., 2009b; Vempala, 2010). In this section we first summarize known results about such distributions that are useful for our analysis and then prove a novel structural statement that will be key to our analysis (Theorem 4). In Section 6 we describe extensions to nearly log-concave distributions as well.

Definition 1 *A distribution over R^d is log-concave if $\log f(\cdot)$ is concave, where f is its associated density function. It is isotropic if its mean is the origin and its covariance matrix is the identity.*

Log-concave distributions form a broad class of distributions: for example, the Gaussian, Logistic, and uniform distribution over any convex set are log-concave distributions. The following lemma summarizes known useful facts about isotropic log-concave distributions (most are from (Lovasz and Vempala, 2007); the upper bound on the density is from (Klivans et al., 2009b)).

Lemma 2 *Assume that D is log-concave in R^d and let f be its density function.*

(a) If D is isotropic then $\mathbb{P}_{x\sim D}[||X|| \geq \alpha\sqrt{d}] \leq e^{-\alpha+1}$. If $d = 1$ then: $\mathbb{P}_{x\sim D}[X \in [a,b]] \leq |b-a|$.

(b) If D is isotropic, then $f(x) \geq 2^{-7d}2^{9d||x||}$ whenever $0 \leq ||x|| \leq 1/9$. Furthermore, $2^{-7d} \leq f(0) \leq d(20d)^{d/2}$, and $f(x) \leq A(d)\exp(-B(d)||x||)$, where $A(d)$ is $2^{8d}d^{d/2}e$ and $B(d)$ is $\frac{2^{-7d}}{2(d-1)(20(d-1))^{(d-1)/2}}$, for all x of any norm.

(c) All marginals of D are log-concave. If D is isotropic, its marginals are isotropic as well.

(d) If $\mathbb{E}[\|X\|^2] = C^2$, then $\mathbb{P}[||X|| \geq RC] \leq e^{-R+1}$.

(e) If D is isotropic and $d = 1$ we have $f(0) \geq 1/8$ and $f(x) \leq 1$ for all x.

Throughout our paper we will use the fact that there exists a universal constant c such that the probability of disagreement of any two homogeneous linear separators is lower bounded by the c times the angle between their normal vectors. This follows by projecting the region of disagreement in the space given by the two normal vectors, and then using properties of log-concave distributions in 2-dimensions. The proof is implicit in earlier works (e.g., (Vempala, 2010)); for completeness, we include a proof in Appendix B.

Lemma 3 *Assume D is an isotropic log-concave in R^d. Then there exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v,u) \leq \mathrm{d}_D(u,v)$.*

To analyze our active and passive learning algorithms we provide a novel characterization of the region of disagreement of two linear separators under a log-concave measure:

Theorem 4 *For any $c_1 > 0$, there is a $c_2 > 0$ such that the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u,v) = \alpha < \pi/2$. If D is isotropic log-concave in R^d, then:*

$$\mathbb{P}_{x\sim D}[\mathrm{sign}(u\cdot x) \neq \mathrm{sign}(v\cdot x) \text{ and } |v\cdot x| \geq c_2\alpha] \leq c_1\alpha. \tag{4}$$

Proof Choose $c_1, c_2 > 0$. We will show that, if c_2 is large enough relative to $1/c_1$, then (4) holds. Let $b = c_2\alpha$. Let E be the set whose probability we want to bound. Since the event under consideration only concerns the projection of x onto the span of u and v, Lemma 2(c) implies we can assume without loss of generality that $d = 2$.

Next, we claim that each member x of E has $||x|| \geq b/\alpha = c_2$. Assume without loss of generality that $v \cdot x$ is positive. (The other case is symmetric.) Then $u \cdot x < 0$, so the angle of x with u is obtuse, i.e. $\theta(x, u) \geq \pi/2$. Since $\theta(u, v) = \alpha$, this implies that $\theta(x, v) \geq \pi/2 - \alpha$. But $x \cdot v \geq b$, and v is unit length, so $||x|| \cos\theta(x, v) \geq b$, which, since $\theta(x, v) \geq \pi/2 - \alpha$, implies $||x|| \cos(\pi/2 - \alpha) \geq b$; This, since $\cos(\pi/2 - \alpha) \leq \alpha$ for all $\alpha \in [0, \pi/2]$, in turn implies $||x|| \geq b/\alpha = c_2$. This implies that, if $B(r)$ is a ball of radius r in R^2, that

$$\mathbb{P}[E] = \sum_{i=1}^{\infty} \mathbb{P}[E \cap (B((i+1)c_2) - B(ic_2))]. \tag{5}$$

To obtain the desired bound, we carefully bound each term in the RHS. Choose $i \geq 1$.

Let $f(x_1, x_2)$ be the density of D. We have

$$\mathbb{P}[E \cap (B((i+1)c_2) - B(ic_2))] = \int_{(x_1,x_2)\in B((i+1)c_2)-B(ic_2)} 1_E(x_1, x_2) f(x_1, x_2)\, dx_1 dx_2.$$

Applying the density upper bound from Lemma 2 with $d = 2$, there are constants C_1 and C_2 such that

$$\begin{aligned}\mathbb{P}[E \cap (B((i+1)c_2) - B(ic_2))] &\leq \int_{(x_1,x_2)\in B((i+1)c_2)-B(ic_2)} 1_E(x_1, x_2) C_1 \exp(-c_2 C_2 i)\, dx_1 dx_2 \\ &= C_1 \exp(-c_2 C_2 i) \int_{(x_1,x_2)\in B((i+1)c_2)-B(ic_2)} 1_E(x_1, x_2)\, dx_1 dx_2.\end{aligned}$$

If we include $B(ic_2)$ in the integral again, we get

$$\mathbb{P}[E \cap (B((i+1)c_2) - B(ic_2))] \leq C_1 \exp(-c_2 C_2 i) \int_{(x_1,x_2)\in B((i+1)c_2)} 1_E(x_1, x_2)\, dx_1 dx_2.$$

Now, we exploit the fact that the integral above is a rescaling of a probability with respect to the uniform distribution. Let C_3 be the volume of the unit ball in R^2. Then, we have

$$\mathbb{P}[E \cap (B((i+1)c_2) - B(ic_2))] \leq C_1 \exp(-c_2 C_2 i) C_3 (i+1)^2 c_2^2 \alpha/\pi = C_4 c_2^2 \alpha (i+1)^2 \exp(-c_2 C_2 i),$$

for $C_4 = C_1 C_3/\pi$. Returning to (5), we get

$$\mathbb{P}[E] = \sum_{i=1}^{\infty} C_4 c_2^2 \alpha (i+1)^2 \exp(-c_2 C_2 i) = C_4 c_2^2 \times \frac{4e^{2c_2C_2} - 3e^{c_2C_2} + 1}{\left(e^{c_2C_2} - 1\right)^3} \times \alpha.$$

Since $\lim_{c_2\to\infty} c_2^2 \times \frac{4e^{2c_2C_2} - 3e^{c_2C_2} + 1}{\left(e^{c_2C_2} - 1\right)^3} = 0$, this completes the proof. ■

We note that a weaker result of this type was proven (via different techniques) for the uniform distribution in the unit ball in (Balcan et al., 2007). In addition to being more general, Theorem 4 is tighter and more refined even for this specific case – this improvement is essential for obtaining tight bounds for polynomial time algorithms for passive learning (Section 5) and better bounds for active learning as well.

4. Active Learning

In this section we analyze a margin-based algorithm for actively learning linear separators under log-concave distributions (Balcan et al., 2007) (Algorithm 1). Lower bounds proved in Section 7 show that this algorithm needs exponentially fewer labeled examples than any passive learning algorithm.

This algorithm has been previously proposed and analyzed in (Balcan et al., 2007) for the special case of the uniform distribution in the unit ball. In this paper we analyze it for the much more general class of log-concave distributions.

Algorithm 1 Margin-based Active Learning

Input: a sampling oracle for D, a labeling oracle, sequences $m_k > 0$, $k \in Z^+$ (sample sizes) and $b_k > 0$, $k \in Z^+$ (cut-off values).
Output: weight vector $\hat{w}_s$.

- Draw m_1 examples from D, label them and put them in $W(1)$.
- **iterate** $k = 1, \ldots, s$
  - find a hypothesis $\hat{w}_k$ with $\|\hat{w}_k\|_2 = 1$ consistent with all labeled examples in $W(k)$.
  - let $W(k+1) = W(k)$.
  - until m_{k+1} additional data points are labeled, draw sample x from D
    - if $|\hat{w}_k \cdot x| \geq b_k$, then reject x,
    - else, ask for label of x, and put into $W(k+1)$.

Theorem 5 *Assume D is isotropic log-concave in R^d. There exist constants C_1, C_2 s.t. for $d \geq 4$, and for any $\epsilon, \delta > 0$, $\epsilon < 1/4$, using Algorithm 1 with $b_k = \frac{C_1}{2^k}$ and $m_k = C_2\left(d + \ln \frac{1+s-k}{\delta}\right)$, after $s = \lceil \log_2 \frac{1}{c\epsilon} \rceil$ iterations, we find a separator of error at most ϵ with probability $1-\delta$. The total number of labeled examples needed is $O((d + \log(1/\delta) + \log\log(1/\epsilon))\log(1/\epsilon))$.*

Proof Let c be the constant from Lemma 3. We will show, using 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 $\hat{w}$ consistent with the data in the working set $W(k)$ has $\mathrm{err}(\hat{w}) \leq c2^{-k}$, so that, in particular, $\mathrm{err}(\hat{w}_k) \leq c2^{-k}$.

The case where $k = 1$ follows from the standard VC bounds (see e.g.,(Vapnik and Chervonenkis, 1971)). Assume now the claim is true for $k-1$ ($k > 1$), and consider the kth iteration. Let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$, and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By the induction hypothesis, we know that, with probability at least $1 - \frac{\delta}{2}\sum_{i<k-1}\frac{1}{(1+s-i)^2}$, all $\hat{w}$ consistent with $W(k-1)$, including $\hat{w}_{k-1}$, have errors at most $c2^{-(k-1)}$. Consider an arbitrary such $\hat{w}$. By Lemma 3 we have $\theta(\hat{w}, w^*) \leq 2^{-(k-1)}$ and $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(k-1)}$, so $\theta(\hat{w}_{k-1}, \hat{w}) \leq 4 \times 2^{-k}$. Applying Theorem 4, there is a choice of C_1 (the constant such that $b_{k-1} = C_1/2^{k-1}$) that satisfies $\mathbb{P}((\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{4}$ and $\mathbb{P}((\hat{w}_{k-1} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{4}$. So

$$\mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \frac{c2^{-k}}{2}. \tag{6}$$

Now let us treat the case that $x \in S_1$. Since we are labeling m_k data points in S_1 at iteration $k-1$, classic Vapnik-Chervonenkis bounds (1971) imply that, if C_2 is a large enough absolute constant, then with probability $1 - \delta/(4(1+s-k)^2)$, for all $\hat{w}$ consistent with the data in $W(k)$,

$$\mathrm{err}(\hat{w}|S_1) = \mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0 \mid x \in S_1) \leq \frac{c2^{-k}}{4b_k} = \frac{c}{4C_1}. \tag{7}$$

Finally, since S_1 consists of those points that, after projecting onto the direction $\hat{w}_{k-1}$, fall into an interval of length $2b_k$, Lemma 2 implies that $\mathbb{P}(S_1) \leq 2b_k$. Putting this together with (6) and (7), with probability $1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$, we have $\mathrm{err}(\hat{w}) \leq c2^{-k}$, completing the proof. ■

5. Passive Learning

In this section we show how an analysis that was inspired by active learning leads to optimal (up to constant factors) bounds for polynomial-time algorithms for passive learning.

Theorem 6 *Assume that D is zero mean and log-concave in R^d. There exists an absolute constant C_3 s.t. for $d \geq 4$, and for any $\epsilon, \delta > 0$, $\epsilon < 1/4$, any algorithm that outputs a hypothesis that correctly classifies $m = \frac{C_3(d+\log(1/\delta))}{\epsilon}$ examples finds a separator of error at most ϵ with probability $\geq 1 - \delta$.*

PROOF SKETCH: We focus here on the case that D is isotropic. We can treat the non-isotropic case by observing that the two cases are equivalent; one may pass between them by applying the whitening transform. (See Appendix C for details.)

While our analysis will ultimately provide a guarantee for any learning algorithm that always outputs a consistent hypothesis, we will use intermediate hypothesis of Algorithm 1 in the analysis.

Let c be the constant from Lemma 3. While proving Theorem 5, we proved that, if Algorithm 1 is run with $b_k = \frac{C_1}{2^k}$ and $m_k = C_2\left(d + \ln\frac{1+s-k}{\delta}\right)$, that for all $k \leq s$, with probability $\geq 1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$ any $\hat{w}$ consistent with the data in $W(k)$ has $\mathrm{err}(\hat{w}) \leq c2^{-k}$. Thus, after $s = O(\log(1/\epsilon))$ iterations, with probability at least $\geq 1 - \delta$, any linear classifier consistent with *all* the training data has error $\leq \epsilon$, since any such classifier is consistent with the examples in $W(s)$.

Now, let us analyze the number of examples used, including those examples whose labels were not requested by Algorithm 1. Lemma 2 implies that there is a positive constant c_1 such that $\mathbb{P}(S_1) \geq c_1 b_k$: again, S_1 consists of those points that fall into an interval of length $2b_k$ after projecting onto $\hat{w}_{k-1}$. The density is lower bounded by a constant when $b_k \leq 1/9$, and we can use the bound for $1/9$ when $b_k > 1/9$. The expected number of examples that we need before we find m_k elements of S_1 is therefore at most $\frac{m_k}{c_1 b_k}$. Using a Chernoff bound, if we draw $\frac{2m_k}{c_1 b_k}$ examples, the probability that we fail to get m_k members of S_1 is at most $\exp(-m_k/6)$, which is at most $\delta/(4(1+s-k)^2)$ if C_2 is large enough. So, the total number of examples needed, $\sum_k \frac{2m_k}{c_1 b_k}$, is at most a constant factor more than

$$\sum_{k=1}^{s} 2^k\left(d + \log\left(\frac{1+s-k}{\delta}\right)\right) = O(2^s(d+\log(1/\delta))) + \sum_{k=1}^{s} 2^k \log(1+s-k)$$
$$= O\left(\frac{d+\log(1/\delta)}{\epsilon}\right) + \sum_{k=1}^{s} 2^k \log(1+s-k).$$

We can show $\sum_{k=1}^{s} 2^k \log(1+s-k) = O(1/\epsilon)$, completing the proof. ■

We conclude this section by pointing out several important facts and implications of Theorem 6 and its proof.

(1) The separator in Theorem 6 (and the one in Theorem 5) can be found in *polynomial time*, for example by using linear programming.

(2) The analysis of Theorem 6 also bounds the number of unlabeled examples needed by the active learning algorithm of Theorem 5. This shows that an algorithm can request a nearly optimally small number of labels without increasing the total number of examples required by more than a constant factor. Specifically, in round k, we only need $2^k(d + \ln[(1 + s - k)/\delta])$ unlabeled examples (whp), where $s = O(\log(1/\epsilon))$, so the total number of unlabeled examples needed over all rounds is $O(d/\epsilon + \log(1/\delta)/\epsilon)$.

6. More Distributions

In this section we consider learning with respect to a more general class of distributions. We start by providing a general set of conditions on a set $\mathcal{D}$ of distributions that is sufficient for efficient passive and active learning w.r.t. distributions in $\mathcal{D}$. We now consider nearly log-concave distributions, an interesting, more general class containing log-concave distributions, considered previously in (Applegate and Kannan, 1991) and (Caramanis and Mannor, 2007). We then prove that isotropic nearly log-concave distributions satisfy our sufficient conditions; in Appendix D, we also show how to remove the assumption that the distribution is isotropic.

Definition 7 *A set $\mathcal{D}$ of distributions is* admissible *if it satisfies the following:*

- *There exists c such that for any $D \in \mathcal{D}$ and any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v, u) \leq \mathrm{d}_D(u, v)$.*
- *For any $c_1 > 0$, there is a $c_2 > 0$ such that the following holds for all $D \in \mathcal{D}$. Let u and v be two unit vectors in R^d s.t. $\theta(u, v) = \alpha < \pi/2$. Then $\mathbb{P}_{x\sim D}[\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x), |v \cdot x| \geq c_2\alpha] \leq c_1\alpha$.*
- *There are positive constants c_3, c_4, c_5 such that, for any $D' \in \mathcal{D}$, for any projection D of D' onto a one-dimensional subspace, the density f of D satisfies $f(x) < c_3$ for all x and $f(x) > c_4$ for all x with $|x| < c_5$.*

The proofs of Theorem 5 and Theorem 6 can be used without modification to show:

Theorem 8 *If $\mathcal{D}$ is admissible, then arbitrary $f \in \mathbb{C}$ can be learned with respect to arbitrary distributions in $\mathcal{D}$ in polynomial time in the active learning model from $O((d+\log(1/\delta)+\log\log(1/\epsilon))\log(1/\epsilon))$ labeled examples, and in the passive learning model from $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ examples.*

6.1. The nearly log-concave case

Definition 9 *A density function $f : \mathbb{R}^n \to \mathbb{R}$ is β log-concave if for any $\lambda \in [0, 1]$, $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^n$, we have $f(\lambda x_1 + (1 - \lambda)x_2) \geq e^{-\beta} f(x_1)^{\lambda} f(x_2)^{1-\lambda}$.*

Clearly, a density function f is log-concave if it is 0-log-concave. An example of a $O(1)$-log-concave distribution is a mixture of two log-concave distributions whose covariance matrices are I, and whose means μ_1 and μ_2 have $||\mu_1 - \mu_2|| = O(1)$.

In this section we prove that for any sufficiently small constant $\beta \geq 0$, the class of isotropic β log-concave distribution in R^d is admissible and has light tails (this second fact is useful for analyzing the disagreement coefficient in Sections 8). In doing so we provide several new properties for such distributions, which could be of independent interest. Detailed proofs of our claims appear in Appendix D.

We start by showing that for any isotropic β log-concave density f there exists a log-concave density $\tilde{f}$ whose center is within $e(C - 1)\sqrt{Cd}$ of f's center and that satisfies $f(x)/C \leq \tilde{f}(x) \leq$

$Cf(x)$, for C as small as $e^{\beta \log d}$. The fact C depends only exponentially in $\log d$ (as opposed to exponentially in d) is key for being able to argue that such distributions have light tails.

Lemma 10 *For any isotropic β log-concave density function f there exists a log-concave density function $\tilde{f}$ that satisfies $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$ and $\left\| \int x(f(x) - \tilde{f}(x))dx \right\| \leq e(C-1)\sqrt{Cd}$, for $C = e^{\beta \lceil \log_2(d+1) \rceil}$. Moreover, we have $1/C \leq \int (u \cdot x)^2 \tilde{f}(x)dx \leq C$ for every unit vector u.*

PROOF SKETCH: Note that if the density function f is β log-concave we have that $h = \ln f$ satisfies that for any $\lambda \in [0,1]$, $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^n$, we have $h(\lambda x_1 + (1-\lambda)x_2) \geq -\beta + \lambda h(x_1) + (1-\lambda)h(x_2)$. Let $\hat{h}$ be the function whose subgraph is the convex hull of the subgraph of h.By using Caratheodory's theorem[2] we can show that $\hat{h}(x) = \max_{\sum_{i=1}^{d+1} \alpha_i = 1, \alpha_i \geq 0, x = \sum_{i=1}^{d+1} \alpha_i x_i} \sum_{i=1}^{d+1} \alpha_i h(x_i)$. This implies $h(x) \leq \hat{h}(x)$ and we can prove by induction on $\log_2(d+1)$ that $h(x) \geq \hat{h}(x) - \beta \lceil \log_2(d+1) \rceil$. If we further normalize $e^{\hat{h}}$ to make it a density function, we obtain $\tilde{f}$ that is log-concave and satisfies $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$, where $C = e^{\beta \lceil \log_2(d+1) \rceil}$. This implies that for any x we have $|f(x) - \tilde{f}(x)| \leq (C-1)\tilde{f}(x)$.

Using this fact and concentration properties of $\tilde{f}$ (in particular Lemma 2), we can show that the center of $\tilde{f}$ is close to the center of f, as desired. ∎

Theorem 11 *Assume β is a sufficiently small non-negative constant and let $\mathcal{D}$ be the set of all isotropic β log-concave distributions. (a) $\mathcal{D}$ is admissible. (b) Any $D \in \mathcal{D}$ has light tails. That is: $\mathbb{P}(||X|| > R\sqrt{Cd}) < Ce^{-R+1}$, for $C = e^{\beta \lceil \log_2(d+1) \rceil}$.*

PROOF SKETCH: (a) Choose $D \in \mathcal{D}$. As in Lemma 3, consider the plane determined by u and v and let $proj_{u,v}(x)$ denote the projection operator that given $x \in R^d$, orthogonally projects x onto this plane. If $D_2 = proj_{u,v}(D)$ then $\mathrm{d}_D(u,v) = \mathrm{d}_{D_2}(u',v')$. By using the Prekopa-Leindler inequality (Gardner, 2002) one can show that D_2 is β log-concave (see e.g., (Caramanis and Mannor, 2007)). Moreover, if D is isotropic, than D_2 is isotropic as well. By Lemma 10 we know that there exists a C-isotropic log-concave distribution $\tilde{D}_2$ centered at z, $\|z\| \leq \epsilon$, satisfying $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$ and $1/C \leq \int (u \cdot x)^2 f(x)dx \leq C$ for every unit vector u, for constants $C = e^{\beta}$ and $\epsilon = e(C-1)\sqrt{2C}$. For β sufficiently small we have $(1/20+\epsilon)/\sqrt{1/C - \epsilon^2} \leq 1/9$. Using this, by applying the whitening transform (see Theorem 16 in Appendix D), we can show $\tilde{f}_2(x) \geq c$, for $\|x\| \leq 1/20$, which implies $f_2(x) \geq c/C$, for $\|x\| \leq 1/20$. Using a reasoning as in Lemma 3 we get $c\theta(v,u) \leq \mathrm{d}_D(u,v)$. The generalization of Theorem 4 follows from a similar proof, except using Theorem 16. The density bounds in the $n = 1$ case also follow from Theorem 16 as well.

(b) Since X is isotropic, we have $\mathbb{E}_f[X \cdot X] = d$ (where f is its associated density). By Lemma 10, there exists a log-concave density $\tilde{f}$ such that $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$, for $C = e^{\beta \lceil \log_2(d+1) \rceil}$. This implies $E_{\tilde{f}}[X \cdot X] \leq Cd$. By Lemma 2 we get that that under $\tilde{f}$, $\mathbb{P}(||X|| > R\sqrt{Cd}) < e^{-R+1}$, so under f we have $\mathbb{P}(||X|| > R\sqrt{Cd}) < Ce^{-R+1}$. ∎

Using Theorem 8 and Theorem 11(a) we obtain:

Theorem 12 *Let $\beta \geq 0$ be a sufficiently small constant. Assume that D is an isotropic β log-concave distribution in R^d. Then arbitrary $f \in \mathbb{C}$ can be learned with respect to D in polynomial time in the active learning model from $O((d + \log(1/\delta) + \log\log(1/\epsilon))\log(1/\epsilon))$ labeled examples, and in the passive learning model from $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ examples.*

2. Caratheodory's theorem states that if a point x of R^d lies in the convex hull of a set P, then there is a subset $\hat{P}$ of P consisting of $d+1$ or fewer points such that x lies in the convex hull of $\hat{P}$.

7. Lower Bounds

In this section we give lower bounds on the label complexity of passive and active learning of homogeneous linear separators when the underlying distribution is β log-concave, for a sufficiently small constant β. These lower bounds are information theoretic, applying to any procedure, that might not be necessarily computationally efficient. The proof is in Appendix E.

Theorem 13 *For a small enough constant β we have: (1) for any β log-concave distribution D whose covariance matrix has full rank, the sample complexity of learning origin-centered linear separators under D in the passive learning model is $\Omega\left(\frac{d}{\epsilon}+\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$; (2) the sample complexity of active learning of linear separators under β log-concave distributions is $\Omega\left(d\log\left(\frac{1}{\epsilon}\right)\right)$.*

Note that, if the covariance matrix of D does not have full rank, the number of dimensions is effectively less than d, so our lower bound essentially applies for all log-concave distributions.

8. The inseparable case: Disagreement-based active learning

We consider two closely related distribution dependent capacity notions: the Alexander capacity and the disagreement coefficient; they have been widely used for analyzing the label complexity of non-aggressive active learning algorithms (Hanneke, 2007; Dasgupta et al., 2007; Koltchinskii, 2010; Hanneke, 2011; Beygelzimer et al., 2010). We begin with the definitions. For $r>0$, define $B(w,r)=\{u\in\mathbb{C}:\mathbb{P}_D(\text{sign}(u\cdot x)\neq\text{sign}(w\cdot x))\leq r\}$. For any $\mathcal{H}\subseteq\mathbb{C}$, define the region of disagreement as $\text{DIS}(\mathcal{H})=\{x\in X:\exists w,u\in\mathcal{H}\text{ s.t. }\text{sign}(u\cdot x)\neq\text{sign}(w\cdot x))\}$. Define the Alexander capacity function $\text{cap}_{w^*,D}(\cdot)$ for $w^*\in\mathbb{C}$ w.r.t. D as: $\text{cap}_{w^*,D}(r)=\frac{\mathbb{P}_D(\text{DIS}(B(w^*,r)))}{r}$. Define the disagreement coefficients for $w^*\in\mathbb{C}$ w.r.t. D as: $\text{dis}_{w^*,D}(\epsilon)=\sup_{r\geq\epsilon}[\text{cap}_{w^*,D}(r)]$.

The following is our bound in the disagreement coefficient. Its proof is in Appendix F.

Theorem 14 *Let $\beta\geq 0$ be a sufficiently small constant. Assume that D is an isotropic β log-concave distribution in R^d. For any w^*, for any ϵ, $\text{cap}_{w^*,D}(\epsilon)$ is $O(d^{1/2+\frac{\beta}{2\ln 2}}\log(1/\epsilon))$. Thus $\text{dis}_{w^*,D}(\epsilon)=O(d^{1/2+\frac{\beta}{2\ln 2}}\log(1/\epsilon))$.*

Theorem 14 immediately leads to concrete bounds on the label complexity of several algorithms in the literature (Hanneke, 2007; Cohn et al., 1994; Balcan et al., 2006; Koltchinskii, 2010; Dasgupta et al., 2007). For example, by composing it with a result of (Dasgupta et al., 2007), we obtain a bound of $\tilde{O}(d^{3/2}(\log^2(1/\epsilon)+(\nu/\epsilon)^2))$ for agnostic active learning when D is isotropic log-concave in R^d; that is we only need $\tilde{O}(d^{3/2}(\log^2(1/\epsilon)+(\nu/\epsilon)^2)))$ label requests to output a classifier of error at most $\nu+\epsilon$, where $\nu=\min_{w\in\mathbb{C}}\text{err}(w)$.

9. The Tsybakov condition

In this section we consider a variant of the Tsybakov noise condition (Mammen and Tsybakov, 1999). We assume that the classifier h that minimizes $\mathbb{P}_{(x,y)\sim D_{XY}}(h(x)\neq y)$ is a linear classifier, and that, for the weight vector w^* of the optimal classifier, there exist known parameters $\alpha,a>0$ such that, for all w, we have

$$a(\text{d}_D(w,w^*))^{1/(1-\alpha)}\leq\text{err}(w)-\text{err}(w^*).$$

By generalizing Theorem 4 so that it provides a stronger bound for larger margins, and combining the result with the other lemmas of this paper and techniques from (Balcan et al., 2007), we get the following.

Theorem 15 *Let $s = O(\log(1/\epsilon))$. Assume that the distribution D_{XY} satisfies the Tsybakov noise condition for constants $\alpha \in [0, 1)$ and $a \geq 0$, and that the marginal D on R^d is isotropic log-concave. (1) If $\alpha = 0$, we can find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$ using $O(\log(1/\epsilon))(d + \log(s/\delta))$ labeled examples in the active learning model, and $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ labeled examples in the passive learning model. (2) If $\alpha > 0$, we can find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$ using $O((1/\epsilon)^{2\alpha} \log^2(1/\epsilon))(d + \log(s/\delta))$ labeled examples in the active learning model.*

In the case $\alpha = 0$ (that is more general than the Massart noise condition) our analysis leads to optimal bounds for active and passive learning of linear separators under log-concave distributions, improving the dependence on d over previous best known results (Hanneke and Yang, 2012; Giné and Koltchinskii, 2006). Our analysis for Tsybakov noise ($\alpha \geq 0$) leads to bounds on active learning with improved dependence on d over previous known results (Hanneke and Yang, 2012) in this case as well. Proofs and further details appear in Appendix G.

10. Discussion and Open Questions

The label sample complexity of our active learning algorithm for learning homogeneous linear separators under isotropic logconcave distributions is $O((d+\log(1/\delta)+\log\log(1/\epsilon))\log(1/\epsilon))$, while our lower bound for this setting is $\Omega\left(d \log\left(\frac{1}{\epsilon}\right)\right)$. Our upper bound is achieved by an algorithm that uses a polynomial number of unlabeled training examples, and polynomial time. If an unbounded amount of computation time and an unbounded number of unlabeled examples are available, it seems to be easy to learn to accuracy ϵ using $O(d \log(1/\epsilon))$ label requests, no matter what the value of δ. (Roughly, the algorithm can construct an ϵ-cover to initialize a set of candidate hypotheses, then repeatedly wait for an unlabeled example that evenly splits the current list of candidates, and ask its label, eliminated roughly half of the candidates.) It would be interesting to know what is the best label complexity for a polynomial-time algorithm, or even an algorithm that is constrained to use a polynomial number of unlabeled examples.

Conceptually, our analysis of ERM for passive learning under (nearly) log-concave distributions is based on a more aggressive localization than those considered previously in the literature. It would be very interesting to extend this analysis, as well as our analysis for active learning to arbitrary distributions, and more general concept spaces.

Acknowledgements We thank Steve Hanneke for a number of useful discussions.
This work was supported in part by NSF grant CCF-0953192, AFOSR grant FA9550-09-1-0538, and a Microsoft Research Faculty Fellowship.

References

K.S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 1987.

N. Alon. A non-linear lower bound for planar epsilon-nets. *FOCS*, pages 341–346, 2010.

D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *STOC*, 1991.

P. Assouad. Plongements lipschitziens dans. *R* . *Bull. Soc. Math. France*, 111(4):429–448, 1983.

M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *COLT*, 2007.

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, 2008.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 2005.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 2005.

N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *JCSS*, 2009.

C. Caramanis and S. Mannor. An inequality for nearly log-concave distributions with applications to learning. *IEEE Transactions on Information Theory*, 2007.

R. Castro and R. Nowak. Minimax bounds for active learning. In *COLT*, 2007.

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 2010.

K. L. Clarkson and K. Varadarajan. Improved approximation algorithms for geometric set cover. *Discrete Comput. Geom.*, 37(1):43–58, 2007.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *ICML*, 1994.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, 2005.

S. Dasgupta. Active learning. *Encyclopedia of Machine Learning*, 2011.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *COLT*, 2005.

S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007.

O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 2012.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 1989.

Y. Freund, H.S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

E. J. Friedman. Active learning for smooth problems. In *COLT*, 2009.

R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.*, 2002.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

A. Gonen, S. Sabato, and S. Shalev-Shwartz. Efficient pool-based active learning of halfspaces. In *ICML*, 2013.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke and L. Yang. Surrogate losses in passive and active learning, 2012. http://arxiv.org/abs/1207.3772.

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

David Haussler and Emo Welzl. Epsilon nets and simplex range queries. *Disc. Comp. Geometry*, 2: 127–151, 1987.

A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual Symposium on the Foundations of Computer Science (FOCS)*, 2005.

M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *JMLR*, 2009a.

A. R. Klivans, P. M. Long, and A. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *RANDOM*, 2009b.

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

J. Komlós, J. Pach, and G. Woeginger. Almost tight bounds on epsilon-nets. *Discrete and Computational Geometry*, 7:163–173, 1992.

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 1993.

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

P. M. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 2003.

L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 2007.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27: 1808–1829, 1999.

P. Massart and E. Nedelec. Risk bounds for statistical learning. *The Annals of Statistics*, 2006.

S. Mendelson. Estimating the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

R. Nowak. The Geometry of Generalized Binary Search. *IEEE Transactions on Information Theory*, 2011.

J. Pach and P.K. Agarwal. *Combinatorial Geometry*. John Wiley and Sons, 1995.

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *NIPS*, 2011.

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

S. van de Geer. *Empirical processes in M-estimation*. Cambridge Series in Statistical and Probabilistic Methods, 2000.

A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, 1996.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. N. Vapnik. *Estimation of Dependencies based on Empirical Data*. Springer Verlag, 1982.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *JACM*, 57(6), 2010.

Appendix A. Additional Related Work

Learning with noise. Alexander Capacity and the Disagreement Coefficient Roughly speaking the Alexander capacity (Alexander., 1987; Giné and Koltchinskii, 2006) quantifies how fast the region of disagreement of the set of classifiers at distance r of the optimal classifier collapses as a function r; [3] the disagreement coefficient (Hanneke, 2007) additionally involves the supremum of r over a range of values. (Friedman, 2009) provides guarantees on these quantities (for sufficiently small r) for general classes of functions in $\mathbb{R}^d$ if the underlying data distribution is sufficiently smooth. Our analysis implies much tighter bounds for linear separators under log-concave distributions (matching what was known for the much less general case of nearly uniform distribution over the unit sphere); furthermore, we also analyze the nearly log-concave case where we allow an arbitrary number of discontinuities, a case not captured by the (Friedman, 2009) conditions at all. This immediately implies concrete bounds on the labeled data complexity of several algorithms in the literature including the A^2 algorithm (Balcan et al., 2006) and the DHM algorithm (Dasgupta et al.,

3. The region of disagreement $\mathrm{DIS}(\mathbb{C})$ of a set of classifiers $\mathbb{C}$ is the of set of instances x s.t. for each $x \in \mathrm{DIS}(\mathbb{C})$ there exist two classifiers $f, g \in \mathbb{C}$ that disagree about the label of x.

2007), with implications for the purely agnostic case (i.e., arbitrary forms of noise), as well as the Koltchinskii's algorithm (Koltchinskii, 2010) and the CAL algorithm (Balcan et al., 2006; Hanneke, 2007, 2011). Furthermore, in the realizable case and under Tsybakov noise, we show even better bounds, by considering aggressive active learning algorithms.

Note that as opposed to the realizable case, all existing active learning algorithms analyzed under Massart and Tsybakov noise conditions using the learning model analyzed in this paper (including our algorithms in Theorem 15), as well as those for the agnostic setting, are not known to run in time $\mathrm{poly}(d, 1/\epsilon)$. In fact, even ignoring the optimality of sample complexity, there are no known algorithms for passive learning that run in time $\mathrm{poly}(d, 1/\epsilon)$ for general values of ϵ, even for the Massart noise condition and under log-concave distributions. Existing works on agnostic passive learning under log-concave distributions either provide running times $d^{\mathrm{poly}(1/\epsilon)}$ (e.g., the work of (Kalai et al., 2005)) or can only achieve values of ϵ that are significantly larger than the noise rate (Klivans et al., 2009a).

Other Work on Active Learning Several papers (Cesa-Bianchi et al., 2010; Dekel et al., 2012) present efficient online learning algorithms in the selective sampling framework, where labels must be actively queried before they are revealed. Under the assumption that the label conditional distribution is linear function determined by a fixed target vector, they provide bounds on the regret of the algorithm and on the number of labels it queries when faced with an adaptive adversarial strategy of generating the instances. As pointed by (Dekel et al., 2012), these results can be converted to a statistical setting when the instances x_t are drawn i.i.d and they further assume a margin condition. In this setting they obtain exponential improvement in label complexity over passive learning. While very interesting, these results are incomparable to ours; their techniques significantly exploit the linear noise condition to get these improvements – note that such an improvement would not be possible in the realizable case (as pointed for example in (Gonen et al., 2013)).

(Nowak, 2011) considers an interesting abstract "generalized binary search" problem with applications to active learning; while these results apply for more general concept spaces, it is not clear how to implement the resulting procedures in polynomial time and by using access to only a polynomial number of unlabeled samples from the underlying distribution (as required by the active learning model). Another interesting recent work is that of (Gonen et al., 2013), which study active learning of linear separators via an aggressive algorithm using a margin condition, using a general approximation guarantee on the number of labels requested; note that while these results work for potentially more general distributions, as opposed to ours, they do not come with explicit (tight) bounds on the label complexity.

ϵ-nets, Learning, and Geometry Small ϵ-nets are useful for many applications, especially in Computational Geometry (see (Pach and Agarwal, 1995)). The same fundamental techniques of (Vapnik and Chervonenkis, 1971; Vapnik, 1982) have been applied to establish the existence of small ϵ-nets (Haussler and Welzl, 1987) and to bound the sample complexity of learning (Vapnik, 1982; Blumer et al., 1989), and a number of interesting upper and lower bounds on the smallest possible size of ϵ-nets have been obtained (Komlós et al., 1992; Clarkson and Varadarajan, 2007; Alon, 2010).

Our analysis implies a $O(d/\epsilon)$ upper bound on the size of an ϵ-net for a set of regions of disagreement between all possible linear classifiers and the target, when the distribution is zero-mean and log-concave. In particular, since in Theorem 6 we prove that any hypothesis consistent with the training data has error rate $\leq \epsilon$ with probability $1-\delta$, setting δ to a constant gives a proof of a $O(d/\epsilon)$ bound on the size of an ϵ-net for the following set: $\{\{x : (w \cdot x)(w^* \cdot x) < 0\} \ : \ w \in R^n\}$.

Appendix B. Proof of Lemma 3

Lemma 3. *Assume D is an isotropic log-concave in R^d. Then there exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v,u) \leq \mathrm{d}_D(u,v)$.*

Proof Consider two unit vectors u and v. Let $proj_{u,v}(x)$ denote the projection operator that, given $x \in R^d$, orthogonally projects x onto the plane determined by u and v. That is, if we define an orthogonal coordinate system in which coordinates $1,2$ lie in this plane and coordinates $3,\ldots,d$ are orthogonal to this plane, then $x' = proj_{u,v}(x_1,\ldots,x_d) = (x_1,x_2)$. Also, given distribution D over R^d, define $proj_{u,v}(D)$ to be the distribution given by first picking $x \sim D$ and then outputting $x' = proj_{u,v}(x)$. That is, $proj_{u,v}(D)$ is just the marginal distribution over coordinates $1,2$ in the above coordinate system. Notice that if $x' = proj_{u,v}(x)$ then $u \cdot x = u' \cdot x'$ where $u' = proj_{u,v}(u)$ and $v' = proj_{u,v}(v)$. So, if $D_2 = proj_{u,v}(D)$ then $\mathrm{d}_D(u,v) = \mathrm{d}_{D_2}(u',v')$.

By Lemma 2(c), we have that if D is isotropic and log-concave, then D_2 is as well. Let A to be the region of disagreement between u' and v' intersected with the ball of radius $1/9$ in R^2. The probability mass of A under D_2 is at least the volume of A times $\inf_{x\in A} D_2(x)$. So, using Lemma 2(b)

$$\mathrm{d}_{D_2}(u',v') \geq \mathrm{vol}(A) \inf_{x\in A} D_2(x) \geq c\theta(u,v),$$

as desired. ■

Appendix C. Passive Learning

Theorem 6. *Assume that D is zero mean and log-concave in R^d. There exists an absolute constant C_3 s.t. for $d \geq 4$, and for any $\epsilon,\delta > 0$, $\epsilon < 1/4$, any algorithm that outputs a hypothesis that correctly classifies $m = \frac{C_3(d+\log(1/\delta))}{\epsilon}$ examples finds a separator of error at most ϵ with probability $\geq 1-\delta$.*

Proof First, let us prove the theorem in the case that D is isotropic. We will then treat the general case at the end of the proof.

While our analysis will ultimately provide a guarantee for any learning algorithm that always outputs a consistent hypothesis, we will use intermediate hypothesis of Algorithm 1 in the analysis.

Let c be the constant from Lemma 3. While proving Theorem 5, we proved that, if Algorithm 1 is run with $b_k = \frac{C_1}{2^k}$ and $m_k = C_2\left(d + \ln\frac{1+s-k}{\delta}\right)$, that for all $k \leq s$, with probability $\geq 1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$ any $\hat{w}$ consistent with the data in $W(k)$ has $\mathrm{err}(\hat{w}) \leq c2^{-k}$. Thus, after $s = O(\log(1/\epsilon))$ iterations, with probability at least $\geq 1-\delta$, any linear classifier consistent with *all* the training data has error $\leq \epsilon$, since any such classifier is consistent with the examples in $W(s)$.

Now, let us analyze the number of examples used, including those examples whose labels were not requested by Algorithm 1. Lemma 2 implies that there is a positive constant c_1 such that $\mathbb{P}(S_1) \geq c_1 b_k$: again, S_1 consists of those points that fall into an interval of length $2b_k$ after projecting onto $\hat{w}_{k-1}$. The density is lower bounded by a constant when $b_k \leq 1/9$, and we can use the bound for $1/9$ when $b_k > 1/9$.

The expected number of examples that we need before we find m_k elements of S_1 is therefore at most $\frac{m_k}{c_1 b_k}$. Using a Chernoff bound, if we draw $\frac{2m_k}{c_1 b_k}$ examples, the probability that we fail to

get m_k members of S_1 is at most $\exp(-m_k/6)$, which is at most $\delta/(4(1+s-k)^2)$ if C_2 is large enough. So, the total number of examples needed, $\sum_k \frac{2m_k}{c_1 b_k}$, is at most a constant factor more than

$$\begin{aligned}
&\sum_{k=1}^{s} 2^k \left(d + \log\left(\frac{1+s-k}{\delta}\right)\right) \\
&= O(2^s(d+\log(1/\delta))) + \sum_{k=1}^{s} 2^k \log(1+s-k) \\
&= O\left(\frac{d+\log(1/\delta)}{\epsilon}\right) + \sum_{k=1}^{s} 2^k \log(1+s-k).
\end{aligned}$$

We claim that $\sum_{k=1}^{s} 2^k \log(1+s-k) = O(1/\epsilon)$. We have

$$\begin{aligned}
&\sum_{k=1}^{s} 2^k \log(1+s-k) \leq \sum_{k=1}^{s} 2^k (3+s-k) \\
&\leq \int_{k=1}^{s+1} 2^k (3+s-k) \\
&\qquad \text{(since } 2^k(3+s-k) \text{ is increasing for } k \leq s+1) \\
&= \frac{2(2^s-1)(1+\ln(4)) - s\ln 2}{\ln^2 2} = O(1/\epsilon),
\end{aligned}$$

completing the proof in the case that D is isotropic.

Now let us treat the case in which D is not isotropic. Suppose that Σ is the covariance matrix of D, so that $\Sigma^{-1/2}$ is the "whitening transform". Suppose, for $m = \frac{C_3(d+\log(1/\delta))}{\epsilon}$, an algorithm is given a sample S of examples $(x_1, y_1), ..., (x_m, y_m)$ for $x_1, ..., x_m$ drawn according to D, and y_m labeled by a target hypothesis with weight vector v. Note that w is consistent with S if and only if $w^T\Sigma^{1/2}$ is consistent with $(\Sigma^{-1/2}x_1, y_1), ..., (\Sigma^{-1/2}x_m, y_m)$ (so those examples are consistent with $v^T\Sigma^{1/2}$). So our analysis of the isotropic case implies that, with probability $1-\delta$, for any w consistent with $(x_1, y_1), ..., (x_m, y_m)$, we have

$$\mathbb{P}(\text{sign}((w^T\Sigma^{1/2})(\Sigma^{-1/2}x)) \neq \text{sign}((v^T\Sigma^{1/2})(\Sigma^{-1/2}x))) \leq \epsilon,$$

which of course means that $\mathbb{P}(\text{sign}(w^T x) \neq \text{sign}(v^T x)) \leq \epsilon$. ■

Appendix D. More Distributions

D.1. Isotropic Nearly Log-concave distributions

Lemma 10. *For any isotropic β log-concave density function f there exists a log-concave density function $\tilde{f}$ that satisfies $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$ and $\left\| \int x(f(x) - \tilde{f}(x))dx \right\| \leq e(C-1)\sqrt{Cd}$, for $C = e^{\beta\lceil \log_2(d+1)\rceil}$. Moreover, we have $1/C \leq \int (u\cdot x)^2 \tilde{f}(x)dx \leq C$ for every unit vector u.*

Proof Note that if the density function f is β log-concave we have that $h = \ln f$ satisfies that for any $\lambda \in [0,1]$, $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^n$, we have $h(\lambda x_1 + (1-\lambda)x_2) \geq -\beta + \lambda h(x_1) + (1-\lambda)h(x_2)$.

Let $\hat{h}$ be the function whose subgraph is the convex hull of the subgraph of h. That is, $\hat{h}(x)$ is the maximum of all values of $\sum_{i=1}^{k} \alpha_i h(u_i)$ for any $u_1, ..., u_k \in R^d$ and $\alpha_1, ..., \alpha_k \in [0,1]$ such that $\sum_{i=1}^{k} \alpha_i = 1$ and $x = \sum_{i=1}^{k} \alpha_i u_i$. Note that, if the components of u_i are $u_{i,1}, ..., u_{1,d}$, we can get $\hat{h}(x)$ by starting with

$$T = \{(u_{1,1}, ..., u_{1,d}, h(u_1)), ..., (u_{k,1}, ..., u_{k,d}, h(u_k))\}$$

taking the convex combination of the members of T with mixing coefficients $\alpha_1, ..., \alpha_k$, and then reading off the last component. Caratheodory's theorem[4] implies that we can get the same result using a mixture of at most $d+1$ members of T. In other words, we can assume without loss of generality that $k = d+1$, so that

$$\hat{h}(x) = \max_{\sum_{i=1}^{d+1} \alpha_i = 1, \alpha_i \geq 0, x = \sum_{i=1}^{d+1} \alpha_i x_i} \sum_{i=1}^{d+1} \alpha_i h(x_i). \tag{8}$$

Because of the case where $(\alpha_1, ..., \alpha_{d+1})$ concentrates all its weight on one component, we have $h(x) \leq \hat{h}(x)$.

We also claim that

$$h(x) \geq \hat{h}(x) - \beta \lceil \log_2(d+1) \rceil. \tag{9}$$

We will prove this by induction on $\log_2(d+1)$, treating the case in which $d+1$ is a power of 2. (By padding with zeroes if necessary, we may assume without loss of generality that $d+1$ is a power of 2.) The base case, in which $d = 1$, follows immediately from the definitions. Let $k = d+1$. Assume that $x = a_1 x_1 + a_2 x_2 + ... + a_k x_k$, $\sum_{i=1}^{k} a_i = 1$, $a_i \geq 0$. We can write this as:

$$x = (a_1 + a_2)x_{1,2} + (a_3 + a_4)x_{3,4} + ...(a_{k-1} + a_k)x_{k-1,k}$$

where $x_{i,i+1} = \frac{a_i}{a_i + a_{i+1}} x_i + \frac{a_{i+1}}{a_i + a_{i+1}} x_{i+1}$, for all i. Now, by induction we have:

$$\begin{aligned}
h(x) &\geq -\beta \log(k/2) + (a_1 + a_2)h(x_{1,2}) + \\
&\quad ... + (a_{k-1} + a_k)h(x_{k-1,k}) \\
&\geq -\beta \log(k/2) \\
&- (a_1 + a_2)\beta + a_1 h(x_1) + a_2 h(x_2) \\
&- (a_3 + a_4)\beta + a_3 h(x_3) + a_4 h(x_4) + \\
&\quad ... \\
&- (a_{k-1} + a_k)\beta + a_{k-1} h(x_{k-1}) + a_k h(x_k) \\
&= -\beta \log(k) + a_1 h(x_1) + a_2 h(x_2) + a_3 h(x_3) + ...a_k h(x_k).
\end{aligned}$$

The last inequality follows from the fact that $\sum_{i=1}^{n} a_i = 1$.

So, we have proved (9). If we further normalize $e^{\hat{h}}$ to make it a density function, we obtain $\tilde{f}$ that is log-concave and satisfies $f(x)/C \leq \tilde{f}(x) \leq Cf(x)$, where $C = e^{\beta \lceil \log_2(d+1) \rceil}$. This implies that for any x we have $|f(x) - \tilde{f}(x)| \leq (C-1)\tilde{f}(x)$.

4. Caratheodory's theorem states that if a point x of R^d lies in the convex hull of a set P, then there is a subset $\hat{P}$ of P consisting of $d+1$ or fewer points such that x lies in the convex hull of $\hat{P}$.

We now show that the center of $\tilde{f}$ is close to the center of f. We have:

$$\begin{aligned}\left\|\int x(f(x)-\tilde{f}(x))dx\right\| &\leq \int \|x\|\,|f(x)-\tilde{f}(x)|dx \\ \leq (C-1)\int \|x\|\,\tilde{f}(x)dx &= (C-1)\int_{r=0}^{\infty}\mathbb{P}_{\tilde{f}}[\|X\|\geq r]dr.\end{aligned}$$

Using concentration properties of $\tilde{f}$ (in particular Lemma 2) we get

$$\begin{aligned}\left\|\int x(f(x)-\tilde{f}(x))dx\right\| &\leq (C-1)\int_{r=0}^{\infty} e^{-\frac{r}{\sqrt{Cd}}+1}dr \\ &= e(C-1)\sqrt{Cd},\end{aligned}$$

as desired. ■

Theorem 16 *(i) Let $f: R^2 \to R$ be the density function of a log-concave distribution centered at z and with covariance matrix $A = \mathbb{E}_f[(X-z)(X-z)^T]$. Assume f satisfies $\|z\| \leq \xi$ and $1/C \leq \int (u\cdot x)^2 f(x)dx \leq C$ for every unit vector u, for $C \geq 1$ constant close to 1. We have: (a) Assume $(1/20+\xi)/\sqrt{1/C-\xi^2} \leq 1/9$. Then there exist an universal constant c s.t. we have $f(x) \geq c$, for all x with $0 \leq ||x|| \leq 1/20$. (b) Assume $C \leq 1+1/5$. There exist universal constants c_1 and c_2 such that $f(x) \leq C_1 \exp(-C_2||x||)$ for all x.*

(ii) Let $f: R \to R$ be the density function of a log-concave distribution centered at ξ with standard deviation $\sigma = \sqrt{\mathrm{Var}_f(X)}$. Then $f(x) \leq 1/\sigma$ for all x. If furthermore f satisfies $1/C \leq \mathbb{E}_f[X^2] \leq C$ for $C \geq 1$ and $\xi/\sqrt{1/C-\xi^2} \leq 1/9$, then we have $f(0) \geq c$ for some universal constant c.

Proof (i) Let $Y = A^{-1/2}(X-z)$. Then Y is a log-concave distribution in the isotropic position. Moreover, the density function of g is given by $g(y) = \det(A^{1/2})f(A^{1/2}y+z)$. Let $M = \mathbb{E}[XX^T]$. We have

$$A = \mathbb{E}[(X-z)(X-z)^T] = \mathbb{E}[XX^T] - zz^T = M - zz^T.$$

Also, the fact $1/C \leq \int (u\cdot x)^2 f(x)dx \leq C$ for every unit vector u is equivalent to

$$1/C \leq u^T\mathbb{E}[XX^T]u \leq C$$

for every unit vector u. Using $v=(1,0)$, $v=(0,1)$, and $v=(1/\sqrt{2},1/\sqrt{2})$ we get that $M_{1,1} \in [1/C, C]$, $M_{2,2} \in [1/C, C]$, and $M_{1,2} = M_{2,1} \in [1/C - C, C - 1/C]$. We also have $\|z\| \leq \xi$ and $\det(A^{1/2}) = \sqrt{\det(A)}$. All these imply that

$$\sqrt{(1/C-\xi^2)^2 - (C-1/C)^2} \leq \det(A^{1/2}) \leq C.$$

(a) For $x = A^{1/2}y+z$ we have $\|x-z\|^2 = (x-z)(x-z)^T = \|y\|^2 v^T Av$, where $v = (1/\|y\|)y$ is a unit vector, so $\|y\| \leq \|x-z\|/\sqrt{1/C-\xi^2}$. If $\|x\| \leq 1/20$ we have $\|y\| \leq 1/9$, so by Lemma 2 we have $g(y) \geq c_1$, so $f(y) \geq c$, for some universal constants c_1, c_2, as desired.

(b) We have $f(x) = \frac{1}{\det(A^{1/2})} g(A^{-1/2}(x-z))$. By Lemma 2 (b) we have

$$f(x) \leq \frac{1}{\det A^{1/2}} \exp\left[-c\left\|A^{-1/2}(x-z)\right\|\right].$$

By the triangle inequality we further obtain:

$$f(x) \leq \frac{1}{\det\left(A^{1/2}\right)} \exp\left[c\left\|A^{-1/2}z\right\|\right] \exp\left[-c\left\|A^{-1/2}x\right\|\right].$$

For $C \leq 1+1/5$, we can show that $\left\|A^{-1/2}x\right\| \geq (1/\sqrt{2})\|x\|$. It is enough to show $\left\|A^{-1/2}x\right\|^2 \geq (1/2)\|x\|^2$, or that $2\|v\|^2 \geq \left\|A^{1/2}v\right\|$, where $v = A^{-1/2}x$ (so $x = A^{1/2}v$). This is equivalent to $2v^Tv \geq v^TAv$, which is true since the matrix $2I - A$ is positive semi-definite.

(ii) Define $Y = (X - z)/\sigma$. We have $\mathbb{E}[Y] = 0$ and $\mathbb{E}[Y^2] = 1$. The density g of Y is given by $g(y) = \sigma f(\sigma y + z)$. Now, since g is isotropic and log-concave, we can apply Lemma 2(e) to g. So $g(y) \leq 1$ for all y. So, $\sigma f(\sigma y + z) \leq 1$ for all y, which implies $f(x) \leq 1/\sigma$ for all x. The second part follows as in Theorem 16. ■

D.2. More covariance matrices

In this section, we extend Theorem 5 to the case of arbitrary covariance matrices.

Theorem 17 *If all distributions in $\mathcal{D}$ are zero-mean and log-concave in R^d, then arbitrary $f \in \mathbb{C}$ be learned in polynomial time from arbitrary distributions in $\mathcal{D}$ in the active learning model from $O((d + \log(1/\delta) + \log\log(1/\epsilon))\log(1/\epsilon))$ labeled examples, and in the passive learning model from $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ examples.*

Our proof is through a series of lemma. First, (Lovasz and Vempala, 2007) have shown how to reduce to the nearly isotropic case.

Lemma 18 ((Lovasz and Vempala, 2007)) *For any constant $\kappa > 0$, there is a polynomial time algorithm that, given polynomially many samples from a log-concave distribution D, outputs an estimate Σ of the covariance matrix of D such that, with probability $1 - \delta$ the distribution D' obtained by sampling x from D and producing $\Sigma^{-1/2}x$ has $\frac{1}{1+\kappa} \leq \mathbb{E}((u \cdot x)^2) \leq 1 + \kappa$ for all unit vectors u.*

As a result of Lemma 18, we can assume without loss of generality that the distribution D satisfies $\frac{1}{1+\kappa} \leq \mathbb{E}((u \cdot x)^2) \leq 1 + \kappa$ for an arbitrarily small constant κ. By Theorem 16, this implies that, without loss of generality, there are constants $c_1, ..., c_4$ such that, for the density f of any one or two-dimensional marginal D' of D, we have

$$f(x) \geq c_1 \text{ for all } x \text{ with } ||x|| \leq c_2, \tag{10}$$

and for all x,

$$f(x) \leq c_3 \exp(-c_4||x||). \tag{11}$$

We will show that these imply that $\mathcal{D}$ is admissible.

Lemma 19 *(a) There exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v, u) \leq \mathrm{d}_D(u, v)$.*

(b) For any $c_6 > 0$, there is a $c_7 > 0$ such that the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u, v) = \alpha < \pi/2$. Then

$$\mathbb{P}_{x\sim D}[\text{sign}(u \cdot x) \neq \text{sign}(v \cdot x), |v \cdot x| \geq c_7\alpha] \leq c_6\alpha.$$

Proof (a) Projecting D onto a subspace can only reduce the norm of its mean, and its variance in any direction. Therefore, as in the proof of Lemma 3, we may assume without loss of generality that $d = 2$. Here, let us define A to be the region of disagreement between u' and v' intersected with the ball B_{c_2} of radius c_2 in R^2. Then we have $\mathrm{d}_{D_2}(u', v') \geq \mathrm{vol}(A) \inf_{x \in A} D_2(x) \geq \mathrm{vol}(B_{c_2}) c_1 \theta(u, v)$. (b) This proof basically amounts to observing that everything that was needed for the proof of Theorem 4 is true for D, because of (10) and (11). ■

Armed with Lemma 19, to prove Theorem 17, we can just apply Theorem 8.

Appendix E. Lower Bounds

The proof of our lower bounds (Theorem 13) relies on a lower bound on the packing numbers $M_D(\mathbb{C}, \epsilon)$. Recall that the ϵ-packing number, $M_D(\mathbb{C}, \epsilon)$, is the maximal cardinality of an ϵ-separated set with classifiers from $\mathbb{C}$, where we say that $w_1, ..., w_N$ are ϵ-separated w.r.t $\mathcal{D}$ if $\mathrm{d}_D(w_i, w_j) > \epsilon$ for any $i \neq j$.

Lemma 20 *There is a positive constant c such that, for all $\beta < c$, the following holds. Assume that D is β log-concave in R^d, and that its covariance matrix has full rank. For all sufficiently small ϵ, $d \in N$, we have $M_D(\mathbb{C}, \epsilon) \geq \frac{\sqrt{d}}{2} \left(\frac{c}{2\epsilon}\right)^{d-1} - 1$.*

Proof We first prove the lemma in the case that D is isotropic. The proof in this case follows the outline of a proof for the special case of the uniform distribution in (Long, 1995).

Let UBALL_d be the uniform distribution on the surface of the unit ball in $\mathbb{R}^d$. By Theorem 11, there exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(v, u) \leq \mathrm{d}_D(u, v)$. This implies that for a fixed u the probability that a randomly chosen v has $\mathrm{d}_D(u, v) \leq \epsilon$ is upper bounded by the volume of those vectors in the interior of the unit ball whose angle is at most ϵ/c divided by the volume of the unit ball. Using known bounds on this ratio (see (Long, 1995)) we have $\mathbb{P}_{v \in \mathrm{UBALL}_d}[\mathrm{d}_D(u, v) \leq \epsilon] \leq \frac{1}{\sqrt{d}} \left(\frac{2\epsilon}{c}\right)^{d-1}$, so $\mathbb{P}_{u,v \in \mathrm{UBALL}_d}[\mathrm{d}_D(u, v) \leq \epsilon] \leq \frac{1}{\sqrt{d}} \left(\frac{2\epsilon}{c}\right)^{d-1}$. That means that for a fixed s if we pick s normal vectors at random from the unit ball, then the expected number of pairs of half-spaces that are ϵ-close according to D is at most $\frac{s^2}{\sqrt{d}} \left(\frac{2\epsilon}{c}\right)^{d-1}$. Removing one element of each pair from S yields a set of $s - \frac{s^2}{\sqrt{d}} \left(\frac{2\epsilon}{c}\right)^{d-1}$ halfspaces that are ϵ-separated. Setting $s = \frac{\sqrt{d}}{(2\epsilon/c)^{d-1}}$, leads the desired result.

To handle the non-isotropic case, suppose that Σ is the covariance matrix of D, so that $\Sigma^{-1/2}$ is the whitening transform. Let D' be the whitened version of D, i.e. the distribution obtained by first choosing x from D, and then producing $\Sigma^{-1/2}x$. We have $\mathrm{d}_D(v, w) = \mathrm{d}_{D'}(v\Sigma^{1/2}, w\Sigma^{1/2})$ (because $\mathrm{sign}(v \cdot x) \neq \mathrm{sign}(w \cdot x)$ iff $\mathrm{sign}((v\Sigma^{1/2}) \cdot (\Sigma^{-1/2}x)) \neq \mathrm{sign}((w\Sigma^{1/2}) \cdot (\Sigma^{-1/2}x))$). So we can use an ϵ-packing w.r.t. D' to construct an ϵ-packing of the same size w.r.t. D. ■

Now we are ready to prove Theorem 13.

Theorem 13. *For a small enough constant β we have: (1) for any β log-concave distribution D whose covariance matrix has full rank, the sample complexity of learning origin-centered linear separators under D in the passive learning model is $\Omega\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$; (2) the sample complexity of active learning of linear separators under β log-concave distributions is $\Omega\left(d \log\left(\frac{1}{\epsilon}\right)\right)$.*

Proof First, let us consider passive PAC learning. It is known (Long, 1995) that, for any distribution D, the sample complexity of passive PAC learning origin-centered linear separators w.r.t. D is at

least

$$\frac{d-1}{e}\left(\frac{M_D(\mathbb{C},2\epsilon)}{4}\right)^{1/(d-1)}.$$

Applying Lemma 20 gives an $\Omega(d/\epsilon)$ lower bound. It is known (Long, 1995) that, if for each ϵ, there is a pair of classifier v, w such that $\mathrm{d}_D(v,w)=\epsilon$, then the sample complexity of PAC learning is $\Omega((1/\epsilon)\log(1/\delta))$; this requirement is satisfied by D.

Now let us consider the sample complexity of active learning. As shown in (Kulkarni et al., 1993), in order to output a hypothesis of error at most ϵ with probabality at least $1-\delta$, where $\delta \leq 1/2$ and active learning algorithm that is allowed to make arbitrary yes-no queries must make $\Omega(\log M_D(\mathbb{C},\epsilon))$ queries. Using this together with Lemma 20 we get the desired result. ■

Appendix F. The inseparable case: Disagreement-based active learning

Theorem 14. *Let $\beta \geq 0$ be a sufficiently small constant. Assume that D is an isotropic β log-concave distribution in R^d. For any w^*, for any ϵ, $\mathrm{cap}_{w^*,D}(\epsilon)$ is $O(d^{1/2+\frac{\beta}{2\ln 2}}\log(1/\epsilon))$. Thus $\mathrm{dis}_{w^*,D}(\epsilon)=O(d^{1/2+\frac{\beta}{2\ln 2}}\log(1/\epsilon))$.*

Proof Roughly, we will show that almost all x classified by a large enough margin by w^* are not in $\mathrm{DIS}(B(w^*,r))$, because all hypotheses agree with w^* about how to classify such x, and therefore all pairs of hypotheses agree with each other. Consider w such that $d(w,w^*)\leq r$; by Theorem 11 we have $\theta(w,w^*)\leq cr$. Define $C=e^{\beta\lceil\log_2(d+1)\rceil}$ as in the proof of Theorem 11. For any x such that $||x||\leq\sqrt{dC}\log(1/r)$ we have

$$\begin{aligned}(w\cdot x - w^*\cdot x) &< ||w-w^*||\times||x|| \\ &\leq cr\sqrt{dC}\log(1/r).\end{aligned}$$

Thus, if x also satisfies $|w^*\cdot x|\geq cr\sqrt{dC}\log(1/r)$ we have $(w^*\cdot x)(w\cdot x)>0$. Since this is true for all w, any such x is not in $\mathrm{DIS}(B(h,r))$. By Theorem 11 we have, for a constant c_2, that

$$\mathbb{P}_{x\sim D}(|w^*\cdot x|\leq cr\sqrt{Cd}\log(1/r))\leq c_2 r\sqrt{Cd}\log(1/r).$$

Moreover, by Theorem 11 we also have

$$\mathbb{P}_{x\sim D}[||x||\geq cr\sqrt{Cd}\log(1/r)]\leq r.$$

These both imply $\mathrm{cap}_{w^*,D}(\epsilon)=O(C^{1/2}\sqrt{d}\log(1/\epsilon))$. ■

Appendix G. Massart and Tsybakov noise

In this section we analyze label complexity for active learning under the popular Massart and Tsybakov noise conditions, proving Theorem 15.

We consider a variant of the Tsybakov noise condition (Mammen and Tsybakov, 1999). We assume that the classifier h that minimizes $\mathbb{P}_{(x,y)\sim D_{XY}}(h(x)\neq y)$ is a linear classifier, and that, for

the weight vector w^* of that optimal classifier, there exist known parameters $\alpha, a > 0$ such that, for all w, we have

$$a(\mathrm{d}_D(w, w^*))^{1/(1-\alpha)} \leq \mathrm{err}(w) - \mathrm{err}(w^*). \tag{12}$$

By generalizing Theorem 4 so that it provides a stronger bound for larger margins, and combining the result with the other lemmas of this paper and techniques from (Balcan et al., 2007), we get the following.

Theorem 15. *Let $s = O(\log(1/\epsilon))$. Assume that the distribution D_{XY} satisfies the Tsybakov noise condition for constants $\alpha \in [0, 1)$ and $a \geq 0$, and that the marginal D on R^d is isotropic log-concave. (1) If $\alpha = 0$, we can find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$ using $O(\log(1/\epsilon))(d+\log(s/\delta))$ labeled examples in the active learning model, and $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ labeled examples in the passive learning model. (2) If $\alpha > 0$, we can find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$ using $O((1/\epsilon)^{2\alpha} \log^2(1/\epsilon))(d + \log(s/\delta))$ labeled examples in the active learning model.*

Note that the case where $\alpha = 0$ is more general than the well-known Massart noise condition (Massart and Nedelec, 2006). In this case, for active learning, Theorem 15 improves over the previously best known results (Hanneke and Yang, 2012) by a (disagreement coefficient) $\mathrm{dis}_{w^*,D}(\epsilon)$ factor. For passive learning, the bound on the total number of examples needed improves by $\log(\mathrm{cap}_{w^*,D}(\epsilon))$ factor the previously known best bound of (Giné and Koltchinskii, 2006). It is consistent with recent lower bounds of (Raginsky and Rakhlin, 2011) that include $\log(\mathrm{cap}_{w^*,D}(\epsilon))$ because those bounds are for a worst-case domain distribution, subject to a constraint on $\mathrm{cap}_{w^*,D}(\epsilon)$.

When $\alpha > 0$, the previously best result for active learning (Hanneke and Yang, 2012) is

$$O((1/\epsilon)^{2\alpha} \mathrm{dis}_{w^*,D}(\epsilon)(d \log(\mathrm{dis}_{w^*,D}(\epsilon)) + \log(1/\delta)).$$

Combining this with our new bound on $\mathrm{dis}_{w^*,D}(\epsilon)$ (Theorem 14) we get a bound of

$$O((1/\epsilon)^{2\alpha} d^{3/2} \log(1/\epsilon)(\log(d) + \log\log(1/\epsilon)) + \log(1/\delta))$$

for log-concave distributions. So our Theorem 15 saves roughly a factor of $\sqrt{d}$, at the expense of an extra $\log(1/\epsilon)$ factor.

We note that the results in this section can also be extended to nearly log-concave distributions by making use of our results in Section 6.1.

G.1. Proof of Theorem 15

We are now ready to discuss the proof of Theorem 15. As in (Balcan et al., 2007), we will use a different algorithm in the inseparable case (Algorithm 2).

G.1.1. MASSART NOISE ($\alpha = 0$)

We start by analyzing Algorithm 2 in the case that $\alpha = 0$; the resulting assumption is more general than the well-known Massart noise condition.

From the log-concavity assumption, the proof of Theorem 5, with slight modifications, proves that there exists c such that for all w we have

$$ca\theta(w, w^*) \leq \mathrm{err}(w) - \mathrm{err}(w^*). \tag{13}$$

Algorithm 2 Margin-based Active Learning (non-separable case)

Input: a sampling oracle for D, and a labeling oracle a sequence of sample sizes $m_k > 0$, $k \in Z^+$; a sequence of cut-off values $b_k > 0$, $k \in Z^+$ a sequence of hypothesis space radii $r_k > 0$, $k \in Z^+$; a sequence of precision values $\epsilon_k > 0$, $k \in Z^+$
Output: weight vector $\hat{w}_s$.

- Pick random $\hat{w}_0$: $\|\hat{w}_0\|_2 = 1$.
- Draw m_1 examples from D_X, label them and put into W.
  - **iterate** $k = 1, \ldots, s$
    - find $\hat{w}_k \in B(\hat{w}_{k-1}, r_k)$ ($\|\hat{w}_k\|_2 = 1$) to approximately minimize training error: $\sum_{(x,y)\in W} I(\hat{w}_k \cdot xy) \leq \min_{w \in B(\hat{w}_{k-1}, r_k)} \sum_{(x,y)\in W} I(w \cdot xy) + m_k \epsilon_k$.
    - clear the working set W
    - until m_{k+1} additional data points are labeled, draw sample x from D_X
      - if $|\hat{w}_k \cdot x| \geq b_k$, reject x
      - otherwise, ask for label of x, and put into W

    end iterate

We prove by induction on k that after $k \leq s$ iterations, we have

$$\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq ca2^{-k}$$

with probability $1 - \frac{\delta}{2}\sum_{i<k} \frac{1}{(1+s-i)^2}$. The case $k = 1$ follows from classic bounds (Vapnik, 1998).

Assume now the claim is true for $k-1$ ($k \geq 2$). Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By induction hypothesis, we know that with probability at least $1 - \frac{\delta}{2}\sum_{i<k-1} \frac{1}{(1+s-i)^2}$ $\hat{w}_{k-1}$ has excess errors at most $ca2^{-(k-1)}$, implying, using (13), that $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(k-1)}$. By assumption, $\theta(\hat{w}_{k-1}, \hat{w}_k) \leq 2^{-(k-1)}$.

From Theorem 4, recalling that a is a constant, we have both:

$$\mathbb{P}((\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) < 0, x \in S_2) \leq ca2^{-k}/4$$
$$\mathbb{P}((\hat{w}_{k-1} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq ca2^{-k}/4.$$

Taking the sum, we obtain:

$$\mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq ca2^{-k}/2. \tag{14}$$

Therefore:

$$\begin{aligned} \mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) &\leq (\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1))\mathbb{P}(S_1) \\ &+ \mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \\ &\leq (\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1))c_3 b_{k-1} \\ &+ ca2^{-k}/2. \end{aligned}$$

By standard Vapnik-Chervonenkis bounds, we can choose C s.t. with m_k samples, we obtain

$$\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1) \leq ca2^{-k}/(c_3 b_{k-1})$$

with probability $1-(\delta/2)/(1+s-i)^2$. Therefore $\mathrm{err}(\hat{w}_k)-\mathrm{err}(w^*) \leq ca2^{-k}$ with probability $1-\frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$, as desired.

The bound on the total number of examples, labeled and unlabeled, follows the same line of argument as Theorem 6, except with the constants of this analysis.

G.1.2. TSYBAKOV NOISE ($\alpha > 0$)

We now treat the more general Tsybakov noise.

For this analysis, we need a generalization of Theorem 4 that provides a stronger bound on the probably of large-margin errors, using a stronger assumption on the margin.

Theorem 21 *There is a positive constant c such that the following holds. Let u and v be two unit vectors in R^d, and assume that $\theta(u,v)=\eta<\pi/2$. Assume that D is isotropic log-concave in R^d. Then, for any $b \geq c\eta$, we have*

$$\mathbb{P}_{x\sim D}[\mathrm{sign}(u\cdot x)\neq \mathrm{sign}(v\cdot x) \text{ and } |v\cdot x|\geq b] \leq C_5\eta\exp(-C_6 b/\eta), \tag{15}$$

for absolute constants C_5 and C_6.

Proof Arguing as in the proof of Lemma 3, we may assume without loss of generality that $d=2$.

Next, we claim that each member x of E has $||x|| \geq b/\eta$. Assume without loss of generality that $v\cdot x$ is positive. (The other case is symmetric.) Then $u\cdot x<0$, so the angle of x with u is obtuse, i.e. $\theta(x,u)\geq\pi/2$. Since $\theta(u,v)=\eta$, this implies that

$$\theta(x,v)\geq \pi/2-\eta. \tag{16}$$

But $x\cdot v\geq b$, and v is unit length, so $||x||\cos\theta(x,v)\geq b$, which, using (16), implies $||x||\cos(\pi/2-\eta)\geq b$, which, since $\cos(\pi/2-\eta)\leq\eta$ for all $\eta\in[0,\pi/2]$, in turn implies $||x||\geq b/\eta$. This implies that, if $B(r)$ is a ball of radius r in R^2, that

$$\mathbb{P}[E]=\sum_{i=1}^{\infty}\mathbb{P}[E\cap(B((i+1)(b/\eta))-B(i(b/\eta)))]. \tag{17}$$

Let us bound one of the terms in RHS. Choose $i\geq 1$.

Let $f(x_1,x_2)$ be the density of D. We have

$$\begin{aligned}&\mathbb{P}[E\cap(B((i+1)(b/\eta))-B(i(b/\eta)))]\\ &=\int_{(x_1,x_2)\in B((i+1)(b/\eta))-B(i(b/\eta))} 1_E(x_1,x_2)f(x_1,x_2)\,dx_1dx_2.\end{aligned}$$

Let $R_i=B((i+1)(b/\eta))-B(i(b/\eta)$. Applying the density upper bound from Lemma 2 with $d=2$, there are constants C_1 and C_2 such that

$$\begin{aligned}&\mathbb{P}[E\cap(B((i+1)(b/\eta))-B(i(b/\eta)))]\\ &\leq\int_{(x_1,x_2)\in R_i}1_E(x_1,x_2)C_1\exp(-(b/\eta)C_2 i)dx_1dx_2\\ &=C_1\exp(-(b/\eta)C_2 i)\cdot\\ &\int_{(x_1,x_2)\in R_i}1_E(x_1,x_2)\,dx_1dx_2.\end{aligned}$$

If we include $B(i(b/\eta))$ in the integral again, we get

$$\begin{aligned}&\mathbb{P}[E \cap (B((i+1)(b/\eta)) - B(i(b/\eta)))]\\ &\leq C_1 \exp(-(b/\eta)C_2 i) \int_{(x_1,x_2)\in B((i+1)(b/\eta))} 1_E(x_1,x_2)\, dx_1 dx_2.\end{aligned}$$

Now, we exploit the fact that the integral above is a rescaling of a probability with respect to the uniform distribution. Let C_3 be the volume of the unit ball in R^2. Then, we have

$$\begin{aligned}&\mathbb{P}[E \cap (B((i+1)(b/\eta)) - B(i(b/\eta)))]\\ &\leq C_1 \exp(-(b/\eta)C_2 i) C_3 (i+1)^2 (b/\eta)^2 \eta/\pi\\ &= C_4 (b/\eta)^2 \eta (i+1)^2 \exp(-(b/\eta)C_2 i),\end{aligned}$$

for $C_4 = C_1 C_3/\pi$. Returning to (17), we get

$$\begin{aligned}\mathbb{P}[E] &= \sum_{i=1}^{\infty} C_4 (b/\eta)^2 \eta (i+1)^2 \exp(-(b/\eta)C_2 i)\\ &= C_4 (b/\eta)^2 \eta \sum_{i=1}^{\infty} (i+1)^2 \exp(-(b/\eta)C_2 i)\\ &= C_4 (b/\eta)^2 \times \frac{4e^{2(b/\eta)C_2} - 3e^{(b/\eta)C_2} + 1}{\left(e^{(b/\eta)C_2} - 1\right)^3} \times \eta.\end{aligned}$$

Now, if $b/\eta > 4/C_2$, we have

$$\begin{aligned}\mathbb{P}[E] &\leq C_4 (b/\eta)^2 \times \frac{5e^{2(b/\eta)C_2}}{\left(e^{(b/\eta)C_2}/2\right)^3} \times \eta\\ &\leq C_5 \eta \times (b/\eta)^2 \exp(-(b/\eta)C_2) \text{(where } C_5 = 40C_4)\\ &= C_5 \eta \times \exp(-(b/\eta)C_2 + 2\ln(b/\eta))\\ &\leq C_5 \eta \times \exp(-(b/\eta)C_2/2),\end{aligned}$$

completing the proof. ■

Now we are ready to prove Theorem 15 in the case that $\alpha > 0$.

Under the noise condition 12 and from the log-concavity assumption, we obtain that there exists c such that for all w we have:

$$ac^{1/(1-\alpha)}\theta(w,w^*)^{1/(1-\alpha)} \leq \mathrm{err}(w) - \mathrm{err}(w^*).$$

Let us denote by $\tilde{c} = ac^{1/(1-\alpha)}$. For all w, we have:

$$\tilde{c}\theta(w,w^*)^{1/(1-\alpha)} \leq \mathrm{err}(w) - \mathrm{err}(w^*). \tag{18}$$

We prove by induction on k that after $k \leq s$ iterations, we have

$$\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq \tilde{c}2^{-k}$$

with probability $1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$. The case $k = 1$ follows from classic bounds.

Assume now the claim is true for $k-1$ ($k \geq 2$). Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By the induction hypothesis, we know that with probability at least $1 - \delta\sum_{i<k-1}\frac{1}{(1+s-i)^2}$, $\hat{w}_{k-1}$ has excess errors at most $\tilde{c}2^{-(k-1)(1-\alpha)}$, implying

$$\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(k-1)(1-\alpha)}.$$

By assumption, $\theta(\hat{w}_{k-1}, \hat{w}_k) \leq 2^{-(k-1)(1-\alpha)}$.

Applying Theorem 21, we have both:

$$\mathbb{P}((\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) < 0, x \in S_2) \leq \tilde{c}2^{-k}/4$$
$$\mathbb{P}((\hat{w}_{k-1} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \tilde{c}2^{-k}/4$$

Taking the sum, we obtain:

$$\mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \leq \tilde{c}2^{-k}/2. \tag{19}$$

Therefore:

$$\begin{aligned}\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) &\leq (\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1))\mathbb{P}(S_1) \\ &\quad +\mathbb{P}((\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2) \\ &\leq (\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1))b_k \\ &\quad +\tilde{c}2^{-k}/2.\end{aligned}$$

By standard bounds, we can choose C_1, C_2 and C_3 s.t. with m_k samples, we obtain $\mathrm{err}(\hat{w}_k|S_1) - \mathrm{err}(w^*|S_1) \leq \epsilon_k \leq \frac{\tilde{c}2^{-k}}{2b_k}$ with probability $1 - (\delta/2)/(1+s-i)^2$. Therefore $\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq \tilde{c}2^{-k}$ with probability $1 - \frac{\delta}{2}\sum_{i<k}\frac{1}{(1+s-i)^2}$, as desired, completing the proof of Theorem 15.