An Inequality for Nearly Log-concave Distributions with Applications to Learning

Constantine Caramanis, *Member, IEEE,* Shie Mannor, *Member, IEEE*

***Abstract*— We prove that given a nearly log-concave distribution, in any partition of the space to two well separated sets, the measure of the points that do not belong to these sets is large. We apply this isoperimetric inequality to derive lower bounds on the generalization error in learning. We further consider regression problems and show that if the inputs and outputs are sampled from a nearly log-concave distribution, the measure of points for which the prediction is wrong by more than ϵ_0 and less than ϵ_1 is (roughly) linear in $\epsilon_1 - \epsilon_0$, as long as ϵ_0 is not too small, and ϵ_1 not too large. We also show that when the data are sampled from a nearly log-concave distribution, the margin cannot be large in a strong probabilistic sense.**

***Index Terms*— classification, generalization error, margin, statistical learning theory**

I. Introduction

Large margin classifiers (e.g., [1], [2] to name but a few recent books) have become an almost ubiquitous approach in supervised machine learning. The plethora of algorithms that maximize the margin, and their impressive success (e.g., [3] and references therein) may lead one to believe that obtaining a large margin is synonymous with successful generalization and classification. In this paper we directly consider the question of how much weight the margin must carry. We show that essentially if the margin between two classes is large, then the weight of the "no-man's land" between the two classes must be large as well. Our probabilistic assumption is that the data are sampled from a nearly log-concave distribution. Under this assumption, we prove that for any partition of the space into two sets such that the distance between those two sets is t, the measure of the "no man's land" outside the two sets is lower bounded by t times the minimum of the measure of the two sets times a dimension-free constant. The direct implication of this result is that a large margin is unlikely when sampling data from such a distribution.

Our modelling assumption is that the underlying distribution has a β-log-concave density. While this assumption may appear restrictive, we note that many "reasonable" functions belong to this family. We discuss this assumption in Section II, and point out some interesting properties of β-log-concave functions.

In Section III we prove an inequality stating that the measure (under a β-log-concave density) of the "no-man's land" is large if the sets are well separated. This result relies essentially on the Prékopa-Leindler inequality which is a generalization of the Brunn-Minkowski inequality (we refer the reader to the excellent survey [4]). We note that the isoperimetric inequality we prove (Theorem 2) was stated in [5] for volumes, and in [6] for continuous β-log-concave distributions, in the context of efficient sampling from convex bodies. However, the proof sketched in [6] relies in an essential way on having a continuous density ([7]). We provide a complete proof of the more general result using the Ham-Sandwich Theorem (as in [5], but using a different method) and a different reduction argument. We further point out a few natural extensions.

In Section IV we specialize the isoperimetric inequality to provide lower bounds for the generalization error in classification under the assumption that the classifier will be tested using a β-log-concave distribution, which did not necessarily generate the data. While this assumption is not in line with the standard PAC (Probably Approximately Correct) learning formulation (see, e.g., [8]), it is applicable to the setup where data are sampled from one distribution and performance is judged by another. Suppose, for instance, that the generating distribution evolves over time, while the true classifier remains fixed. We may have access to a training set generated by a distribution quite different from the one we use to test our classifier. Another important motivation is the case where the data were indeed generated by the true distribution, but a portion of the data were erased, or lost.

In the absence of further information about the generating distribution or its evolution, or the data erasure, it becomes natural to ask "how bad" the training data may be. We show that if there is a large (in a geometric sense) family of classifiers that agree with the training points, then for any choice of classifier there exists another classifier compared to which the generalization error is relatively large.

In Section V we investigate regression problems. We consider several regression models and lower bound the measure of a tube around the prediction with inner radius ϵ_0 and outer radius ϵ_1. The measure of this tube represents the probability of a prediction error between ϵ_0 and ϵ_1 (equivalently, this is the weight of samples that become erroneous when we change the sensitivity parameter from ϵ_1 to ϵ_0 when using ϵ-sensitive error). Our results imply that the margins of the tube carry a significant portion of the measure. We start from a simple additive model where the noise is independent of the value of the independent variable. We show that the weight of the tube around the true regressor is bounded from below by

A preliminary version appeared in the 17th Annual Conference on Learning Theory (COLT) 2004, Lecture Notes in Computer Science 3120, pages 534-548.

C. Caramanis is with the Department of Electrical and Computer Engineering at The University of Texas at Austin, Austin, TX 78712 USA (email: cmcaram@ece.utexas.edu). He was partially supported by the National Science Foundation under Grant ANI-0335256.

S. Mannor is with the Department of Electrical and Computer Engineering, McGill University, Montréal, Québec, Canada (email: shie@ece.mcgill.ca). He was partially supported by NSERC and The Canada Research Chairs Program.

$\epsilon_1 - \epsilon_0$ times a constant (as long as ϵ_0 is not too small, and ϵ_1 not too large). We then consider a setup where the noise and the independent variable are drawn from a distribution which is jointly β-log-concave and show that the result extends to this setup. This setup is particularly interesting because it applies to linear prediction when the measure is generated by some other (unknown) linear function. We then extend the result to a conditional β-log-concave distribution and show that similar results still hold even if the independent variable is drawn first from a β-log-concave distribution, and then the dependent variable is drawn from another β-log-concave distribution (with, perhaps, a different β).

In Section VI we consider the standard statistical machine learning setup, and show that for any classifier the probability of a large margin (with respect to that specific classifier) decreases exponentially fast to 0 with the number of samples, if the data are sampled from a β-log-concave distribution. It is important to note that the β-log-concave assumption applies to the input space. If we use classification methods such as kernel methods that use Mercer kernels ([9], [10]), the margin is typically measured in the feature space. Since the induced distribution in the feature space is not necessarily β-log-concave our results do not directly hold. If, however, the kernel map is Lipschitz continuous with constant L, then we can relate the "functional" margin in the feature space to the "geometric" margin in the input space, and our results carry over directly.

Some recent results such as [11], [12] argue that the success of large margin classifiers is remarkable since most classes cannot have a useful embedding in some Hilbert space. Our results provide a different angle, as we show that having a large margin is unlikely to start with. Moreover, if there happens to be a large margin, it may well result in a large error (which is proportional to the margin). A notable feature of our bounds is that they are dimension-free and are therefore immune to the curse of dimensionality (this is essentially due to the β-log-concave assumption). We note the different flavor of our results from the "classical" lower bounds (e.g., [13], [14]) that are mostly concerned with the PAC setup and where the sample complexity is the main object of interest. We do not address the sample complexity directly in this work.

II. Nearly Log-Concave Functions

We assume throughout the paper that generalization error is measured using a nearly log-concave distribution. In this section we define such distributions and highlight some of their properties.

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

$$f(\lambda x_1 + (1-\lambda)x_2) \geq e^{-\beta} f(x_1)^{\lambda} f(x_2)^{1-\lambda}. \tag{II.1}$$

A function f is *log-concave* if it is 0-log-concave.

A distribution is called β-log-concave if it is defined by a β-log-concave density function. Henceforth, we refer to both β-log-concave distributions and their associated β-log-concave densities.

The class of log-concave distributions itself is rather rich. For example, it includes the Gaussian, Uniform, Logistic, and Exponential distributions. We refer the reader to [15] for an extensive list of such distributions, sufficient conditions for a distribution to be log-concave, and ways to "produce" log-concave distributions from other log-concave distributions. The class of β-log-concave distributions is considerably richer since we allow a factor of $e^{-\beta}$ in Ineq. (II.1). For example, while log-concavity implies a distribution must have a continuous density, this is not the case for β-log-concave distributions. They can even have densities with arbitrarily many discontinuities, and without well-defined derivative. Nevertheless, we see that while they are not regular in this respect, they have enough structure that much can be said about them. We now provide some results that are useful in the sequel. We start from the following observation (see, e.g., [6]).

Lemma 1: The support of a β-log-concave density is a convex set. Also, β-log-concave densities are bounded on bounded sets.

Densities that are β-log-concave are not necessarily unimodal, but possess a unimodal quality, in the sense of Lemma 2 below. This simple lemma captures the properties of β-log-concavity that are central to our main results and subsequent applications. It implies that if we have a β-log-concave distribution on an interval, there cannot be any big "holes" or "valleys" in the mass distribution. Thus if we divide the interval into three intervals, if the middle interval is large, it must also carry a lot of the weight. In higher dimensions, essentially this says that if we consider two subsets, then if the distance between the two sets is large, the mass of the "no-man's land" will also be large. This is essentially the content of Theorem 2 below.

Lemma 2: Suppose that $f(x) : [u_1, u_2] \to \mathbb{R}$ is β-log-concave on an interval $[u_1, u_2]$. Let $u_1 < x_1 < x_2 < u_2$. Then for any $x \in [x_1, x_2]$, at least one of the following holds:

$$f(x) \geq f(y) \cdot e^{-\beta}, \quad \text{for all } y \in [u_1, x_1],$$

or

$$f(x) \geq f(y) \cdot e^{-\beta}, \quad \text{for all } y \in [x_2, u_2].$$

Proof. Consider any $x \in [x_1, x_2]$, and suppose, in order to obtain a contradiction, that there exists $y_1 \in [u_1, x_1]$ and $y_2 \in [x_2, u_2]$, such that $f(x) < f(y_i)e^{-\beta}$, for $i = 1, 2$. Then, there exists $\lambda \in [0,1]$ such that $x = \lambda y_1 + (1-\lambda)y_2$, and thus by β-log-concavity of f, we have:

$$\begin{aligned} f(x) &= f(\lambda y_1 + (1-\lambda)y_2) \\ &\geq e^{-\beta} f(y_1)^{\lambda} f(y_2)^{1-\lambda} \\ &> e^{-\beta} (f(x)e^{\beta})^{\lambda} (f(x)e^{\beta})^{1-\lambda} = f(x), \end{aligned}$$

a contradiction. □

The following inequality has many uses in geometry, statistics, and analysis (see [16] for a proof, and [4] for more context, uses, and references). Note that it is stated with respect to a specific $\lambda \in (0,1)$ and not to all λ.

Theorem 1 (Prékopa-Leindler Inequality): Let $0 < \lambda < 1$, and h, g_1, g_2 be nonnegative integrable functions on $\mathbb{R}^n$, such that $h((1-\lambda)x + \lambda y) \geq g_1(x)^{1-\lambda} g_2(y)^{\lambda}$, for every $x, y \in \mathbb{R}^n$.

Then

$$\int_{\mathbb{R}^n} h(x)\,dx \geq \left(\int_{\mathbb{R}^n} g_1(x)\,dx\right)^{1-\lambda}\left(\int_{\mathbb{R}^n} g_2(x)\,dx\right)^{\lambda}.$$

The following lemma plays a key part in the reduction technique we use below. It essentially says that the projection of a β-log-concave distribution is still β-log-concave. Recall that the orthogonal projection of a set $K \subseteq \mathbb{R}^{n+m}$ onto $\mathbb{R}^n$ is defined as $K|_{\mathbb{R}^n} \stackrel{\triangle}{=} \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m \text{ s.t. } (x,y) \in K\}$.

Lemma 3: Let $f(x,y)$ be a β-log-concave density on a convex set $K \subseteq \mathbb{R}^{n+m}$. For every x in $K|_{\mathbb{R}^n}$, consider the section $K(x) \stackrel{\triangle}{=} \{(x,y) \in \mathbb{R}^{n+m} : (x,y) \in K\}$. Then the density $F(x) \stackrel{\triangle}{=} \int_{K(x)} f(x,y)\,dy$ is β-log-concave on $K|_{\mathbb{R}^n}$.

PROOF. This is a consequence of the Prékopa-Leindler inequality (as in [4] for log-concave functions). Fix $x_1, x_2 \in K|_{\mathbb{R}^n}$. Define the functions $g_i(y) = f(x_i, y)$ for $i = 1,2$. Therefore $g_i(\cdot)$ is defined on $K(x_i)$, $i = 1,2$. For $\lambda \in (0,1)$, let $x = \lambda x_1 + (1-\lambda)x_2$, and define the function $h(y) = f(x,y)$ defined on $K(x)$. By the convexity of K, $\lambda K(x_1) + (1-\lambda)K(x_2) \subseteq K(x)$. In particular, for any $y_i \in K(x_i)$, $i = 1,2$, the point $(x,y) = \lambda(x_1,y_1) + (1-\lambda)(x_2,y_2) \in K(x)$. By the β-log-concavity of $f(x,y)$,

$$\begin{array}{rcl} f(x,y) & = & f(\lambda(x_1,y_1) + (1-\lambda)(x_2,y_2)) \\ & \geq & f(x_1,y_1)^{\lambda} \cdot f(x_2,y_2)^{1-\lambda} e^{-\beta}, \end{array}$$

and therefore this implies $h(\lambda y_1 + (1-\lambda)y_2) \geq g_1(y_1)^{\lambda} \cdot g_2(y_2)^{1-\lambda} e^{-\beta}$.

Denoting the indicator function by $\chi(\cdot)$, we have

$$\begin{array}{l} h(\lambda y_1 + (1-\lambda)y_2)\chi_{K(x)}(y) \geq (g_1(y_1) \cdot \chi_{K(x_1)}(y_1))^{\lambda} \cdot \\ \qquad\qquad\qquad\qquad (g_2(y_2) \cdot \chi_{K(x_2)}(y_2))^{1-\lambda} e^{-\beta}. \end{array}$$

But then the functions $H(y) = e^{\beta} h(y) \cdot \chi_{K(x)}(y)$, $G_1(y) = g_1(y) \cdot \chi_{K(x_1)}(y)$, and $G_2(y) = g_2(y) \cdot \chi_{K(x_2)}(y)$, satisfy the hypotheses of the Prékopa-Leindler Theorem, and thus we can write $F(\lambda x_1 + (1-\lambda)x_2) = F(x)$ as

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}^m} f(x,y) \cdot \chi_{K(x)}(y)\,dy = \int_{\mathbb{R}^m} h(y) \cdot \chi_{K(x)}(y)\,dy \\ \displaystyle \geq e^{-\beta}\left(\int_{\mathbb{R}^m} g_1(y)\chi_{K(x_1)}(y)\,dy\right)^{\lambda}\left(\int_{\mathbb{R}^m} g_2(y)\chi_{K(x_2)}(y)\,dy\right)^{1-\lambda} \\ \displaystyle = \quad e^{-\beta}\left(\int_{\mathbb{R}^m} f(x_1,y) \cdot \chi_{K(x_1)}(y)\,dy\right)^{\lambda} \cdot \\ \displaystyle \qquad\qquad \left(\int_{\mathbb{R}^m} f(x_2,y) \cdot \chi_{K(x_2)}(y)\,dy\right)^{1-\lambda} \\ = \quad F(x_1)^{\lambda} \cdot F(x_2)^{1-\lambda} e^{-\beta}. \end{array}$$

Since this holds for all $\lambda \in (0,1)$, $F(x)$ is β-log-concave. □

There are quite a few interesting properties of β-log-concave distributions. For example, the convolution of a β_1-log-concave and a β_2-log-concave density is $(\beta_1 + \beta_2)$-log-concave; Gaussian mixtures are β-log-concave; and mixtures of distributions with bounded Radon-Nikodym derivative are also β-log-concave. Additional discussion of these and other properties of β-log-concave distributions is beyond the scope of this paper.

III. ISOPERIMETRIC INEQUALITIES

In this section we prove our main result concerning β-log-concave distributions. We show that if two sets are well separated, then the "no man's land" between them has large measure relative to the measure of the two sets. Results of this nature exist in the literature for log-concave distributions. Recent results along these lines (e.g., [17]; for a survey see [18]) use a powerful localization lemma proved in [19] that requires a continuity assumption (for related results using generalized localization theorems, we refer the reader to [20] and [21]). Here, we provide a different proof that requires no such regularity.

We first prove the result for bounded sets and then provide two immediate corollaries. Let $d(x,y)$ denote the Euclidean distance in $\mathbb{R}^n$. We define the distance between two sets K_1 and K_2 as $d(K_1,K_2) \stackrel{\triangle}{=} \inf_{x \in K_1, y \in K_2} d(x,y)$ and the diameter of a set K as $\operatorname{diam}(K) \stackrel{\triangle}{=} \sup_{x,y \in K} d(x,y)$. Given a density f we say that $\mu(K) = \int_K f(x)\,dx$ is the induced measure. A decomposition of a closed set $K \subseteq \mathbb{R}^n$ to a collection of closed sets $K_1, K_2, \ldots, K_\ell$ satisfies that: $\bigcup_{i=1}^{\ell} K_i = K$ and $\nu(K_i \cap K_j) = 0$ for all $i \neq j$ where ν is the Lebesgue measure on $\mathbb{R}^n$.

Theorem 2: Let K be a closed and bounded convex set with non-zero diameter in $\mathbb{R}^n$ with a decomposition $K = K_1 \cup B \cup K_2$. For any β-log-concave density $f(x)$, the induced measure μ satisfies

$$\mu(B) \geq e^{-\beta}\frac{d(K_1,K_2)}{\operatorname{diam}(K)}\min\{\mu(K_1),\mu(K_2)\}.$$

We remark that this bound is dimension-free. The ratio $d(K_1,K_2)/\operatorname{diam}(K)$ is necessary, as essentially it adjusts for any scaling of the problem. We further note that the minimum $\min\{\mu(K_1),\mu(K_2)\}$ might be quite small; however, this appears to be unavoidable (e.g., consider the tail of a Gaussian, which is log-concave).

The proof proceeds by induction on the dimension n. The steps are as follows.

1. We prove the base case, $n = 1$, in Lemma 4. Here, the set K is an interval. The key tool we use is Lemma 2.
2. The inductive step uses a projection argument to reduce to $n-1$ dimensions. Lemma 5 reduces to the case of an "ϵ-flat" set, i.e., a set contained in an ellipse whose smallest axis is smaller than some $\epsilon > 0$.
3. Once we have reduced to the ϵ-flat case, we complete the induction by projecting to $n-1$ dimensions where the result holds by inductive hypothesis. By properly performing the projection, we show that if the result holds for the projection, it holds for the original set.

We abbreviate $t = d(K_1,K_2)$. The theorem trivially holds if $t = 0$, so we can assume that $t > 0$. From Lemma 1 above, we know that the support of $f(x)$ is convex. Thus, we can assume without loss of generality that since K is compact, $f(x)$ is strictly positive on the interior of K.

Step 1:

Lemma 4: Theorem 2 holds for $n = 1$.

PROOF. If $n = 1$, then K is some interval, $K = [u_1, u_2]$, with $\operatorname{diam}(K) = |u_2 - u_1|$. Since $t = d(K_1,K_2) > 0$, no point of K_1 is within a distance (strictly less than) t from any point of K_2. Furthermore, there must be at least one interval $(b_1, b_2) \subseteq B$ such that $|b_2 - b_1| \geq t$, and such that $(b_1, b_2) \cap (K_1 \cup K_2) = \emptyset$. Fix some $\epsilon > 0$, with $\epsilon < t/2$. Define the ϵ-expansion sets $\hat{K}_1 \triangleq \{x \in K : d(x, K_1) \leq \epsilon\}$, and $\hat{K}_2 \triangleq \{x \in K : d(x, K_2) \leq \epsilon\}$. Define $\hat{B}$ to be the closure of the complement in K of $\hat{K}_1 \cup \hat{K}_2$. Each set is a union of a finite number of closed intervals, and thus we have the decomposition $[u_1, u_2] = \bigcup_{i=1}^{m} [r_{i-1}, r_i]$, where each interval $[r_{i-1}, r_i]$ is either a $\hat{K}_1$-interval, a $\hat{K}_2$-interval, or a $\hat{B}$-interval. We modify the sets so that if the $\hat{B}$-interval $[r_{i-1}, r_i]$ is sandwiched by two $\hat{K}_i$-intervals ($i = 1, 2$) then we add that interval to $\hat{K}_i$. If the $\hat{B}$-interval is either the first interval $[r_0, r_1]$, or the last interval, $[r_{m-1}, r_m]$, then we add it to whichever set $\hat{K}_i$ is to its right, or left, respectively.

The three resulting sets $\hat{K}_1, \hat{K}_2$, and $\hat{B}$ are closed, intersect at most at a finite number of points, and thus are a decomposition of K. Each set is a union of a finite number of closed intervals. Furthermore, $\hat{t} = d(\hat{K}_1, \hat{K}_2) \geq t - 2\epsilon$, and $\hat{K}_1 \supseteq K_1$, $\hat{K}_2 \supseteq K_2$, and $\hat{B} \subseteq B$. By our modifications above, each $\hat{B}$-interval must have length at least $\hat{t}$.

Consider any $\hat{B}$-interval $[r_{i-1}, r_i]$. Let x^* be a maximizer[1] of $f(x)$ on $[u_1, u_2]$, and $x_{\min}$ a minimizer of $f(x)$ on $[r_{i-1}, r_i]$. Suppose that $x^* \geq x_{\min}$. Then by Lemma 2, for any $y \leq r_{i-1}$, we must have $f(x_{\min}) \geq f(y)e^{-\beta}$. Therefore,

$$\begin{aligned} e^{-\beta}\mu([u_1, r_{i-1}]) &= e^{-\beta} \int_{u_1}^{r_{i-1}} f(x)\,dx \\ &\leq (r_{i-1} - u_1) f(x_{\min}) \\ &\leq \operatorname{diam}(K) \cdot f(x_{\min}) \\ &\leq \frac{\operatorname{diam}(K)}{(r_i - r_{i-1})} \int_{r_{i-1}}^{r_i} f(x)\,dx \\ &\leq \frac{\operatorname{diam}(K)}{\hat{t}} \mu([r_{i-1}, r_i]). \end{aligned}$$

If instead we have $x^* \leq x_{\min}$, then in a similar manner we obtain the inequality

$$e^{-\beta}\mu([r_i, u_2]) \leq \frac{\operatorname{diam}(K)}{\hat{t}} \mu([r_{i-1}, r_i]).$$

Therefore, in general, for any $\hat{B}$-interval (r_{i-1}, r_i),

$$\mu([r_{i-1}, r_i]) \geq e^{-\beta} \frac{\hat{t}}{\operatorname{diam}(K)} \min\{\mu([u_1, r_{i-1}]), \mu([r_i, u_2])\}.$$

Suppose, without loss of generality, that $[r_0, r_1]$ is a K_1-interval. Consider the first $\hat{B}$-interval $[r_1, r_2]$. If $\mu([r_1, r_2]) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu([r_2, u_2])$, then $\mu(\hat{B}) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu(\hat{K}_2)$ and we are done. So let us assume that $\mu([r_1, r_2]) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu([u_1, r_1])$. Similarly, for the last $\hat{B}$-interval (r_{m-2}, r_{m-1}), we can assume that $\mu([r_{m-2}, r_{m-1}]) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu([r_{m-1}, u_2])$ otherwise the result immediately follows. This implies that there must be two consecutive $\hat{B}$-intervals, say (r_{j-1}, r_j) and (r_{j+1}, r_{j+2}) such that

$$\mu([r_{j-1}, r_j]) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu([u_1, r_{j-1}]),$$

and

$$\mu([r_{j+1}, r_{j+2}]) \geq e^{-\beta}(\hat{t}/\operatorname{diam}(K))\mu([r_{j+2}, u_2]).$$

Since $[u_1, r_{j-1}] \cup [r_{j+2}, u_2]$ contains either all of $\hat{K}_1$ or $\hat{K}_2$, combining these two inequalities, and using the fact that $\hat{K}_i \supseteq K_i$, and $\hat{B} \subseteq B$, we obtain

$$\begin{aligned} \mu(B) \geq \mu(\hat{B}) &\geq \mu([r_{j-1}, r_j] \cup [r_{j+1}, r_{j+2}]) \\ &\geq e^{-\beta} \frac{\hat{t}}{\operatorname{diam}(K)} (\mu([u_1, r_{j-1}]) + \mu([r_{j+2}, u_2])) \\ &\geq e^{-\beta} \frac{\hat{t}}{\operatorname{diam}(K)} \min\{\mu(\hat{K}_1), \mu(\hat{K}_2)\} \\ &\geq e^{-\beta} \frac{t - 2\epsilon}{\operatorname{diam}(K)} \min\{\mu(K_1), \mu(K_2)\}. \end{aligned}$$

Since this holds for every $\epsilon > 0$, the result follows. □

Step 2: We now prove the n-dimensional case. The first part of our inductive step is to show that it is enough to consider an "ϵ-flat" set K. To make this precise, we use the *Löwner-John Ellipsoid* of a set K. This is the minimum volume ellipsoid E containing K (see, e.g., [22]). This ellipsoid is unique. The key property we use is that if we shrink E from its center by a factor of n, then it is contained in K. We define an ϵ-flat set to be such that the smallest axis of its Löwner-John Ellipsoid has length no more than ϵ.

Lemma 5: Suppose the theorem *fails by* δ on K, for some $\delta > 0$, i.e.

$$(1 + \delta)\mu(B) \leq e^{-\beta} \frac{t}{\operatorname{diam}(K)} \min\{\mu(K_1), \mu(K_2)\}. \quad \text{(III.2)}$$

Then for any $\epsilon > 0$, there exists some ϵ-flat set $\tilde{K} \subseteq K$ with decomposition $\tilde{K} = \tilde{K}_1 \cup \tilde{B} \cup \tilde{K}_2$, such that $\tilde{K}_i \subseteq K_i$, $\tilde{B} \subseteq B$, $d(\tilde{K}_1, \tilde{K}_2) \geq t$, and $\operatorname{diam}(\tilde{K}) \leq d$, and such that the theorem fails by δ, i.e., Ineq. (III.2) holds for $\tilde{K}, \tilde{K}_1, \tilde{K}_2, \tilde{B}$.

PROOF. Let K, K_1, K_2, B and δ be as in the statement above. Pick some $\epsilon > 0$ much smaller than t. Suppose that all axes of the Löwner-John ellipsoid of K are greater than ϵ. A powerful consequence of the Borsuk-Ulam Theorem, the so-called Ham-Sandwich Theorem (see, e.g., [23]) says that in $\mathbb{R}^n$, given n Borel measures $\mu_k, k = 1, \ldots, n$, such that the weight of any hyperplane under each measure is zero, there exists a hyperplane H that bisects each measure, i.e., $\mu_k(H^+) = \mu_k(H^-) = \frac{1}{2}\mu_k(\mathbb{R}^n)$ for each k, where H^+, H^- denote the two half-spaces defined by H. Now, since we have $n \geq 2$, the Ham-Sandwich Theorem guarantees that there exists some hyperplane H that bisects (in terms of the measure μ) both K_1 and K_2. Let K' and K'' be the two parts of K defined by H (K and B are not necessarily bisected), and similarly define K_1', K_1'', K_2', K_2'', and B', B''. The minimum distance cannot decrease, i.e., $d(K_1', K_2') \geq t$, and $d(K_1'', K_2'') \geq t$, and the diameter of K cannot be smaller than either the diameter of K' or K''. Consequently, if the

[1] As in Lemma 2, f may not be continuous, so we may only be able to find a point x^* ($x_{\min}$) that is infinitesimally close to the supremum (infimum) of f. For convenience of exposition, we assume f is continuous. This assumption can be removed with an argument parallel to that given in Lemma 2.

theorem holds, or fails by less than δ, for both K' and K'', then

$$\begin{aligned}(1+\delta)\mu(B) &= (1+\delta)\mu(B') + (1+\delta)\mu(B'')\\ &\geq e^{-\beta}\frac{t}{\mathrm{diam}(K)}\left(\min\{\mu(K_1'),\mu(K_2')\} + \right.\\ &\qquad\qquad \left.\min\{\mu(K_1''),\mu(K_2'')\}\right)\\ &= e^{-\beta}\frac{t}{\mathrm{diam}(K)}\min\{\mu(K_1),\mu(K_2)\}.\end{aligned}$$

Therefore the theorem must fail by δ for either K' or K''. We note that this is the *same* δ as above. Call the set for which the theorem does not hold $K^{(1)}$, and similarly define $K_1^{(1)}, K_2^{(1)}$ and $B^{(1)}$. We continue bisecting $K^{(j)}$ in this way, always focusing on the side for which the theorem fails by δ, thus obtaining a sequence of nested sets

$$K \supseteq K^{(1)} \supseteq \cdots \supseteq K^{(j)} \supseteq \cdots.$$

We claim that eventually the smallest axis of the Löwner-John ellipsoid will be smaller than ϵ. If this is not the case, then the set K always contains a ball of radius ϵ/n. This follows from the properties of the Löwner-John ellipsoid. Therefore, letting $B_{\epsilon/n}(x_0)$ denote the ball of radius ϵ/n centered at x_0, we have

$$\begin{aligned}\mu(K^{(j)}) &= \int_{K^{(j)}} f(x)\,dx\\ &\geq \inf_{B_{\epsilon/n}(x_0)\subseteq K}\left(\int_{B_{\epsilon/n}(x_0)} f(x)\,dx\right) \geq \eta > 0,\end{aligned}$$

for some $\eta > 0$, independent of j. We know that $\eta > 0$ by our initial assumption that $f(x)$ is non-zero on K.

However, by our choice of hyperplanes, the sets $K_1^{(j)}, K_2^{(j)}$ are bisected with respect to the measure μ. Thus $\mu(K_1^{(j)}) = 2^{-j}\mu(K_1)$, and $\mu(K_2^{(j)}) = 2^{-j}\mu(K_2)$, and the measure of each set $K_1^{(j)}, K_2^{(j)}$ becomes arbitrarily small as j increases. Since the measure of $K^{(j)}$ does not also become arbitrarily small, the measure of $B^{(j)}$ must also be bounded away from zero. In particular,

$$\mu(B^{(j)}) \geq \eta - 2^{-j}(\mu(K_1)+\mu(K_2)),$$

and thus for

$$j \geq \log_2(2(\mu(K_1)+\mu(K_2))/\eta),$$

we have

$$\mu(B^{(j)}) \geq \eta/2 \geq \min\{\mu(K_1^{(j)}),\mu(K_2^{(j)})\}.$$

This contradicts our assumption that the theorem fails on all elements of our nested chain of sets. The contradiction completes the proof of the lemma. □

Step (3): We now perform the projection, proving the inductive step. We put the steps together to complete the proof.

Proof of Theorem 2: The proof is by induction on the number of dimensions. By Lemma 4 above, the statement holds for $n = 1$. Assume that the result holds for n dimensions. Suppose we have $K \subseteq \mathbb{R}^{n+1}$, with the decomposition $K = K_1 \cup B \cup K_2$ satisfying the assumptions of the theorem. We show that for every $\delta > 0$:

$$(1+\delta)\mu(B) \geq e^{-\beta}\frac{t}{\mathrm{diam}(K)}\min\{\mu(K_1),\mu(K_2)\}.$$

Taking δ to zero yields our result. Let E be the Löwner-John ellipsoid of K. By Lemma 5 above, we can assume that the Löwner-John ellipsoid of K has at least one axis of length no more than ϵ. Figure 1 illustrates the bisecting process of Lemma 5, and also the essential reason why the bisection allows us to project to one fewer dimensions. We take ϵ smaller than $t/2$, and also such that $\sqrt{t^2-4\epsilon^2} > t/(1+\delta)$. Assume that the $(n+1)^{st}$ coordinate direction is parallel to the shortest axis of the ellipsoid, and the first n coordinate directions span the same plane as the other n axes of the ellipse (changing coordinates if necessary). Call the last coordinate y, so that we refer to points in $\mathbb{R}^{n+1}$ as (x,y), for $x \in \mathbb{R}^n$, and $y \in \mathbb{R}$. Let Π denote the plane spanned by the other n axes, and let $K_\Pi = \pi(K)$ denote the projection of K onto Π. Since $\epsilon < t/2$, no point in K_Π is the image of points in both K_1 and K_2, otherwise the two pre-images would be at most $2\epsilon < t$ apart. This allows us to define the sets

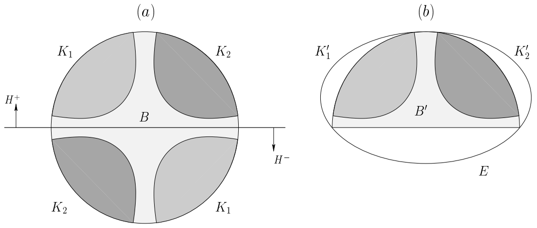


Fig. 1. The inductive step works by projecting K onto one less dimension. In (a) above, a projection on the horizontal axis would yield a distance of zero between the projected K_1 and K_2. Once we bisect to obtain (b), we see that a projection onto the horizontal axis would not affect the minimum distance between K_1 and K_2.

$$\begin{aligned}\hat{K}_1 &\triangleq \{(x,y)\in K : \pi(x,y)\in\pi(K_1)\},\\ \hat{K}_2 &\triangleq \{(x,y)\in K : \pi(x,y)\in\pi(K_2)\},\\ \hat{B} &\triangleq \{(x,y)\in K : \pi(x,y)\notin\pi(K_1)\cup\pi(K_2)\}.\end{aligned}$$

Note that $\mu(\hat{K}_i) \geq \mu(K_i)$, $i = 1,2$, and $\mu(\hat{B}) \leq \mu(B)$. Again we have a decomposition $K = \hat{K}_1 \cup \hat{B} \cup \hat{K}_2$. On K_Π, we also have a decomposition: $K_\Pi = \pi(\hat{K}_1)\cup\pi(\hat{B})\cup\pi(\hat{K}_2)$. Since we project with respect to the L^2 norm, by the Pythagorean Theorem, $d(\pi(\hat{K}_1),\pi(\hat{K}_2)) \geq \sqrt{t^2-4\epsilon^2}$. In addition, $\mathrm{diam}(K_\pi) \leq \mathrm{diam}(K)$.

For $x \in K_\Pi$, define the section $K(x) = \{(x,y)\in\mathbb{R}^{n+1} : (x,y)\in K\}$. We define a function on $K_\Pi \subseteq \mathbb{R}^n$: $F(x) \triangleq \int_{K(x)} f(x,y)\,dy$, where $f(x,y)$ is our β-log-concave density on $\mathbb{R}^{n+1}$. We have

$$\int_{\pi(\hat{K}_i)} F(x)\,dx = \int_{\hat{K}_i} f(x,y)\,dx\,dy = \mu(\hat{K}_i), \quad i = 1,2,$$

and similarly for $\hat{B}$. By Lemma 3, $F(x)$ is β-log-concave.

Therefore, by the inductive hypothesis, we have that

$$\begin{aligned}\mu(B) &\geq \mu(\hat{B}) = \int_{\hat{B}} f(x,y)\,dx\,dy = \int_{\pi(\hat{B})} F(x)\,dx \\ &\geq e^{-\beta}\frac{\sqrt{t^2-4\epsilon^2}}{\operatorname{diam}(K_\pi)} \min\left\{\int_{\pi(\hat{K}_1)} F(x)\,dx, \int_{\pi(\hat{K}_2)} F(x)\,dx\right\} \\ &= e^{-\beta}\frac{\sqrt{t^2-4\epsilon^2}}{\operatorname{diam}(K_\pi)} \min\left\{\int_{\hat{K}_1} f(x,y)\,dx\,dy, \int_{\hat{K}_2} f(x,y)\,dx\,dy\right\} \\ &> e^{-\beta}\frac{t/(1+\delta)}{\operatorname{diam}(K)} \min\{\mu(\hat{K}_1), \mu(\hat{K}_2)\},\end{aligned}$$

and thus

$$(1+\delta)\mu(B) \geq (t/\operatorname{diam}(K)) \min(\mu(K_1), \mu(K_2)).$$

Since this holds for every $\delta > 0$, the result follows. □

Corollaries 1, 2, and Proposition 1 below offer some flexibility for obtaining a tighter lower bound on $\mu(B)$.

Corollary 1: Let K be a closed and bounded convex set with a decomposition $K = K_1 \cup B \cup K_2$ as in Theorem 2 above. Let $f(x)$ be any density (not necessarily β-log-concave) that is bounded away from zero on K, say $f(x) > \eta$ for $x \in K$. Then the induced measure μ satisfies

$$\mu(B) \geq \eta \cdot \frac{d(K_1,K_2)}{\operatorname{diam}(K)} \min\{\nu(K_1), \nu(K_2)\}.$$

where ν denotes Lebesgue measure.

PROOF. Consider the uniform distribution on K. Since it is log-concave, Theorem 2 applies with $\beta = 0$. Since the Lebesgue measure ν is just a scaled uniform distribution, $\nu(B) \geq (d(K_1,K_2)/\operatorname{diam}(K)) \min\{\nu(K_1), \nu(K_2)\}$. The corollary follows since $\mu(B) \geq \eta\nu(B)$. □

The lower bound on $\mu(B)$ which we obtain from Theorem 2, depends inversely on the diameter of the set K, which we take to be bounded. This poses two potential problems. First, if the set K is unbounded, then the theorem cannot be applied, and the isoperimetric inequality, as stated, is meaningless. Second, even if K is bounded, the inequality may be rendered quite weak if the diameter is very large. Specifically, the problem arises if K has a very large diameter, while most of the mass of the distribution is contained in a small-diameter subset of K, with light tails putting very little mass on the rest of K. A Gaussian is a prime example of a β-log-concave (in fact 0-log-concave) distribution with this behavior.

The following two results address both issues by truncating K, and then applying Theorem 2 to the truncation. First we give a corollary that does not assume any further knowledge about the density $f(x)$. Then in Proposition 1 we give a corollary that replaces the diamater in Theorem 2 by the second moment of $f(x)$.

Corollary 2: Fix $\epsilon > 0$. Let K be a closed, convex, but not necessarily bounded set. Let $K = K_1 \cup B \cup K_2$ be a decomposition of K. Let f be a β-log-concave density with induced measure μ, such that there exists $d(\epsilon) > 0$ for which $(1-\epsilon)\mu(K_1) \leq \mu(K_1 \cap B_{d(\epsilon)})$, $(1-\epsilon)\mu(K_2) \leq \mu(K_2 \cap B_{d(\epsilon)})$, and $(1-\epsilon)\mu(B) \leq \mu(B \cap B_{d(\epsilon)})$, where $B_{d(\epsilon)}$ is a ball with radius $d(\epsilon)$ around the origin. Then

$$\mu(B) \geq e^{-\beta}(1-\epsilon)^2 \frac{d(K_1,K_2)}{d(\epsilon)} \min\{\mu(K_1), \mu(K_2)\}.$$

PROOF. We have that $\mu(K \cap B_{d(\epsilon)}) \geq (1-\epsilon)\mu(K)$. Let $P = \mu(K \cap B_{d(\epsilon)})$, and note that $P \geq 1-\epsilon$. Consider the measure $\hat{\mu}$ defined on $K \cap B_{d(\epsilon)}$ by the density $\hat{f}(x) = f(x)/P$. It follows that $\hat{f}$ is β-log-concave. We now apply Theorem 2 on $\hat{f}$ to obtain that:

$$\hat{\mu}(B \cap B_{d(\epsilon)}) \geq e^{-\beta}(t/d(\epsilon)) \min\{\hat{\mu}(K_1 \cap B_{d(\epsilon)}), \hat{\mu}(K_2 \cap B_{d(\epsilon)})\},$$

where $t \geq d(K_1,K_2)$. It follows that

$$\hat{\mu}(K_1 \cap B_{d(\epsilon)}) \geq (1-\epsilon)\mu(K_1),$$

and similarly for K_2, and also

$$\mu(B)/(1-\epsilon) \geq \mu(B)/P \geq \hat{\mu}(B \cap B_{d(\epsilon)}).$$

The result now follows by some algebra. □

If most of the mass of the distribution is contained in a small-diameter set, so that the trace of the covariance matrix is not too big, then it is possible to obtain a similar result, replacing the term $\operatorname{diam}(K)$ in the denominator by a term involving the covariance.

Proposition 1: Let K, K_1, K_2, B and f and μ be as above, and let $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ be the mean of the density $f(x)$, and σ^2 the trace of the covariance of $f(x)$ (which we assume to be finite):

$$\sigma^2 \triangleq \int_K \left(\sum_{i=1}^{n} (x_i - \bar{x}_i)^2\right) f(x)\,dx = \int_K ||x-\bar{x}||_2^2\, f(x)\,dx.$$

Then the induced measure μ satisfies

$$\mu(B) \geq e^{-\beta}\frac{d(K_1,K_2)}{4\sigma\sqrt{2}} \min\{\mu(K_1)^{3/2}, \mu(K_2)^{3/2}\}.$$

PROOF. Let us assume first that $\mu(K_1) \leq \mu(K_2)$. We require the following generalization of the Chebychev inequality to multiple dimensions. For such generalized inequalities, see, e.g., [24] or [25], and references therein. Here we use the inequality

$$\mu(K \setminus B_l(\bar{x})) = (\mu(\{x : ||x-\bar{x}||_2 \geq l\}) \leq \frac{\sigma^2}{l^2}, \qquad \text{(III.3)}$$

where $B_l(\bar{x})$ denotes the l-ball about the mean, $\bar{x}$. Setting the right hand side of Ineq. (III.3) equal to $\mu(K_1)/2$, we find

$$l = \sigma\sqrt{\frac{2}{\mu(K_1)}}.$$

Now let $\hat{K}_1, \hat{K}_2$ and $\hat{B}$ denote the truncations of K_1, K_2 and B, and let $\hat{\mu}$ denote the truncated and renormalized measure. Using $2l$ as the diameter of the truncated set, and observing that $d(\hat{K}_1, \hat{K}_2) \geq d(K_1, K_2)$, and then applying Theorem 2, we find

$$\hat{\mu}(\hat{B}) \geq e^{-\beta}\frac{d(K_1,K_2)}{2l} \min\{\hat{\mu}(\hat{K}_1), \hat{\mu}(\hat{K}_2)\}. \qquad \text{(III.4)}$$

Noting that $\mu(B_l(\bar{x})) \geq 1 - \mu(K_1)/2 \geq 3/4$, we have

$$\begin{aligned} \mu(B) &\geq \mu(B \cap B_l(\bar{x})) = \mu(B_l(\bar{x})) \frac{\mu(B \cap B_l(\bar{x}))}{\mu(B_l(\bar{x}))} \\ &\geq \frac{3}{4}\hat{\mu}(\hat{B}) \\ \hat{\mu}(\hat{K}_1) &\geq \frac{4}{3}\mu(K_1 \cap B_l(\bar{x})) \\ &\geq \frac{4}{3}(\mu(K_1) - \mu(K_1)/2) = \frac{2}{3}\mu(K_1) \\ \hat{\mu}(\hat{K}_2) &\geq \frac{2}{3}\mu(K_2). \end{aligned}$$

Together with inequality (III.4) we have

$$\begin{aligned} \mu(B) &\geq e^{-\beta}\frac{d(K_1, K_2)}{4l}\min\{\mu(K_1), \mu(K_2)\} \\ &= e^{-\beta}\frac{d(K_1, K_2)}{4\sigma\sqrt{2}}\mu(K_1)^{3/2}. \end{aligned}$$

A similar inequality results when $\mu(K_2) \leq \mu(K_1)$, whence the result follows. □

IV. Lower Bounds on Generalization Error

In this section we obtain lower bounds on the generalization error of classification problems. The generalization error is the weight of the region where the chosen classifier and the true classifier differ. This in turn is related to the weight of the no-man's land. Appealing to the isoperimetric inequality of Theorem 2, we use the size (in the geometric sense of distance between sets) of the no-man's land, to obtain bounds on the weight it must carry. Thus we show that the size of the no-man's land can be a tractable measure providing good bounds on the measure of the set where two classifiers differ. We also point out that in the absence of β-log-concavity, no such bounds are valid.

Lower bounds on the generalization error in classification require a careful definition of the probabilistic setup. In this section we consider a generic setup where proper learning is possible. We consider the standard classification problem where data points $x \in \mathbb{R}^n$ and labels $y \in \{-1, 1\}$ are given, and not necessarily generated according to any particular distribution. We assume that we are given a set of classifiers $\mathcal{H}$ which are functions from $\mathbb{R}^n$ to $\{-1, 1\}$. For now, by a slight abuse of notation, we use $\mathcal{H}$ to refer both to the full family of classifiers, and the subset of classifiers that have zero error on the training data. Thus when speaking of linear classifiers, it is understood that by $\mathcal{H}$ we mean the subset of linear classifiers that correctly classify the training data. In our model, the performance of the classifier is measured using some probability measure induced by a β-log-concave density f. We note that this model deviates from the "classical" statistical machine learning setup.

Given a density f, the disagreement of a classifier $h \in \mathcal{H}$ with another classifier h' is defined as:

$$\begin{aligned} \Delta(h; h') &\triangleq \int_{\mathbb{R}^n} \frac{1}{2}(1 - h(x)h'(x))f(x)dx \\ &= \mu\{x \in \mathbb{R}^n : h(x) \neq h'(x)\}, \end{aligned}$$

where μ is the probability measure induced by f. If there exists a true classifier h^{true} (not necessarily in $\mathcal{H}$) such that $y = h^{true}(x)$ then the error of h is $\Delta(h; h^{true})$. For a classifier h, let $K^+(h) \triangleq \{x \in K : h(x) = 1\}$, and similarly $K^-(h) \triangleq \{x \in K : h(x) = -1\}$. Given a pair of classifiers h_1 and h_2 we define the distance between them as $\operatorname{dist}(h_1, h_2)$, given by

$$\max\left\{d\left(K^+(h_1), K^-(h_2)\right), d\left(K^-(h_1), K^+(h_2)\right)\right\}.$$

We note that $\operatorname{dist}(h_1, h_2)$ may equal zero even if the classifiers are rather different. However, in some cases, $\operatorname{dist}(h_1, h_2)$ provides a useful measure of difference; see Proposition 2 below. We consider later generalizations of '$\operatorname{dist}(\cdot, \cdot)$' which are interesting exactly when the classifiers are different, but the distance defined above is zero.

Suppose we have to choose a classifier from a set $\mathcal{H}$. This may occur if, for example, we are given sample data points and there are several classifiers that classify the data correctly. The following theorem states that if the set of classifiers we choose from is too large, then the error might be large as well. Note that we have to scale the error lower bound by the minimal weight of the positively/negatively labelled region.

Theorem 3: Suppose that f is β-log-concave defined on a closed and bounded set K with nonzero diameter[2]. Then for every $h \in \mathcal{H}$, for every $\epsilon > 0$[3], there exists $h' \in \mathcal{H}$ such that

$$\begin{aligned} \Delta(h; h') &\geq \frac{e^{-\beta}P_1}{\operatorname{diam}(K)}\left(\sup_{h_1 \in \mathcal{H}} \operatorname{dist}(h, h_1) - \epsilon\right) && \text{(IV.5)} \\ &\geq \frac{e^{-\beta}P_1}{\operatorname{diam}(K)}\frac{1}{2}\left(\sup_{h_1, h_2 \in \mathcal{H}} \operatorname{dist}(h_1, h_2) - \epsilon\right), && \text{(IV.6)} \end{aligned}$$

where $P_1 = \inf_{\tilde{h} \in \mathcal{H}} \min\{\mu(K^+(\tilde{h})), \mu(K^-(\tilde{h}))\}$.
Without the β-log-concavity assumption, this result need not hold. Indeed, without it, we may have classifiers with $\operatorname{dist}(h, h')$ large, but with little or zero weight on the region in which they differ.

Proof. If $\sup_{h_1, h_2 \in \mathcal{H}} \operatorname{dist}(h_1, h_2) = 0$, the result follows, so we can assume this is not the case. For every $\epsilon > 0$ we can choose $h' \in \mathcal{H}$ such that $\operatorname{dist}(h, h') \geq \sup_{h_1 \in \mathcal{H}} \operatorname{dist}(h, h_1) - \epsilon$. We consider the case where $\operatorname{dist}(h, h') = d(K^+(h), K^-(h'))$; the other case where $\operatorname{dist}(h, h') = d(K^-(h), K^+(h'))$ follows in a symmetric manner. Let $B = K \setminus (K^+(h) \cup K^-(h'))$. It follows from Theorem 2 that

$$\mu(B) \geq e^{-\beta}\frac{\operatorname{dist}(h, h')}{\operatorname{diam}(K)}\min\left\{\mu(K^+(h)), \mu(K^-(h'))\right\}. \tag{IV.7}$$

From here the first inequality of the theorem follows. Now for the second inequality, similarly to the above, for every $\varepsilon > 0$ we can pick h_1, h_2 so that $\operatorname{dist}(h_1, h_2) \geq \sup_{h_1', h_2'} \operatorname{dist}(h_1, h_2) - \varepsilon$. By Theorem 2, letting $B = K \setminus (K^+(h_1) \cup K^-(h_2))$ inequality (IV.7) holds with h_1, h_2 in place of h, h'. Now, $\Delta(h; h_1) \geq \int_B \chi_{\{h(x) \neq h_1(x)\}} f(x)dx$ and $\Delta(h; h_2) \geq \int_B \chi_{\{h(x) \neq h_2(x)\}} f(x)dx$. Since $h_1(x) \neq h_2(x)$ on B, then either $\Delta(h; h_1) \geq \mu(B)/2$ or $\Delta(h; h_2) \geq \mu(B)/2$. Since $P_1 \leq \mu(K^+(h_1))$ and $P_1 \leq \mu(K^-(h_2))$,

[2] Unless explicitly noted, we assume throughout that K is closed and bounded with nonzero diameter.

[3] If $\mathcal{H}$ is compact in an appropriate sense, then we can set $\epsilon = 0$.

and by substituting in Ineq. (IV.7), we obtain that $\Delta(h, h_i) \geq e^{-\beta} \operatorname{dist}(h_1, h_2) P_1/(2 \operatorname{diam}(K))$ for $i = 1$ or $i = 2$. □

The following example demonstrates the power of Theorem 3 in the context of linear classification. Consider an input-output sequence $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ arising from some unknown source (not necessarily β-log-concave) as in the classical binary classification problem. Define $X_N^+ = \{x_i : y_i = 1\}$ and $X_N^- = \{x_i : y_i = -1\}$. Suppose that the true error is measured according to a β-log-concave distribution, and that X_N^+ and X_N^- are linearly separable. Recall that a linear classifier h is a function given by $h(x) = \operatorname{sign}(\langle x, u\rangle + b)$, where '$\operatorname{sign}$' is the sign function and '$\langle\cdot,\cdot\rangle$' is the standard inner product in $\mathbb{R}^n$. The following proposition provides a lower bound on the true error. We state it for generic sets of vectors, so the data are not assumed to be sampled from any concrete source. The lower bound concerns the case where we are faced with a choice from a set of classifiers, all of which agree with the data (i.e., have zero training error). If we commit to any specific classifier, then there exists another classifier (whose training error is zero as well) such that the true error of the classifier we committed to is relatively large if the other classifier happens to equal h^{true}.

Proposition 2: Suppose that we are given two sets of linearly separable vectors $X^+, X^- \subseteq K$ and let $t = d(\operatorname{conv}(X^+), \operatorname{conv}(X^-))$. Then for every linear classifier h that separates X^+ and X^-, and any β-log-concave density f and induced measure μ defined on a bounded set K, there exists another linear classifier h' that separates X^+ and X^- as well, such that

$$\Delta(h; h') \geq e^{-\beta} P_1 t/(2 \operatorname{diam}(K)),$$

where

$$P_1 = \min\{\mu\{x : \langle x, u\rangle \geq \langle x^+, u\rangle\}, \mu\{x : \langle x, u\rangle \leq \langle x^-, u\rangle\}\}$$

for some $x^\pm \in \operatorname{conv}(X^\pm)$ such that $d(x^+, x^-) = t$ and $u = (x^+ - x^-)/2$.

PROOF. Let $\mathcal{H}$ be the set of all hyperplanes that separate X^+ from X^-. It follows by a standard linear programming argument (see [26]) that $\sup_{h_1,h_2\in\mathcal{H}} \operatorname{dist}(h_1, h_2) = t$. This is attained for $h_1(x) = \operatorname{sign}(\langle x, u\rangle - \langle x^+, u\rangle)$ and $h_2(x) = \operatorname{sign}(\langle x, u\rangle - \langle x^-, u\rangle)$. We now apply Theorem 3 to obtain the desired result. Note that P_1 in the declaration of the proposition is tighter than P_1 in Theorem 3. This is the result of calculating $\mu(K^+(h_1))$ and $\mu(K^-(h_2))$ directly (instead of taking the infimum as in Theorem 3). □

Finally, we note that inequality (IV.5) is in general strictly stronger than (IV.6), since the inequality $\sup_{h'} \operatorname{dist}(h, h') \geq \sup_{h_1,h_2} \operatorname{dist}(h_1, h_2)/2$ is usually strict. If, on the other hand, h is the maximum margin classifier, then the two bounds coincide. In the linear case, the maximum margin classifier is the "safest choice." Thus we have a reinterpretation of the maximum margin classifier as the "safest" classifier under worst-case (minimax) assumptions.

A More General Notion of Distance

In the above discussion, we show how the isoperimetric inequality can essentially use the measure $\sup_{h_1,h_2\in\mathcal{H}} \operatorname{dist}(h_1, h_2)$ to obtain bounds on the generalization error. As remarked above, Theorem 3 says nothing if $\sup_{h_1,h_2\in\mathcal{H}} \operatorname{dist}(h_1, h_2) = 0$. Generalizing the notion of 'dist' for classifiers, and considering the distance from a single classifier h' to a family $\mathcal{H}' \subseteq \mathcal{H}$ of classifiers, we can obtain a stronger measure that again allows us to derive a bound on $\sup_{h_1,h_2} \Delta(h_1; h_2)$, the generalization error.

Definition 2: Given a classifier h, and r other classifiers $\mathcal{H}' = \{h_1, \ldots, h_r\}$, define the sets

$$\begin{aligned} K^+ = K^+(\{h\} \cup \mathcal{H}') &= K^+(h) \cap \bigcap_{i=1}^{r} K^+(h_i) \\ K^- = K^-(\{h\} \cup \mathcal{H}') &= K^-(h) \cap \bigcap_{i=1}^{r} K^-(h_i), \end{aligned}$$

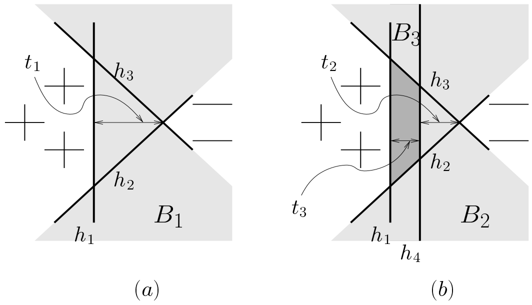
so that $K^+(\{h\} \cup \mathcal{H}')$ is the set of points that all classifiers in $\mathcal{H}'$, and also h, label as '$+$', and similarly for K^-. Now we define the distance measure, $\operatorname{dist}(h, \mathcal{H}')$, from a classifier h to a family of classifiers $\mathcal{H}$, to be the Euclidean distance:

$$\operatorname{dist}(h, \mathcal{H}') = \operatorname{dist}(K^+, K^-).$$

If the intersection is empty, we define dist to be zero.
For $k = 2$, Figure 2 illustrates the generalized concept of dist, and further shows that Theorem 3 holds with $\sup_{h_1,h_2\in\mathcal{H}} \operatorname{dist}(h_1, h_2)$ replaced by the new distance concept,

$$\sup_{\{h,h_1,h_2\}\subseteq\mathcal{H}} \operatorname{dist}(h, \{h_1, h_2\}).$$

Indeed the phenomenon illustrated in Figure 2 holds in general. We can restate Theorem 3 using the distance to a fixed classifier. The statement of Theorem 3 now becomes:



(a) (b)

Fig. 2. In (a) above we have three classifiers, $\{h_1, h_2, h_3\}$, so that for any two, $\operatorname{dist}(h_i, h_j) = 0$. Nevertheless, for any $h \in \{h_1, h_2, h_3\}$, there exists $h' \in \{h_1, h_2, h_3\}$ (with $h' \neq h$) so that if $h^{true} = h'$, then $\Delta(h; h') \geq \mu(B_1)/2$, where B_1 is the shaded area. To see this, note for instance that if $h = h_1$, then $B_1 = \Delta(h_1; h_2) \cup \Delta(h_1; h_3)$. We can then get a bound on $\mu(B_1)$ by using the isoperimetric inequality, and the fact that $\operatorname{dist}(h, \{h_i, h_j\}) = t_1$. In (b) we add a fourth classifier h_4. Here we see that if we choose $h = h_4$, then the worst case generalization error is lower bounded by comparing the two distance measures, $\sup_{h_1\in\mathcal{H}} \operatorname{dist}(h, h_1)$ (our previous distance measure) and $(1/2)\sup_{\{h_1,h_2\}\subset\mathcal{H}} \operatorname{dist}(h, \{h_1, h_2\})$. In this example, $\sup_{h_1\in\mathcal{H}} \operatorname{dist}(h, h_1) = t_3$, and $\sup_{\{h_1,h_2\}\subset\mathcal{H}} \operatorname{dist}(h, \{h_1, h_2\}) = t_2$.

Theorem 3′: Under the assumptions of Theorem 3, for every $h \in \mathcal{H}$ there exists $h' \in \mathcal{H}$ such that

$$\begin{aligned}\Delta(h;h') \quad \geq \quad & \frac{e^{-\beta}}{\mathrm{diam}(K)}\max\Bigg\{P_1 \sup_{h_1\in\mathcal{H}} \mathrm{dist}(h,h_1), \\ & \frac{1}{2}P_2 \sup_{\{h_1,h_2\}\subseteq\mathcal{H}} \mathrm{dist}(h,\{h_1,h_2\})\Bigg\}\end{aligned}$$

where we have

$$\begin{aligned}P_1 &= \inf_{h_1\in\mathcal{H}} \min\{\mu(K^+(h_1)),\mu(K^-(h_1))\} \\ P_2 &= \inf_{\{h_1,h_2\}\subset\mathcal{H}} \min\{\mu(K^+(\{h_1,h_2\})),\mu(K^-(\{h_1,h_2\}))\}.\end{aligned}$$

The proof follows directly. This restatement of Theorem 3 is in fact somewhat stronger, since when the two inequalities of Theorem 3 do not coincide, the $\max$ in the restatement picks out the stronger lower bound.

We can obtain a general version of this result.

Theorem 4: Under the assumptions of Theorem 3, for every $h \in \mathcal{H}$ and for any $\epsilon > 0$, there exists $h' \in \mathcal{H}$ such that $\Delta(h;h')$ is bounded below by

$$\frac{e^{-\beta}}{\mathrm{diam}(K)}\max_{r\in\mathbb{N}}\left\{\frac{1}{r}P_r(\sup_{\{h_1,\ldots,h_r\}\subseteq\mathcal{H}} \mathrm{dist}(h,\{h_1,\ldots,h_r\})-\epsilon)\right\},$$

where P_r is given by

$$\inf_{\{h_1,\ldots,h_r\}\subset\mathcal{H}} \min\{\mu(K^+(\{h_1,\ldots,h_r\})),\mu(K^-(\{h_1,\ldots,h_r\}))\}. \tag{IV.8}$$

PROOF. This proof closely follows that of Theorem 3.[4] For every $\epsilon > 0$ and $r \in \mathbb{N}$, choose $\{h_1,\ldots,h_r\} \subseteq \mathcal{H}$ that attain the supremum on the right hand side of (IV.8) within ϵ. Let $K^+ = K^+(h,h_1,\ldots,h_r)$, and $K^- = K^-(h,h_1,\ldots,h_r)$. Now let $B = K \setminus (K^+ \cup K^-)$. By Theorem 2, it follows that

$$\mu(B) \geq e^{-\beta}\frac{\mathrm{dist}(h,\{h_1,\ldots,h_r\})}{\mathrm{diam}(K)}\min\{\mu(K^+),\mu(K^-)\}. \tag{IV.9}$$

Now, we can write B as the union of r possibly overlapping sets, where each set defines the area where h differs with one of the h_i:

$$\begin{aligned}B &= \bigcup_{i=1}^{r}\{h \neq h_i\} \\ &= \left[\bigcup_{i=1}^{r}(K^-(h)\cap K^+(h_i))\right] \cup \left[\bigcup_{i=1}^{r}(K^+(h)\cap K^-(h_i))\right] \\ &= \bigcup_{i=1}^{r}\left[\left(K^-(h)\cap K^+(h_i)\right) \cup \left(K^+(h)\cap K^-(h_i)\right)\right].\end{aligned}$$

The second equality follows by the associativity of unions. For the first equality, we have simply expanded out our definition of K^+ and K^- from above: Take any $x \in B$. Suppose $x \in K^-(h)$. Since $x \in B$, then $x \notin K^-$, and thus there must exist some i for which $x \notin K^-(h_i)$, which means $x \in K^+(h_i)$, and hence $x \in K^-(h) \cap K^+(h_i)$. The reverse inclusion follows similarly.

Consequently, we have

$$\mu(B) \leq r \max_i \left\{\mu((K^-(h)\cap K^+(h_i)) \cup (K^+(h)\cap K^-(h_i)))\right\}.$$

Letting i^* be the maximizing index of the right hand side, we have $\Delta(h,h_{i^*}) \geq \frac{1}{r}\mu(B)$. Substituting in Equation IV.9, we then have that $\Delta(h,h_{i^*})$ is at least

$$\frac{e^{-\beta}}{\mathrm{diam}(K)}\left(\min\{\mu(K^+),\mu(K^-)\}\right)\frac{1}{r}\mathrm{dist}(h,\{h_1,\ldots,h_r\}).$$

This concludes the proof. □

V. Regression Tubes

In this section we consider regression problems, and provide results of a different flavor. Throughout the section we let k be a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. We provide lower bounds on the weight of tubes around k. The probabilistic setup is as follows. We have a probability measure μ with density f on $\mathbb{R}^{n+m}$ that prescribes the probability of getting a pair $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$. The density f has support on the set K. For a specific function $k : \mathbb{R}^n \to \mathbb{R}^m$ we consider the set

$$T^k_{\epsilon_0,\epsilon_1} \stackrel{\triangle}{=} \{(x,y) : \epsilon_0 \leq \|k(x)-y\| \leq \epsilon_1\}.$$

This set represents all the pairs where the prediction of k is off by more than ϵ_0 and less than ϵ_1, or alternatively, the set of pairs whose prediction is converted to zero error when changing the ϵ in an ϵ-insensitive error criterion from ϵ_0 to ϵ_1. Different assumptions on the joint density f lead to different results. We start with a simple case representing an additive independent noise model, and then consider the case where f is β-log-concave jointly in x and y. We finally consider the more complicated case, where f is β-log-concave in x and β'-log-concave in y conditioned on x. We provide a lower bound on the measure of the tube under some continuity assumptions.

As a motivation, consider the classical regression setup where

$$Y = k(X) + N, \tag{V.10}$$

where X is the independent random variable, N is additive noise, and Y is the dependent variable. The results of this section apply to non-additive noise models as well. If the noise N is arbitrary, we cannot hope to obtain a bound on the measure of the intermediate tube in terms of the inner and outer tubes, since the noise may alternate between putting the weight on the inner and the outer tubes. We make certain specific assumptions concerning the continuity of the noise process.

Let us define the projection of a tube for a specific x by

$$T^k_{\epsilon_0,\epsilon_1}(x) \stackrel{\triangle}{=} \{y : \epsilon_0 \leq \|k(x)-y\| \leq \epsilon_1\}.$$

We denote the marginal density by f_x and the conditional density by $f_{y|x}$. The associated measures are then denoted by μ_x and $\mu_{y|x}$.

If the noise in (V.10) is independent of x, we can straightforwardly derive a lower bound on the measure of the intermediate tube.

Corollary 3: Consider the model of Eq. (V.10). Suppose that N is independent of x and has a β-log-concave distribution. Suppose further that N has bounded support K_Y, with

[4] Note again that if $\mathcal{H}$ is compact, we can set $\epsilon = 0$. Furthermore, the $\max$ over $\mathbb{N}$ is attained for some finite $r \in \mathbb{N}$, since dist is bounded uniformly with respect to r in $\mathbb{N}$.

$0 \in K_Y$ (this is always true if N has zero mean). Then for every probability measure on X, we have that:

$$\mu(T^k_{\epsilon_0,\epsilon_1}) \geq \frac{(\epsilon_1-\epsilon_0)e^{-\beta}}{\mathrm{diam}(K_Y)} \min\left\{\mu(T^k_{0,\epsilon_0}), \mu(T^k_{\epsilon_1,\mathrm{diam}(K_Y)})\right\}. \tag{V.11}$$

PROOF. Since N is independent of x and using Theorem 2, we have that for every x:

$$\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) \geq e^{-\beta}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K_Y)} \cdot$$
$$\min\left\{\mu_{y|x}(T^k_{0,\epsilon_0}(x)), \mu_{y|x}(T^k_{\epsilon_1,\mathrm{diam}(K_Y)}(x))\right\}.$$

By definition,

$$\mu(T^k_{0,\epsilon_0}) = \int_{\mathbb{R}^n} f_x(x)\mu_{y|x}(T^k_{0,\epsilon_0}(x))dx,$$

and similarly for $T^k_{\epsilon_0,\epsilon_1}$ and $T^k_{\epsilon_1,\infty}$. Since $\mu_{y|x}(T^k_{0,\epsilon_0}(x))$ is the same for all x, we obtain (V.11). □

It is worth mentioning that Corollary 3 does not require *any* assumptions on X, and in particular the support of f_x is not assumed bounded. For the case of unbounded noise with finite variance (e.g., Gaussian noise) one can use Proposition 1 instead of Theorem 2 and obtain a similar bound (replacing $\mathrm{diam}(K_Y)$ with $4\sqrt{2}\sigma$ and having power of $3/2$ inside the minimum).

We next consider the case where the density $f(x,y)$ is β-log-concave jointly in x and y. This may arise in a situation where Eq. (V.10) holds with N independent of x, but we do not know what is the true k function. In that case we can still consider the measure of the intermediate tube defined by some other function k'. The linear case is particularly simple as the next lemma shows:

Lemma 6: Suppose that the model of Eq. (V.10) holds and that X is β-log-concave. If N is β'-log-concave and if $k(x)$ is linear, we have that $f(x,y)$ is $(\beta+\beta')$-log-concave.

PROOF. We have that $f(x,y) = f_x(x)f_{y|x}(y|x) = f_x(x)g(y-Ax)$, where A is some matrix ($k(x) = Ax$) and g is a β'-log-concave conditional noise density. For $\lambda \in [0,1]$ we have that

$$\begin{aligned}
&f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2))\\
=\ & f(\lambda x_1+(1-\lambda)x_2, \lambda y_1+(1-\lambda)y_2)\\
=\ & f_x(\lambda x_1+(1-\lambda)x_2)\cdot\\
&\qquad g(\lambda y_1+(1-\lambda)y_2 - A(\lambda x_1+(1-\lambda)x_2))\\
=\ & f_x(\lambda x_1+(1-\lambda)x_2)\cdot\\
&\qquad g(\lambda(y_1-Ax_1)+(1-\lambda)(y_2-Ax_2))\\
\geq\ & e^{-\beta}f_x(x_1)^\lambda f_x(x_2)^{1-\lambda}e^{-\beta'}g(y_1-Ax_1)^\lambda g(y_2-Ax_2)^{1-\lambda}\\
=\ & e^{-(\beta+\beta')}f(x_1,y_1)^\lambda f(x_2,y_2)^{1-\lambda}.
\end{aligned}$$

□

Corollary 4: Suppose that $f(x,y)$ is β-log-concave on a bounded set $K \subseteq \mathbb{R}^{n+m}$, with induced measure μ. Assume that k is Lipschitz continuous with constant L, and that $(x,k(x)) \in K$ for every $x \in K|_X$. Then for every $\epsilon_1 > \epsilon_0 > 0$,

$$\mu(T^k_{\epsilon_0,\epsilon_1}) \geq \frac{(\epsilon_1-\epsilon_0)e^{-\beta}}{\sqrt{L^2+1}\,\mathrm{diam}(K)}\min\left\{\mu(T^k_{0,\epsilon_0}), \mu(T^k_{\epsilon_1,\mathrm{diam}(K)})\right\}. \tag{V.12}$$

PROOF. We use Theorem 2 with the decomposition $K_1 = T^k_{0,\epsilon_0}$, $B = T^k_{\epsilon_0,\epsilon_1}$ and $K_2 = T^k_{\epsilon_1,\mathrm{diam}(K)}$. By the Lipschitz continuity of k, we get $d(T^k_{0,\epsilon_0}, T^k_{\epsilon_1,\mathrm{diam}(K)}) \geq (\epsilon_1-\epsilon_0)/\sqrt{L^2+1}$, as follows. Take any $(x_1,y_1) \in T^k_{0,\epsilon_0}$, and $(x_2,y_2) \in T^k_{\epsilon_1,\mathrm{diam}(K)}$. Let $\eta = ||x_1-x_2||$. Then $||k(x_1)-k(x_2)|| \leq L\eta$, and thus $||y_1-y_2|| \geq (\epsilon_1-\epsilon_0)-L\eta$, and thus $d((x_1,y_1),(x_2,y_2))^2 \geq \eta^2 + ((\epsilon_1-\epsilon_0)-L\eta)^2$. Optimizing this bound over η we find that $\eta = (\epsilon_1-\epsilon_0)L/(1+L^2)$ gives the desired bound. □

A direct implication of Corollary 4 and Lemma 6 is that if the true model is linear, and both X and $Y|x$ are β and β' log-concave, respectively, then every function (not necessarily linear) satisfies inequality (V.12).

We now consider a different model where instead of assuming that x and y are jointly β-log-concave, we assume that x is β-log-concave and that y is β'-log-concave conditioned on x. We define $K|_X$ to be the projection of K on the first n dimensions.

Definition 3: A density $f(x,y)$ is β-β' conditional log-concave if the marginal $f(x) = \int_{\mathbb{R}^m} f(x,y)$ is β-log-concave and if the conditional $f(y|x) = f(y,x)/f(x)$ is β'-log-concave for all $x \in K|_X$.

The following theorem asserts that a similar bound to (V.12) can be obtained even for β-β' conditional log-concave distributions. The setup is, however, considerably more general. It includes, for example, regression where the independent parameter, x, is sampled from a uniform distribution, and the dependent parameter equals $y = k(x) + N$, where N is some β-log-concave function that depends on x. We denote by B^ℓ the unit ball in $\mathbb{R}^\ell$.

Theorem 5: Suppose that $f(x,y)$ is β-β' conditional log-concave on a bounded set $K \subseteq \mathbb{R}^{n+m}$, with induced measure μ. Fix $\epsilon_1 > \epsilon_0 > 0$. Assume further that there exist constants $C > 0$, $\delta_0 > 0$, and $\rho > 0$ such that for all $(x,y) \in T^k_{\epsilon_0,\epsilon_1}$, $\delta, \delta' < \hat{\delta} \leq \delta_0$, $u_x \in B^m$, and $u_y \in B^n$:

$$\left|1 - \frac{f_{y|x}(y+\delta u_y|x+\delta' u_x)}{f_{y|x}(y|x)}\right| \leq C\hat{\delta}^\rho.$$

Assume that k is Lipschitz continuous with constant L. Then: $\mu(T^k_{\epsilon_0,\epsilon_1})$ is lower bounded by

$$\frac{1}{18}\frac{(\epsilon_1-\epsilon_0)e^{-(\beta+\beta')}}{\mathrm{diam}(K)}\min\left\{1, \frac{\min\{\delta_0, 1/(2C)^{1/\rho}\}}{\mathrm{diam}(K)\max\{1,L\}}\right\}\cdot$$
$$\min\left\{\mu(T^k_{0,\epsilon_0}), \mu(T^k_{\epsilon_1,\mathrm{diam}(K)})\right\}.$$

PROOF. Fix positive $\epsilon_0 < \epsilon_1$. For a set $\mathcal{X} \subseteq K|_X$ (this is a set in $\mathbb{R}^n$) we denote the extension to a set in $\mathbb{R}^n \times \mathbb{R}^m$ by $\mathrm{ext}_K(\mathcal{X}) = \{(x,y) : x \in \mathcal{X} \text{ and } (x,y) \in K\}$. We now define two sets:

$$\begin{aligned}
\mathcal{X}^{in} &= \left\{x : \mu_{y|x}\left(T^k_{0,\epsilon_0}(x)\right) \geq \mu_{y|x}\left(T^k_{\epsilon_1,\mathrm{diam}(K)}(x)\right)\right\}\\
\mathcal{X}^{out} &= \left\{x : \mu_{y|x}\left(T^k_{0,\epsilon_0}(x)\right) \leq \mu_{y|x}\left(T^k_{\epsilon_1,\mathrm{diam}(K)}(x)\right)\right\}.
\end{aligned}$$

Note that $\mathcal{X}^{in} \cup \mathcal{X}^{out} = K|_X$. We consider the following three cases.

Case 1: $\mathcal{X}^{in} = K|_X$, that is, the inner tube is always heavier than the outer tube. In this case, for every x we apply Theorem

2 to the conditional measure, to obtain:

$$\mu_{y|x}\left(T^k_{\epsilon_0,\epsilon_1}(x)\right) \ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu_{y|x}\left(T^k_{\epsilon_1,\mathrm{diam}(K)}(x)\right).$$

Similarly to Corollary 3, integrating over all $x \in K|_X$, we obtain that

$$\begin{aligned}\mu\left(T^k_{\epsilon_0,\epsilon_1}\right) &= \int_{K|_x} f_x(x)\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x))\,dx \\ &\ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu\left(T^k_{\epsilon_1,\mathrm{diam}(K)}\right).\end{aligned}$$

So the result holds.

Case 2: $\mathcal{X}^{out} = K|_X$. In this case the outer tube is heavier for all x. In a similar manner to the previous case we can prove that

$$\begin{aligned}\mu\left(T^k_{\epsilon_0,\epsilon_1}\right) &= \int_{K|_x} f_x(x)\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x))\,dx \\ &\ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu\left(T^k_{0,\epsilon_1}\right).\end{aligned}$$

Case 3: Both $\mathcal{X}^{in} \ne K|_X$ and $\mathcal{X}^{out} \ne K|_X$. In that case it follows from continuity of f and k that the set $\mathcal{X}^{eq} = \mathcal{X}^{in} \cap \mathcal{X}^{out}$ is not empty. For every x in $\mathcal{X}^{eq}$ we have from Theorem 2 that:

$$\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) \ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu_{y|x}(T^k_{0,\epsilon_0}(x)).$$

Since $\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) + 2\mu_{y|x}(T^k_{0,\epsilon_0}(x)) = 1$ for $x \in \mathcal{X}^{eq}$, by solving for $\mu_{y|x}(T^k_{0,\epsilon_0}(x))$, substituting in the inequality above and collecting terms, we obtain that

$$\begin{aligned}\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) &\ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\cdot\frac{1}{2+e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}} \\ &\ge \frac{e^{-\beta'}}{3}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}.\end{aligned}$$

We first consider the continuity of $\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x))$ as a function of x.

$$\begin{aligned}&\left|\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) - \mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x'))\right| \\ &= \left|\int_{T^k_{\epsilon_0,\epsilon_1}(x)} f_{y|x}(y|x) - \int_{T^k_{\epsilon_0,\epsilon_1}(x')} f_{y|x}(y|x')\right| dy \\ &= \left|\int_{T^k_{\epsilon_0,\epsilon_1}(x)} f_{y|x}(y|x) - \int_{T^k_{\epsilon_0,\epsilon_1}(x)} f_{y|x}(y-k(x)+k(x')|x')\right| dy \\ &= \left|\int_{T^k_{\epsilon_0,\epsilon_1}(x)} f_{y|x}(y|x)(1-\frac{f_{y|x}(y-k(x)+k(x')|x')}{f_{y|x}(y|x)})\right| dy \\ &\le \int_{T^k_{\epsilon_0,\epsilon_1}(x)}\Big| f_{y|x}(y|x)\cdot \\ &\qquad (1-\frac{f_{y|x}(y+L\|x-x'\|u_y|x+\|x-x'\|u_x)}{f_{y|x}(y|x)})\Big|\,dy,\end{aligned}$$

where the last inequality is due to the Lipschitz continuity of k and $u_x \in B^m$ and $u_y \in B^n$. It follows from the continuity assumption on $f_{y|x}$ that if $\max\{L,1\}\|x-x'\| \le \delta_0$,

$$\begin{aligned}&\left|\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) - \mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x'))\right| \\ &\quad\le \mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x))C\left(\max\{L,1\}\|x-x'\|\right)^\rho.\end{aligned}$$

Fix $\delta = \min\{\delta_0, 1/(2C)^{1/\rho}\}/\max\{1,L\}$. Then for every x' such that $\|x-x'\| \le \delta$, we have that

$$\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x')) \ge \frac{1}{2}\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)). \tag{V.13}$$

Define $\mathcal{X}^{eq,\delta}$ to be the δ expansion of $\mathcal{X}^{eq}$, that is

$$\mathcal{X}^{eq,\delta} = \{x \in K|_X, \|x-x'\| \le \delta \text{ for some } x \in \mathcal{X}^{eq}\}.$$

Assume first that $\mu(T^k_{0,\epsilon_0}) \le \mu(T^k_{\epsilon_1,\mathrm{diam}(K)})$. We bound $\mu(T^k_{\epsilon_0,\epsilon_1})$ in terms of $\mu(T^k_{0,\epsilon_0})$. We have that:

$$\begin{aligned}&\mu(\mathrm{ext}_K(\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta})\cap T^k_{0,\epsilon_0}) + \mu(\mathrm{ext}_K(\mathcal{X}^{eq,\delta})\cap T^k_{0,\epsilon_0}) \\ &\quad + \mu(\mathrm{ext}_K((\mathcal{X}^{in}\setminus\mathcal{X}^{eq,\delta}))\cap T^k_{0,\epsilon_0}) = \mu(T^k_{0,\epsilon_0}).\end{aligned} \tag{V.14}$$

One of the summands in (V.14) must be at least $\mu(T^k_{0,\epsilon_0})/3$. We consider each case separately:

Case 3.1: $\mu(\mathrm{ext}_K(\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta})\cap T^k_{0,\epsilon_0}) \ge \mu(T^k_{0,\epsilon_0})/3$. In that case we have that

$$\begin{aligned}\mu(T^k_{\epsilon_0,\epsilon_1}) &\ge \int_{\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta}} f_x(x)\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)) \\ &\ge \int_{\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta}} f_x(x)e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu_{y|x}(T^k_{0,\epsilon_0}(x)) \\ &= e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\int_{\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta}} f_x(x)\mu_{y|x}(T^k_{0,\epsilon_0}(x)) \\ &\ge e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\frac{\mu(T^k_{0,\epsilon_0})}{3}.\end{aligned} \tag{V.15}$$

Case 3.2: $\mu(\mathrm{ext}_K(\mathcal{X}^{eq,\delta})\cap T^k_{0,\epsilon_0}) \ge \mu(T^k_{0,\epsilon_0})/3$. As before,

$$\mu(T^k_{\epsilon_0,\epsilon_1}) \ge \int_{\mathcal{X}^{eq,\delta}} f_x(x)\mu_{y|x}(T^k_{\epsilon_0,\epsilon_1}(x)).$$

Since Ineq. (V.13) holds for all $x \in \mathcal{X}^{eq,\delta}$ for our choice of δ, we obtain:

$$\begin{aligned}\mu(T^k_{\epsilon_0,\epsilon_1}) &\ge \int_{\mathcal{X}^{eq,\delta}} f_x(x)\frac{1}{6}e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)} \\ &= \frac{1}{6}e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\mu(\mathrm{ext}_K(\mathcal{X}^{eq,\delta})) \\ &\ge \frac{1}{6}e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\mathrm{diam}(K)}\frac{\mu(T^k_{0,\epsilon_0})}{3}.\end{aligned} \tag{V.16}$$

Case 3.3: $\mu(\mathrm{ext}_K(\mathcal{X}^{in}\setminus\mathcal{X}^{eq,\delta})\cap T^k_{0,\epsilon_0}) \ge \mu(T^k_{0,\epsilon_0})/3$. In that case we have that $\mu_x(\mathcal{X}^{in}\setminus\mathcal{X}^{eq,\delta}) \ge \mu(T^k_{0,\epsilon_0})/3$. Since we assumed that $\mu(T^k_{0,\epsilon_0}) \le \mu(T^k_{\epsilon_1,\mathrm{diam}(K)})$ one of the three cases must hold:

Case 3.3.1: $\mu(\mathrm{ext}_K(\mathcal{X}^{in}\setminus\mathcal{X}^{eq,\delta})\cap T^k_{\epsilon_1,\mathrm{diam}(K)}) \ge \mu(T^k_{0,\epsilon_0})/3$. In this case we can use the same maneuver as in Case 3.1 (applied to $\mathcal{X}^{in}$) and (V.15) holds.

Case 3.3.2: $\mu(\mathrm{ext}_K(\mathcal{X}^{eq,\delta})\cap T^k_{\epsilon_1,\mathrm{diam}(K)}) \ge \mu(T^k_{0,\epsilon_0})/3$. In this case we can use the same maneuver as in Case 3.2 and (V.16) holds.

Case 3.3.3: $\mu(\mathrm{ext}_K(\mathcal{X}^{out}\setminus\mathcal{X}^{eq,\delta})\cap T^k_{\epsilon_1,\mathrm{diam}(K)}) \ge \mu(T^k_{0,\epsilon_0})/3$. In this case we have that $\mu_x(\mathcal{X}^{out}) \ge \mu(T^k_{0,\epsilon_0})/3$ and $\mu_x(\mathcal{X}^{in}) \ge \mu(T^k_{0,\epsilon_0})/3$. We can use Theorem 2 for μ_x and obtain that:

$$\mu_x(\mathcal{X}^{eq,\delta}) \ge e^{-\beta}\frac{\delta}{\mathrm{diam}(K)}\mu(T^k_{0,\epsilon_0})/3.$$

Substituting the value of δ, and similarly to Case 3.2:

$$\begin{aligned} & \mu(T^k_{\epsilon_0,\epsilon_1}) \\ \geq \; & \int_{\mathcal{X}^{eq,\delta}} f_x(x)\frac{1}{6}e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\operatorname{diam}(K)} \\ = \; & \frac{1}{6}e^{-\beta'}\frac{\epsilon_1-\epsilon_0}{\operatorname{diam}(K)}\mu(\operatorname{ext}_K(\mathcal{X}^{eq,\delta})) \\ \geq \; & \frac{1}{6}e^{-(\beta'+\beta)}\frac{\epsilon_1-\epsilon_0}{\operatorname{diam}(K)^2}\frac{\min\{\delta_0, 1/(2C)^{1/\rho}\}}{\max\{1,L\}}\frac{\mu(T^k_{0,\epsilon_0})}{3}. \end{aligned}$$

The case where $\mu(T^k_{0,\epsilon_0}) > \mu(T^k_{\epsilon_1,\operatorname{diam}(K)})$ follows similarly. The result follows by taking the worst case of the five cases. □

Several remarks are in order. First, the boundedness assumption can be relaxed in a similar manner to Corollary 2 or Proposition 1, with appropriate changes. Second, a continuity assumption of k is necessary, and counterexamples where discontinuity invalidates the theorem can be easily derived. Third, as a result of the continuity requirement on f, Theorem 5 is dimension independent. If f is instead Lipschitz continuous, one can still retain a similar result, however, a dimension dependent constant would be needed for Case 3.3.3.

VI. Bounding the Size of the Margin

In this section we consider the problem of computing the likelihood that data generated by a β-log-concave distribution will have a large margin, and again show that this question can be approached using the isoperimetric inequality. We consider the standard machine learning setup, and assume that the data are sampled from a β-log-concave distribution. We examine the geometric margin as opposed to the "functional" margin which is often defined with respect to a real valued function g. In that case classification is performed by considering $h(x) = \operatorname{sign}(g(x))$ and the margin of g at $(x, y) \in \mathbb{R}^n \times \{-1, 1\}$ is defined as $g(x)y$. If such a function g is Lipschitz with a constant L, then for $x \in K^+(h)$ the event that $\{d(x, K^-(h)) < \gamma\}$ is contained in the event that $\{g(x) < \gamma L\}$ (and for $x \in K^-(h)$ if $d(x, K^-(h)) < \gamma$ then $-g(x) < \gamma L$). Consequently, results on the geometric margin can be easily converted to results on the "functional" margin as long as the Lipschitz assumption holds.[5]

Suppose now that we have a classifier h, and we ask the following question: what is the probability that if we sample N vectors $\boldsymbol{X}_N = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from f, they are far away from the boundary between $K^+(h)$ and $K^-(h)$. More precisely, we want to bound the probability of the event $\{\min_{i:\boldsymbol{x}_i \in K^+(h)} d(\boldsymbol{x}_i, K^-(h)) > \gamma\}$, and similarly for negatively labelled samples. We next show that the probability that a sampled point is some distance from the boundary, is almost linear in this distance. An immediate consequence is an exponential concentration inequality.

Proposition 3: Suppose we are given a classifier h defined on a bounded set K. Fix some $\gamma > 0$ and consider the set $B = \{x \in K^-(h) : d(x, K^+(h)) < \gamma\}$. Let f be a β-log-concave density on K with induced measure μ. Then

$$\mu(B) \geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\min\left\{\mu(K^+(h)), \frac{\mu(K^-(h))}{1+\gamma e^{-\beta}/\operatorname{diam}(K)}\right\}.$$

Proof. Consider the decomposition of K to $K_1 = K^+(h)$, B, and $K_2 = K^-(h) \setminus B$. By Theorem 2 we know that $\mu(B) \geq \gamma e^{-\beta}\min\{\mu(K_1), \mu(K_2)\}/\operatorname{diam}(K)$. We also know that $\mu(B) = \mu(K^-(h)) - \mu(K_2)$. So that

$$\begin{aligned} \mu(B) \;\geq\; & \max\{\gamma e^{-\beta}\min\{\mu(K_1), s\}/ \\ & \operatorname{diam}(K), \mu(K^-(h)) - s\}, \end{aligned} \qquad \text{(VI.17)}$$

where $s = \mu(K_2)$. Minimizing over s in the interval $[0, \mu(K^-(h))]$, it is seen that the minimizer s is either at the point where $\mu(K^-(h)) - s = \gamma e^{-\beta}\mu(K_1)/\operatorname{diam}(K)$ or at the point where $\mu(K^-(h)) - s = s\gamma e^{-\beta}/\operatorname{diam}(K)$. Substituting those s in Ineq. (VI.17) and some algebra gives the desired result. □

A similar result holds by interchanging K^+ and K^- throughout Proposition 3 and the definition of B. The following corollary is a two-sided version of Proposition 3. It does not have $\gamma e^{-\beta}/\operatorname{diam}(K)$ inside the minimum.

Corollary 5: Suppose we are given a classifier h defined on a bounded set K. Fix some $\gamma > 0$ ($\operatorname{diam}(K) \geq \gamma$) and consider the set $B^{symm} = \{x \in K^-(h) : d(x, K^+(h)) < \gamma\} \cup \{x \in K^+(h) : d(x, K^-(h)) < \gamma\}$. Let f be a β-log-concave density on K with induced measure μ. Then

$$\mu(B^{symm}) \geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\min\left\{\mu(K^+(h)), \mu(K^-(h))\right\}.$$

Proof. Let $B^+ = \{x \in K^+(h) : d(x, K^-(h)) < \gamma\}$ and $B^- = \{x \in K^-(h) : d(x, K^+(h)) < \gamma\}$. We have that $\mu(B^{symm}) = \mu(B^+) + \mu(B^-)$. From Proposition 3 we have that

$$\mu(B^-) \geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\min\left\{\mu(K^+(h)), \frac{\mu(K^-(h))}{1+\gamma e^{-\beta}/\operatorname{diam}(K)}\right\}$$

and

$$\mu(B^+) \geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\min\left\{\mu(K^-(h)), \frac{\mu(K^+(h))}{1+\gamma e^{-\beta}/\operatorname{diam}(K)}\right\}.$$

If the minimum is obtained by $\mu(K^+(h))$ for $\mu(B^-)$ or by $\mu(K^-(h))$ for $\mu(B^+)$, then the result holds. Suppose that the minimum is obtained by the second term for both $\mu(B^-)$ and $\mu(B^+)$. We therefore have in that case that

$$\begin{aligned} \mu(B^{symm}) &= \mu(B^+) + \mu(B^-) \\ &\geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\frac{\mu(K^-(h)) + \mu(K^+(h))}{1+\gamma e^{-\beta}/\operatorname{diam}(K)} \\ &= \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\frac{1}{1+\gamma e^{-\beta}/\operatorname{diam}(K)} \\ &\geq \gamma\,\frac{e^{-\beta}}{\operatorname{diam}(K)}\frac{1}{2}, \end{aligned}$$

where the last inequality follows since $\gamma \leq \operatorname{diam}(K)$. The result follows. □

Corollary 6: Suppose that N samples $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are drawn independently from a β-log-concave density f defined on a bounded set K. Let h be a

[5] This is the case if, for instance, we consider kernel machines where the resulting classifier is of the form $\operatorname{sgn}(\sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}))$, where the kernel itself is Lipschitz continuous, and we also have that the sum of the coefficients is controlled, i.e., $\sum_i |\alpha_i|$ is finite. An L^1 regularization certainly achieves this. L^2 regularization along with a separate sparseness condition is also sufficient.

classifier. Then for every $\gamma > 0$:

$$\begin{aligned}\Pr\left(\min_{\{i\,:\,\boldsymbol{x}_i \in K^-(h)\}} d(\boldsymbol{x}_i, K^+(h)) > \gamma\right) \\ \leq \quad \exp\left(-N\gamma C \min\left\{\mu(K^+(h)), \frac{\mu(K^-(h))}{1+\gamma C}\right\}\right),\end{aligned}$$

where $\Pr$ is the probability measure of drawing N samples from f and $C = e^{-\beta}/\operatorname{diam}(K)$.

PROOF. The proof follows from Proposition 3 and the inequality $(1-a)^N \leq \exp(-aN)$ for $a \in [0,1]$ and $N \geq 0$.□

Corollary 6 is a dimension-free inequality. It implies that when sampling from a β-log-concave distribution, then for any specific classifier we cannot hope to have a large margin. It does not claim, however, that the empirical margin is small. Specifically, for $\boldsymbol{X}_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ one can consider the probabilistic behavior of the following empirical gap between the classes: $\operatorname{gap}(\boldsymbol{X}_N; h) = \min_{i,j: h(\boldsymbol{x}_i) \neq h(\boldsymbol{x}_j)} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The probability that this quantity is larger than γ cannot be bounded in a dimension-free manner. The reason is that as the number of dimensions grows to infinity the distance between the samples may become bounded away from zero. To see this, consider uniformly distributed samples on the unit ball in $\mathbb{R}^n$. If n is much bigger than N, it is not hard to prove that all the sampled vectors will be (with high probability) equally far apart from each other. So $\operatorname{gap}(\boldsymbol{X}_N; h)$ does not converge to 0 (for every non-trivial h) in the regime where n increases fast enough with N. For every fixed n one can bound the probability that $\operatorname{gap}(\boldsymbol{X}_N; h)$ is large using covering number arguments, as in [27], but such a bound must be dimension-dependent.

A quantity related to the empirical margin, is the margin to a set of classifiers with more than one, indeed possibly infinite, classifiers. In such cases, a uniform bound in the spirit of Corollary 6 is of interest. Specifically, let the empirical margin of a classifier h on sample points $\boldsymbol{X}_N$ be denoted by:

$$\begin{aligned}&\operatorname{margin}(\boldsymbol{X}_N; h) \quad \triangleq \\ &\min\{d((\boldsymbol{X}_N \cap K^-(h)), K^+(h)), d((\boldsymbol{X}_N \cap K^+(h)), K^-(h))\}.\end{aligned}$$

It is of interest to bound $\Pr\left(\sup_{h\in\mathcal{H}} \operatorname{margin}(\boldsymbol{X}_N; h) \geq \gamma\right)$. This bound, necessarily, must depend on the size of the space of classifiers, much like a bound on the empirical gap must depend on the dimension. This is an appropriate bound to consider when $\mathcal{H}$ consists of a set of, in some sense, equally reasonable classifiers. That is to say, if, for example, μ is a one-dimensional Gaussian distribution, then if $\mathcal{H}$ contains a (linear) classifier far out in the tail of the distribution, the uniform bound on the margin will be useless, as that classifier will essentially dominate the probability that there is a large margin.

If $\mathcal{H} = \{h_1, \ldots, h_m\}$, then by an appeal to the union bound, we have

$$\begin{aligned}\Pr\left(\max_{h\in\mathcal{H}} \operatorname{margin}(\boldsymbol{X}_N; h) \geq \gamma\right) \quad \leq \\ m \max_{h\in\mathcal{H}} \left\{\Pr\left(\operatorname{margin}(\boldsymbol{X}_N; h) \geq \gamma\right)\right\}.\end{aligned}$$

Using covering numbers we can extend the use of the union bound to infinite classifier families. We construct the ε-net using the metric structure of the sample space, i.e., using the same distance as that used to compute the size γ of the margin. We define for this purpose the distance φ between two classifiers $h', h'' \in \mathcal{H}$ as

$$\begin{aligned}&\varphi(h', h'') \quad \triangleq \\ &\max\left\{\sup_{x\in K^-(h')} \inf_{y\in K^-(h'')} d(x,y);\ \sup_{x\in K^+(h')} \inf_{y\in K^+(h'')} d(x,y)\right\}.\end{aligned}$$

Note that this definition is symmetric in h' and h''. Therefore, given any $\varepsilon > 0$, and $h \in \mathcal{H}$, the ε-ball about h is the set $B_{\varepsilon,\varphi}(h) \subseteq \mathcal{H}$ of classifiers $h' \in \mathcal{H}$, with $\varphi(h, h') \leq \varepsilon$. Therefore, as usual, an ε-cover $\mathcal{H}_\varepsilon$ of $\mathcal{H}$ is a collection $\{h_1, \ldots, h_{\mathcal{N}_\varepsilon}\} \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$, there exists some $h_j \in \mathcal{H}_\varepsilon$ with $\varphi(h, h_j) \leq \varepsilon$. Then, we have the following corollary of Proposition 3 and Corollary 6.

Corollary 7: Let $\mathcal{H}$ be a family of classifiers. For $\varepsilon > 0$, let $\mathcal{H}_\varepsilon$ be an ε-cover, with covering number $\mathcal{N}_{\varepsilon,\varphi}$. Then,

$$\begin{aligned}\Pr\left(\max_{h\in\mathcal{H}} \operatorname{margin}(\boldsymbol{X}_N; h) \geq \gamma\right) \quad \leq \\ \inf_{\varepsilon, \mathcal{H}_{\varepsilon,\varphi}} \left\{\mathcal{N}_{\varepsilon,\varphi} \max_{h\in\mathcal{H}_\varepsilon} \Pr\left(\operatorname{margin}(\boldsymbol{X}_N; h) \geq \gamma - \varepsilon\right)\right\}\end{aligned}$$

Thus for the best bound obtainable in this fashion, we must find the optimal tradeoff between the fineness of the covering, and the size of the resulting cover.

Computing ε-covers with this metric, and thus the subsequent optimization problem, can be readily done in a number of common cases. For example, this is the case if $\mathcal{H}$ is the set of linear classifiers through the origin, and K is compact, or if $\mathcal{H}$ is the set of classifiers parallel to a given hyperplane.

Acknowledgements

We would like to thank Ravi Kannan for helpful comments. We would also like to thank two anonymous reviewers for very careful comments and observations, and in particular, one reviewer found a simplified proof of Lemma 2, which we use.

REFERENCES

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, England, 2000.

[2] A.J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans, editors. *Advances in Large Margin Classifiers.* MIT Press,, 2000.

[3] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[4] R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.*, 39:355–405, 2002.

[5] L. Lovász and M. Simonovits. Mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proc. 31st Annual Symp. on Found. of Computer Science*, pages 346–355, 1990.

[6] D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *Proc. 23th ACM STOC*, pages 156–163, 1991.

[7] R. Kannan. Personal communication. 2004.

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

[9] R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, Boston, 2002.

[10] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, England, 2000.

[11] S. Ben-David, N. Eiron, and H.U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *Journal of Machine Learning Research*, 3:441–461, 2002.

[12] S. Mendelson. Lipschitz embeddings of function classes. Available from http://web.rsise.anu.edu.au/˜shahar/, 2004.

[13] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[14] V. N. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.

[15] M. Bagnoli and T. Bergstrom. Log–concave probability and its applications. Available from http://citeseer.nj.nec.com/bagnoli89logconcave.html, 1989.

[16] Henstock R and A. N. Macbeath. On the measure of sum sets, I. the theorems of Brunn, Minkowski and Lusternik. *Proc. London Math. Soc.*, 3:182–194, 1953.

[17] L. Lovász and S. Vempala. The geometry of logconcave functions and an $O^*(n^3)$ sampling algorithm. *Microsoft Research Tech. Rep. MSR-TR-2003-04*, 2003.

[18] S. Vempala. Geometric random walks: A survey. *to appear: MSRI Combinatorial and Computational Geometry*, 2004.

[19] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Computational Geometry*, 13:541–559, 1995.

[20] M. Fradelizi and O. Guédon. The extreme points of subsets of s-concave probabilities and a geometric localization theorem. *Discrete Comput. Geom.*, 31(2):327–335, 2004.

[21] M. Fradelizi and O. Guédon. A generalized localization theorem and geometric inqualities for convex bodies. Available from http://www.institut.math.jussieu.fr/˜guedon/Articles/04/FGfinalAIM.pdf, 2006.

[22] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer Verlag, New Jersey, 1993.

[23] J. Matoušek. *Using the Borsuk-Ulam Theorem*. Springer Verlag, Berlin, 2002.

[24] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized chebyshev bounds via semidefinite programming. *submitted.*, 2004.

[25] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal of Optimization*, 15(3):780–804, 2004.

[26] K. Bennett and E. Bredensteiner. Duality and geometry in SVM classifiers. In *Proc. 17th Int. Conf. on Machine Learning*, pages 57–64, 2000.

[27] J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *Computational Learing Theory*, pages 278–285, 1999.

Constantine Caramanis (S'05-M'06) received the A.B. degree in Mathematics from Harvard University, Cambridge, MA, in 1999, and the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, in 2001 and 2006, respectively. He is currently an Assistant Professor of Electrical and Computer Engineering at The University of Texas at Austin, where he is part of the Wireless Networking and Communications Group (WNCG).

His research interests include optimization and control under uncertainty, and adaptive optimization and control, learning theory, and applications to communications, networks, and scheduling.

Shie Mannor (S'00-M'03) received the B.Sc. degree in electrical engineering, the B.A. degree in mathematics (both summa cum laude), and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1996, 1996, and 2002, respectively. During the spring semester of 2002, he was a lecturer at the Technion Electrical Engineering Department. From 2002 to 2004, he was a postdoctoral associate with M.I.T. He is currently an Assistant Professor of Electrical and Computer Engineering at McGill University.

His research interests include machine learning and pattern recognition, planning and control, multi-agent systems, and communications. He was a Fulbright scholar in 2002, and he currently holds a Canada Research Chair in Machine Learning.