

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

Constantine Caramanis* and Shie Mannor

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139
{cmcaram,shie}@mit.edu

Abstract. We prove that given a nearly log-concave density, 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 also show that when the data are sampled from a nearly log-concave distribution, the margin cannot be large in a strong probabilistic sense. 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 $\epsilon_1 - \epsilon_0$.

1 Introduction

Large margin classifiers (e.g., [CS00,SBSS00] 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., [SS02] 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 2, and point out some interesting properties of β -log-concave functions.

* C. Caramanis is eligible for the Best student paper award.

In Section 3 we prove an inequality stating that the measure (under a β -log-concave distribution) 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 [Gar02]). We note that Theorem 2 was stated in [LS90] for volumes, and in [AK91] for β -log-concave distributions, in the context of efficient sampling from convex bodies. However, there are steps in the proof which we were unable to follow. Specifically, the reduction in [AK91] to what they call the “needle-like” case is based on an argument used in [LS90], which uses the Ham-Sandwich Theorem to guarantee not only bisection, but also some orthogonality properties of the bisecting hyperplane. It is not clear to us how one may obtain such guarantees from the Ham-Sandwich Theorem. Furthermore, the solution of the needle-like case in [AK91] relies on a uniformity assumption on the modulation of the distribution, which does not appear evident from the assumptions on the distribution. We provide a complete proof of the result using the Ham-Sandwich Theorem (as in [LS90]) and a different reduction argument. We further point out a few natural extensions.

In Section 4 we specialize the isoperimetric inequality to two different setups. First, we 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 learning formulation, 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. 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. Second, we consider the typical statistical machine learning setup, and show that for any classifier the probability of a large margin (with respect to that 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 a Mercer kernel, the induced distribution in the feature space may not be β -log-concave. If 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. If the kernel map is not Lipschitz, then our results do not directly apply.

In Section 5 we briefly touch on the issue of regression. We show that if we have a regressor, then the measure of a tube around its prediction with inner radius ϵ_0 and outer radius ϵ_1 is bounded from below by $\epsilon_1 - \epsilon_0$ times a constant (as long as ϵ_1 is not too large). The direct implication of this inequality is that the margins of the tube carry a significant portion of the measure.

Some recent results [BES02, Men04] 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., [AB99, Vap98]) 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.

2 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. While we are mostly interested in distributions, it is useful to write the following definitions in terms of a general function on \mathbb{R}^n .

Definition 1. *A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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:*

$$f(\lambda x_1 + (1 - \lambda)x_2) \geq e^{-\beta} f(x_1)^\lambda f(x_2)^{1-\lambda}. \quad (2.1)$$

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

The class of log-concave distributions itself is rather rich. For example, it includes Gaussian, Uniform, Logistic, and Exponential distributions. We refer the reader to [BB89] 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 Eq. (2.1). For example, unlike log-concave distributions, β -log-concave distributions need not be continuous. We now provide some results that are useful in the sequel. We start from the following observation.

Lemma 1. *The support of a β -log-concave function is a convex set. Also, β -log-concave functions are bounded on bounded sets.*

Distributions 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 divide our set into two sets, if the distance between the 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 $f(x)$ 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. Fix $\epsilon > 0$. There is some $x^* \in [u_1, u_2]$ such that $\sup_{x \in [u_1, u_2]} f(x) < f(x^*) + \epsilon$. Suppose $x^* \in [u_1, x_1]$. Then for any $x \in [x_1, x_2]$ and $y \in [x_2, u_2]$, and for some $\lambda \in (0, 1)$ we have $x = \lambda x^* + (1 - \lambda)y$, and by the β -log-concavity of f , we have

$$f(x) \geq f(x^*)^\lambda f(y)^{1-\lambda} e^{-\beta} \geq (f(y) - \epsilon)^\lambda f(y)^{1-\lambda} e^{-\beta}. \quad (2.2)$$

Similarly, if $x^* \in [x_2, u_2]$, then for every $x \in [x_1, x_2]$ and $y \in [u_1, x_1]$, Eq. (2.2) holds. Finally, if $x^* \in [x_1, x_2]$, then for any $x \in [x_1, x^*]$, Eq. (2.2) holds for $y \in [u_1, x_1]$, and for $x \in [x^*, x_2]$, Eq. (2.2) holds for any $y \in [x_2, u_2]$. Take a sequence $\epsilon_i \searrow 0$. We know that for every ϵ_i Eq. (2.2) holds for all $x \in [x_1, x_2]$ and all $y \in [u_1, x_1]$ or all $y \in [x_2, u_2]$. It follows that there exists a sequence $\epsilon_i \searrow 0$ such that for all $x \in [x_1, x_2]$, Eq. (2.2) holds for all $y \in [u_1, x_1]$ or for all $y \in [x_2, u_2]$. Since ϵ_i converges to 0, $f(x) \geq f(y)e^{-\beta}$ in at least one of those domains. \square

The following inequality has many uses in geometry, statistics, and analysis (see [Gar02]). 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. Recall that the orthogonal projection of a set $K \subseteq \mathbb{R}^{n+m}$ onto \mathbb{R}^n is defined as $K|_{\mathbb{R}^n} \triangleq \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m \text{ s.t. } (x, y) \in K\}$.

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

Proof. This is a consequence of the Prékopa-Leindler inequality as in [Gar02], Section 9, for log-concave distributions. Adapting the proof for β -log-concave distributions is straightforward. \square

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 distribution 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. These properties will be provided elsewhere.

3 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. 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 $d(K_1, K_2) \triangleq \inf_{x \in K_1, y \in K_2} d(x, y)$ and the diameter of a set K as $\text{diam}(K) \triangleq \sup_{x, y \in K} d(x, y)$. Given a distribution 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, \dots, K_ℓ 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 distribution $f(x)$, the induced measure μ satisfies that*

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

We remark that this bound is dimension-free. The ratio $d(K_1, K_2)/\text{diam}(K)$ 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 , with base case $n = 1$. To prove the inductive step, first we show that it is enough to consider an “ ϵ -flat” set K , i.e., a set that is contained in an ellipse whose smallest axis is smaller than some $\epsilon > 0$. Next, we show that for an ϵ -flat set K , we can project onto $n - 1$ dimensions where the theorem holds by induction. 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 .

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

Proof. If $n = 1$, then K is some interval, $K = [u_1, u_2]$, with $\text{diam}(K) = |u_2 - u_1|$. Since $t = d(K_1, K_2) > 0$, no point of K_1 is within a distance 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 maximizer¹ 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,

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

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

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

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

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

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 $\mu([r_1, r_2]) \geq e^{-\beta} (\hat{t} / \text{diam}(K)) \mu([r_2, u_2])$, then $\mu(\hat{B}) \geq$

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 exactly parallel to that given in Lemma 2.

$e^{-\beta}(\hat{t}/\text{diam}(K))\mu(\hat{K}_2)$ and we are done. So let us assume that $\mu([r_1, r_2]) \geq e^{-\beta}(\hat{t}/\text{diam}(K))\mu([u_1, r_1])$. Similarly, for the last \hat{B} -interval (r_{m-2}, r_{m-1}) , we can assume that $\mu([r_{m-2}, r_{m-1}]) \geq e^{-\beta}(\hat{t}/\text{diam}(K))\mu([r_{m-1}, u_2])$ 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 $\mu([r_{j-1}, r_j]) \geq e^{-\beta}(\hat{t}/\text{diam}(K))\mu([u_1, r_{j-1}])$ and $\mu([r_{j+1}, r_{j+2}]) \geq e^{-\beta}(\hat{t}/\text{diam}(K))\mu([r_{j+2}, u_2])$. 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

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

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

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. [GLS93]). 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.*

$$(1 + \delta)\mu(B) \leq e^{-\beta} \frac{t}{\text{diam}(K)} \min\{\mu(K_1), \mu(K_2)\}. \quad (3.3)$$

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 $\text{diam}(\tilde{K}) \leq d$, and such that the theorem fails by δ , i.e., Eq. (3.3) 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., [Mat02]) says that in \mathbb{R}^n , given n Borel measures $\mu_k, k = 1, \dots, n$, 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 theorem holds, or fails by less than δ , for both K' and K'' , then

$$\begin{aligned} (1 + \delta)\mu(B) &= (1 + \delta)\mu(B') + (1 + \delta)\mu(B'') \\ &\geq e^{-\beta} \frac{t}{\text{diam}(K)} \left(\min \left\{ \frac{1}{2}\mu(K'_1), \frac{1}{2}\mu(K'_2) \right\} + \min \left\{ \frac{1}{2}\mu(K''_1), \frac{1}{2}\mu(K''_2) \right\} \right) \\ &= e^{-\beta} \frac{t}{\text{diam}(K)} \min\{\mu(K_1), \mu(K_2)\}. \end{aligned}$$

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

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

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

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 $j \geq \log_2(2(\mu(K_1) + \mu(K_2))/\eta)$, $\mu(B^{(j)}) \geq \eta/2 \geq \min\{\mu(K_1^{(j)}), \mu(K_2^{(j)})\}$. 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. \square

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 $\delta > 0$:

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

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

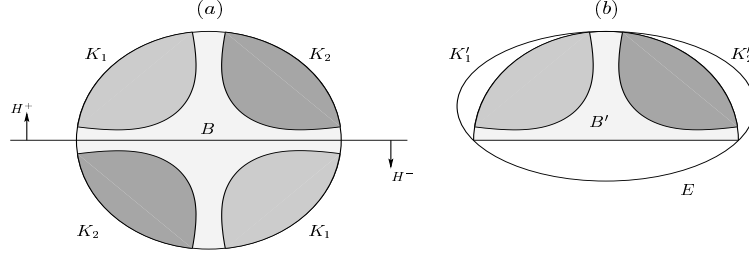


Fig. 1. The inductive step works by projecting K onto one less dimension. In (a) above, a projection on the horizontal axis would yield a distance of zero between the projected K_1 and K_2 . Once we bisect to obtain (b), we see that a projection onto the horizontal axis would not affect the minimum distance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

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

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

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, $\text{diam}(K_\pi) \leq \text{diam}(K)$.

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 function 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

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

and thus $(1+\delta)\mu(B) \geq (t/\text{diam}(K)) \min(\mu(K_1), \mu(K_2))$. Since this holds for every $\delta > 0$, the result follows. \square

Corollaries 1 and 2 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 distribution that is bounded away from zero on K , say $f(x) > \eta$ for $x \in K$. Then the induced measure μ satisfies*

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

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, $\nu(B) \geq (d(K_1, K_2)/\text{diam}(K)) \min\{\nu(K_1), \nu(K_2)\}$. The corollary follows since $\mu(B) \geq \eta\nu(B)$. \square

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 distribution with induced measure μ , such that there exists $d(\epsilon)$ 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 distribution $\hat{f}(x) = f(x)/P$. 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 $\mu(B)/(1-\epsilon) \geq \mu(B)/P \geq \hat{\mu}(B \cap B_{d(\epsilon)})$. The result follows by some algebra. \square

4 Lower Bounds for Classification and the Size of the Margin

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 first 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 $\{-1, 1\}$. Suppose that the performance of the classifier is measured using some β -log-concave distribution f (and associated measure μ). We note that this model deviates from the “classical” statistical machine learning setup. Given a distribution f , the disagreement of a classifier $h \in \mathcal{H}$ with another classifier h' is defined as:

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

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 $K^- \triangleq \{x \in K : h(x) = -1\}$. Given a pair of classifiers h_1 and h_2 we define the distance between them as

$$\text{dist}(h_1, h_2) \triangleq \max\{d(K^+(h_1), K^-(h_2)), d(K^-(h_1), K^+(h_2))\}.$$

We note that $\text{dist}(h_1, h_2)$ may equal zero even if the classifiers are rather different. However, in some cases, $\text{dist}(h_1, h_2)$ provides a useful measure of difference; see Proposition 1 below.

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 bounded set K . Then for every $h \in \mathcal{H}$ there exists $h' \in \mathcal{H}$ such that*

$$\Delta(h; h') \geq \frac{e^{-\beta} P_0}{2 \text{diam}(K)} \sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2),$$

where $P_0 = \inf_{\tilde{h} \in \mathcal{H}} \min\{\mu(K^+(\tilde{h})), \mu(K^-(\tilde{h}))\}$.

Proof. If $\sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2) = 0$, the result follows, so we can assume this is not the case. For every $\epsilon > 0$ we can choose $h_1 \in \mathcal{H}$ and $h_2 \in \mathcal{H}$ such that $\text{dist}(h_1, h_2) \geq \sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2) - \epsilon$. We consider the case where $\text{dist}(h_1, h_2) =$

$d(K^+(h_1), K^-(h_2))$; the other case where $d(K^-(h_1), K^+(h_2)) = \text{dist}(h_1, h_2)$ follows in a symmetric manner. Let $B = K \setminus (K^+(h_1) \cup K^-(h_2))$. It follows by Theorem 2 that

$$\mu(B) \geq e^{-\beta} \frac{\text{dist}(h_1, h_2)}{\text{diam}(K)} \min \{ \mu(K^+(h_1)), \mu(K^-(h_2)) \}. \quad (4.4)$$

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 $P_0 \leq \mu(K^+(h_1))$ and $P_0 \leq \mu(K^-(h_2))$, and by substituting in Eq. (4.4) we obtain that $\Delta(h, h_i) \geq e^{-\beta} \text{dist}(h_1, h_2) P_0 / (2 \text{diam}(K))$ for $i = 1$ or $i = 2$. The result follows by taking ϵ to 0. \square

The following example demonstrates the power of Theorem 3 in the context of linear classification. Consider an input-output sequence $\{(x_1, y_1), \dots, (x_N, y_N)\}$ 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 $h(x) = \text{sign}(\langle x, u \rangle + b)$, where ‘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., 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 1. *Suppose that we are given two sets of linearly separable vectors X^+ and X^- and let $t = d(\text{conv}(X^+), \text{conv}(X^-))$. Then for every linear classifier h that separates X^+ and X^- , and any β -log-concave distribution f and induced measure μ defined on a bounded set K , there exists another linear classifier h' that separates the X^+ and X^- as well, such that $\Delta(h; h') \geq e^{-\beta} P_0 t / (2 \text{diam}(K))$, where $P_0 = \min \{ \mu(\{x : \langle x, u \rangle \geq \langle x^+, u \rangle\}), \mu(\{x : \langle x, u \rangle \leq \langle x^-, u \rangle\}) \}$ for some $x^\pm \in \text{conv}(X^\pm)$ 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 [BB00]) that $\sup_{h_1, h_2 \in \mathcal{H}} \text{dist}(h_1, h_2) = t$. This is attained for $h_1(x) = \text{sign}(\langle x, u \rangle - \langle x^+, u \rangle)$ and $h_2(x) = \text{sign}(\langle x, u \rangle - \langle x^-, u \rangle)$. We now apply Theorem 3 to obtain the desired result. Note that P_0 in the declaration of the proposition is tighter than P_0 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). \square

We now 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 $h(x) = \text{sign}(g(x))$ 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.

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

Proposition 2. *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 distribution on K with induced measure μ . Then*

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

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 $\mu(B) \geq \gamma e^{-\beta} \min\{\mu(K_1), \mu(K_2)\} / \text{diam}(K)$. We also know that $\mu(B) = \mu(K^-(h)) - \mu(K_2)$. So that

$$\mu(B) \geq \max\{\gamma e^{-\beta} \min\{\mu(K_1), s\} / \text{diam}(K), \mu(K^-(h)) - s\}, \quad (4.5)$$

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 $\mu(K^-(h)) - s = \gamma e^{-\beta} \mu(K_1) / \text{diam}(K)$ or at the point where $\mu(K^-(h)) - s = s \gamma e^{-\beta} / \text{diam}(K)$. Substituting those s in Eq. (4.5) and some algebra gives the desired result. \square

A similar result holds by interchanging K^+ and K^- throughout Proposition 2. The following corollary is an immediate application of the above.

Corollary 3. *Suppose that N samples $\mathbf{X}_N = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ are drawn independently from a β -log-concave distribution f defined on a bounded set K . Let h be a classifier. Then for every $\gamma > 0$:*

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

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

Proof. The proof follows from Proposition 2 and the inequality $(1 - a)^N \leq \exp(-aN)$ for $a \in [0, 1]$ and $N \geq 0$. \square

Corollary 3 is a dimension-free inequality. It implies that when sampling from a β -log-concave distribution, 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 $\mathbf{X}_N = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ one can consider the probabilistic behavior of the following empirical gap between the classes: $\text{gap}(\mathbf{X}_N; h) = \min_{i,j:h(\mathbf{x}_i) \neq h(\mathbf{x}_j)} d(\mathbf{x}_i, \mathbf{x}_j)$. 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 that, 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 $\text{gap}(\mathbf{X}_N; h)$ 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 $\text{gap}(\mathbf{X}_N; h)$ is large using covering number arguments, as in [SC99], but such a bound must be dimension-dependent.

We finally note that a uniform bound in the spirit of Corollary 3 is of interest. Specifically, let the empirical margin of a classifier h on sample points \mathbf{X}_N be denoted by:

$$\text{margin}(\mathbf{X}_N; h) \triangleq \min\{d((\mathbf{X}_N \cap K^-(h)), K^+(h)), d((\mathbf{X}_N \cap K^+(h)), K^-(h))\}.$$

It is of interest to bound $\Pr(\sup_{h \in \mathcal{H}} \text{margin}(\mathbf{X}_N; h) \geq \gamma)$. We leave the issue of efficiently bounding the empirical margin to future research.

5 Regression Tubes

Consider a function k from \mathbb{R}^n to \mathbb{R}^m . In this section we provide a result of a different flavor that concerns the weight of tubes around k . The probabilistic setup is as follows. We have a probability measure f on \mathbb{R}^{n+m} that prescribes the probability of getting a pair $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$. For a function $k : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we consider the set

$$T_{\epsilon_0, \epsilon_1}(k) \triangleq \{(x, y) : \epsilon_0 \leq \|k(x) - y\| \leq \epsilon_1\}.$$

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 .

Corollary 4. *Suppose that f 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 . Then for every $\epsilon_1 > \epsilon_0 > 0$*

$$\mu(T_{\epsilon_0, \epsilon_1}(k)) \geq \frac{(\epsilon_1 - \epsilon_0)e^{-\beta}}{L \text{diam}(K)} \min\{\mu(T_{0, \epsilon_0}(k)), \mu(T_{\epsilon_1, \text{diam}(K)}(k))\}.$$

Proof. We use Theorem 2 with the decomposition $K_1 = T_{0,\epsilon_0}(k)$, $B = T_{\epsilon_0,\epsilon_1}(k)$ and $K_2 = T_{\epsilon_1,\text{diam}(K)}(k)$. Note that $d(T_{0,\epsilon_0}, T_{\epsilon_1,\text{diam}(K)}) \geq (\epsilon_1 - \epsilon_0)/L$, since k is Lipschitz with constant L . \square

A result where f is conditionally β -log-concave (i.e., given that x was sampled, the conditional probability of y is β -log-concave) is desirable. This requires some additional continuity assumptions on f , and is left for future research.

Acknowledgements

We thank three anonymous reviewers for thoughtful and detailed comments. Shie Mannor was partially supported by the National Science Foundation under grant ECS-0312921.

References

- [AB99] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- [AK91] D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *Proc. 23th ACM STOC*, pages 156–163, 1991.
- [BB89] M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. Available from citeseer.nj.nec.com/bagnoli89logconcave.html, 1989.
- [BB00] K. Bennett and E. Bredeñsteiner. Duality and geometry in SVM classifiers. In *Proc. 17th Int. Conf. on Machine Learning*, pages 57–64, 2000.
- [BES02] 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.
- [CS00] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, England, 2000.
- [Gar02] R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.*, 39:355–405, 2002.
- [GLS93] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer Verlag, New Jersey, 1993.
- [LS90] 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.
- [Mat02] J. Matoušek. *Using the Borsuk-Ulam Theorem*. Springer Verlag, Berlin, 2002.
- [Men04] S. Mendelson. Lipschitz embeddings of function classes. Available from <http://web.rsise.anu.edu.au/~shahar/>, 2004.
- [SBSS00] A.J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans, editors. *Advances in Large Margin Classifiers*. MIT Press, 2000.
- [SC99] J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *Computational Learning Theory*, pages 278–285, 1999.
- [SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- [Vap98] V. N. Vapnik. *Statistical Learning Theory*. Wiley Interscience, New York, 1998.