Tight Sample Complexity of Large-Margin Learning

Sivan Sabato[1] **Nathan Srebro**[2] **Naftali Tishby**[1]
[1] School of Computer Science & Engineering, The Hebrew University, Jerusalem 91904, Israel
[2] Toyota Technological Institute at Chicago, Chicago, IL 60637, USA
{sivan_sabato,tishby}@cs.huji.ac.il, nati@ttic.edu

Abstract

We obtain a tight distribution-specific characterization of the sample complexity of large-margin classification with L_2 regularization: We introduce the γ-adapted-dimension, which is a simple function of the spectrum of a distribution's covariance matrix, and show distribution-specific upper *and* lower bounds on the sample complexity, both governed by the γ-adapted-dimension of the source distribution. We conclude that this new quantity tightly characterizes the true sample complexity of large-margin classification. The bounds hold for a rich family of sub-Gaussian distributions.

1 Introduction

In this paper we tackle the problem of obtaining a tight characterization of the sample complexity which a particular learning rule requires, in order to learn a particular source distribution. Specifically, we obtain a tight characterization of the sample complexity required for large (Euclidean) margin learning to obtain low error for a distribution $D(X, Y)$, for $X \in \mathbb{R}^d$, $Y \in \{\pm 1\}$.

Most learning theory work focuses on upper-bounding the sample complexity. That is, on providing a bound $\overline{m}(D, \epsilon)$ and proving that when using some specific learning rule, if the sample size is at least $\overline{m}(D, \epsilon)$, an excess error of at most ϵ (in expectation or with high probability) can be ensured. For instance, for large-margin classification we know that if $P_D[\|X\| \leq B] = 1$, then $\overline{m}(D, \epsilon)$ can be set to $O(B^2/(\gamma^2\epsilon^2))$ to get true error of no more than $\ell^*_\gamma + \epsilon$, where $\ell^*_\gamma = \min_{\|w\|\leq 1} P_D(Y\langle w, X\rangle \leq \gamma)$ is the optimal margin error at margin γ.

Such upper bounds can be useful for understanding positive aspects of a learning rule. But it is difficult to understand deficiencies of a learning rule, or to compare between different rules, based on upper bounds alone. After all, it is possible, and often the case, that the true sample complexity, i.e. the actual number of samples required to get low error, is much lower than the bound.

Of course, some sample complexity upper bounds are known to be "tight" or to have an almost-matching lower bound. This usually means that the bound is tight as a worst-case upper bound for a specific class of distributions (e.g. all those with $P_D[\|X\| \leq B] = 1$). That is, there exists *some* source distribution for which the bound is tight. In other words, the bound concerns some quantity of the distribution (e.g. the radius of the support), and is the lowest possible bound *in terms of this quantity*. But this is not to say that for any *specific* distribution this quantity tightly characterizes the sample complexity. For instance, we know that the sample complexity can be much smaller than the radius of the support of X, if the average norm $\sqrt{\mathbb{E}[\|X\|^2]}$ is small. However, $\mathbb{E}[\|X\|^2]$ is also not a precise characterization of the sample complexity, for instance in low dimensions.

The goal of this paper is to identify a simple quantity determined by the distribution that *does* precisely characterize the sample complexity. That is, such that the actual sample complexity for the learning rule on this *specific* distribution is governed, up to polylogarithmic factors, by this quantity.

In particular, we present the γ-adapted-dimension $k_\gamma(D)$. This measure refines both the dimension and the average norm of X, and it can be easily calculated from the covariance matrix of X. We show that for a rich family of "light tailed" distributions (specifically, sub-Gaussian distributions with independent uncorrelated directions – see Section 2), the number of samples required for learning by minimizing the γ-margin-violations is both lower-bounded and upper-bounded by $\tilde{\Theta}(k_\gamma)$. More precisely, we show that the sample complexity $m(\epsilon, \gamma, D)$ required for achieving excess error of no more than ϵ can be bounded from above and from below by:

$$\Omega(k_\gamma(D)) \leq m(\epsilon, \gamma, D) \leq \tilde{O}(\frac{k_\gamma(D)}{\epsilon^2}).$$

As can be seen in this bound, we are *not* concerned about tightly characterizing the dependence of the sample complexity on the desired error [as done e.g. in 1], nor with obtaining tight bounds for very small error levels. In fact, our results can be interpreted as studying the sample complexity needed to obtain error well below random, but bounded away from zero. This is in contrast to classical statistics asymptotic that are also typically tight, but are valid only for very small ϵ. As was recently shown by Liang and Srebro [2], the quantities on which the sample complexity depends on for very small ϵ (in the classical statistics asymptotic regime) can be very different from those for moderate error rates, which are more relevant for machine learning.

Our tight characterization, and in particular the distribution-specific lower bound on the sample complexity that we establish, can be used to compare large-margin (L_2 regularized) learning to other learning rules. In Section 7 we provide two such examples: we use our lower bound to rigorously establish a sample complexity gap between L_1 and L_2 regularization previously studied in [3], and to show a large gap between discriminative and generative learning on a Gaussian-mixture distribution.

In this paper we focus only on large L_2 margin classification. But in order to obtain the distribution-specific lower bound, we develop novel tools that we believe can be useful for obtaining lower bounds also for other learning rules.

Related work

Most work on "sample complexity lower bounds" is directed at proving that under some set of assumptions, there exists a source distribution for which one needs at least a certain number of examples to learn with required error and confidence [4, 5, 6]. This type of a lower bound does not, however, indicate much on the sample complexity of other distributions under the same set of assumptions.

As for distribution-specific lower bounds, the classical analysis of Vapnik [7, Theorem 16.6] provides not only sufficient but also necessary conditions for the learnability of a hypothesis class with respect to a specific distribution. The essential condition is that the ϵ-entropy of the hypothesis class with respect to the distribution be sub-linear in the limit of an infinite sample size. In some sense, this criterion can be seen as providing a "lower bound" on learnability for a specific distribution. However, we are interested in finite-sample convergence rates, and would like those to depend on simple properties of the distribution. The asymptotic arguments involved in Vapnik's general learnability claim do not lend themselves easily to such analysis.

Benedek and Itai [8] show that if the distribution is known to the learner, a specific hypothesis class is learnable if and only if there is a finite ϵ-cover of this hypothesis class with respect to the distribution. Ben-David et al. [9] consider a similar setting, and prove sample complexity lower bounds for learning with any data distribution, for some binary hypothesis classes on the real line. In both of these works, the lower bounds hold for any algorithm, but only for a worst-case target hypothesis. Vayatis and Azencott [10] provide distribution-specific sample complexity upper bounds for hypothesis classes with a limited VC-dimension, as a function of how balanced the hypotheses are with respect to the considered distributions. These bounds are not tight for all distributions, thus this work also does not provide true distribution-specific sample complexity.

2 Problem setting and definitions

Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. D_X will denote the restriction of D to $\mathbb{R}^d$. We are interested in linear separators, parametrized by unit-norm vectors in $\mathbf{B}_1^d \triangleq \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}$.

For a predictor w denote its misclassification error with respect to distribution D by $\ell(w, D) \triangleq \mathbb{P}_{(X,Y)\sim D}[Y\langle w, X\rangle \leq 0]$. For $\gamma > 0$, denote the γ-margin loss of w with respect to D by $\ell_\gamma(w, D) \triangleq \mathbb{P}_{(X,Y)\sim D}[Y\langle w, X\rangle \leq \gamma]$. The minimal margin loss with respect to D is denoted by $\ell^*_\gamma(D) \triangleq \min_{w\in\mathbf{B}^d_1} \ell_\gamma(w, D)$. For a sample $S = \{(x_i, y_i)\}^m_{i=1}$ such that $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, the margin loss with respect to S is denoted by $\hat{\ell}_\gamma(w, S) \triangleq \frac{1}{m}|\{i \mid y_i\langle x_i, w\rangle \leq \gamma\}|$ and the misclassification error is $\hat{\ell}(w, S) \triangleq \frac{1}{m}|\{i \mid y_i\langle x_i, w\rangle \leq 0\}|$. In this paper we are concerned with learning by minimizing the margin loss. It will be convenient for us to discuss transductive learning algorithms. Since many predictors minimize the margin loss, we define:

Definition 2.1. *A* **margin-error minimization algorithm** $\mathcal{A}$ *is an algorithm whose input is a margin γ, a training sample $S = \{(x_i, y_i)\}^m_{i=1}$ and an unlabeled test sample $\tilde{S}_X = \{\tilde{x}_i\}^m_{i=1}$, which outputs a predictor $\tilde{w} \in \operatorname{argmin}_{w\in\mathbf{B}^d_1} \hat{\ell}_\gamma(w, S)$. We denote the output of the algorithm by $\tilde{w} = \mathcal{A}_\gamma(S, \tilde{S}_X)$.*

We will be concerned with the expected test loss of the algorithm given a random training sample and a random test sample, each of size m, and define $\ell_m(\mathcal{A}_\gamma, D) \triangleq \mathbb{E}_{S,\tilde{S}\sim D^m}[\hat{\ell}(\mathcal{A}(S, \tilde{S}_X), \tilde{S})]$, where $S, \tilde{S} \sim D^m$ independently. For $\gamma > 0$, $\epsilon \in [0, 1]$, and a distribution D, we denote the **distribution-specific sample complexity** by $m(\epsilon, \gamma, D)$: this is the minimal sample size such that for any margin-error minimization algorithm $\mathcal{A}$, and for any $m \geq m(\epsilon, \gamma, D)$, $\ell_m(\mathcal{A}_\gamma, D) - \ell^*_\gamma(D) \leq \epsilon$.

Sub-Gaussian distributions

We will characterize the distribution-specific sample complexity in terms of the covariance of $X \sim D_X$. But in order to do so, we must assume that X is not too heavy-tailed. Otherwise, X can have even infinite covariance but still be learnable, for instance if it has a tiny probability of having an exponentially large norm. We will thus restrict ourselves to sub-Gaussian distributions. This ensures light tails in all directions, while allowing a sufficiently rich family of distributions, as we presently see. We also require a more restrictive condition – namely that D_X can be rotated to a product distribution over the axes of $\mathbb{R}^d$. A distribution can always be rotated so that its coordinates are *uncorrelated*. Here we further require that they are *independent*, as of course holds for any multivariate Gaussian distribution.

Definition 2.2 (See e.g. [11, 12])**.** *A random variable X is* **sub-Gaussian with moment** B *(or B-sub-Gaussian) for $B \geq 0$ if*

$$\forall t \in \mathbb{R}, \quad \mathbb{E}[\exp(tX)] \leq \exp(B^2t^2/2). \tag{1}$$

We further say that X is sub-Gaussian with **relative moment** $\rho = B/\sqrt{\mathbb{E}[X^2]}$.

The sub-Gaussian family is quite extensive: For instance, any bounded, Gaussian, or Gaussian-mixture random variable with mean zero is included in this family.

Definition 2.3. *A distribution D_X over $X \in \mathbb{R}^d$ is* **independently sub-Gaussian** *with relative moment ρ if there exists some orthonormal basis $a_1, \ldots, a_d \in \mathbb{R}^d$, such that $\langle X, a_i\rangle$ are independent sub-Gaussian random variables, each with a relative moment ρ.*

We will focus on the family $\mathcal{D}^{\text{sg}}_\rho$ of all independently ρ-sub-Gaussian distributions in arbitrary dimension, for a small fixed constant ρ. For instance, the family $\mathcal{D}^{\text{sg}}_{3/2}$ includes all Gaussian distributions, all distributions which are uniform over a (hyper)box, and all multi-Bernoulli distributions, in addition to other less structured distributions. Our upper bounds and lower bounds will be tight up to quantities which depend on ρ, which we will regard as a constant, but the tightness will not depend on the dimensionality of the space or the variance of the distribution.

3 The γ-adapted-dimension

As mentioned in the introduction, the sample complexity of margin-error minimization can be upper-bounded in terms of the average norm $\mathbb{E}[\|X\|^2]$ by $m(\epsilon, \gamma, D) \leq O(\mathbb{E}[\|X\|^2]/(\gamma^2\epsilon^2))$ [13]. Alternatively, we can rely only on the dimensionality and conclude $m(\epsilon, \gamma, D) \leq \tilde{O}(d/\epsilon^2)$ [7]. Thus,

although both of these bounds are tight in the worst-case sense, i.e. they are the best bounds that rely only on the norm or only on the dimensionality respectively, neither is tight in a distribution-specific sense: If the average norm is unbounded while the dimensionality is small, an arbitrarily large gap is created between the true $m(\epsilon, \gamma, D)$ and the average-norm upper bound. The converse happens if the dimensionality is arbitrarily high while the average-norm is bounded.

Seeking a distribution-specific tight analysis, one simple option to try to tighten these bounds is to consider their minimum, $\min(d, \mathbb{E}[\|X\|^2]/\gamma^2)/\epsilon^2$, which, trivially, is also an upper bound on the sample complexity. However, this simple combination is also not tight: Consider a distribution in which there are a few directions with very high variance, but the combined variance in all other directions is small. We will show that in such situations the sample complexity is characterized not by the minimum of dimension and norm, but by the sum of the number of high-variance dimensions and the average norm in the other directions. This behavior is captured by the γ-*adapted-dimension*:

Definition 3.1. *Let $b > 0$ and k a positive integer.*

*(a). A subset $\mathcal{X} \subseteq \mathbb{R}^d$ is (b, k)-***limited** *if there exists a sub-space $V \subseteq \mathbb{R}^d$ of dimension $d - k$ such that $\mathcal{X} \subseteq \{x \in \mathbb{R}^d \mid \|x'P\|^2 \leq b\}$, where P is an orthogonal projection onto V.*

*(b). A distribution D_X over $\mathbb{R}^d$ is (b, k)-***limited** *if there exists a sub-space $V \subseteq \mathbb{R}^d$ of dimension $d - k$ such that $\mathbb{E}_{X \sim D_X}[\|X'P\|^2] \leq b$, with P an orthogonal projection onto V.*

Definition 3.2. *The γ-***adapted-dimension** *of a distribution or a set, denoted by k_γ, is the minimum k such that the distribution or set is $(\gamma^2 k, k)$ limited.*

It is easy to see that $k_\gamma(D_X)$ is upper-bounded by $\min(d, \mathbb{E}[\|X\|^2]/\gamma^2)$. Moreover, it can be much smaller. For example, for $X \in \mathbb{R}^{1001}$ with independent coordinates such that the variance of the first coordinate is 1000, but the variance in each remaining coordinate is 0.001 we have $k_1 = 1$ but $d = \mathbb{E}[\|X\|^2] = 1001$. More generally, if $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_d$ are the eigenvalues of the covariance matrix of X, then $k_\gamma = \min\{k \mid \sum_{i=k+1}^{d} \lambda_i \leq \gamma^2 k\}$. A quantity similar to k_γ was studied previously in [14]. k_γ is different in nature from some other quantities used for providing sample complexity bounds in terms of eigenvalues, as in [15], since it is defined based on the eigenvalues of the distribution and not of the sample. In Section 6 we will see that these can be quite different.

In order to relate our upper and lower bounds, it will be useful to relate the γ-adapted-dimension for different margins. The relationship is established in the following Lemma , proved in the appendix:

Lemma 3.3. *For $0 < \alpha < 1$, $\gamma > 0$ and a distribution D_X, $k_\gamma(D_X) \leq k_{\alpha\gamma}(D_X) \leq \frac{2k_\gamma(D_X)}{\alpha^2} + 1$.*

We proceed to provide a sample complexity upper bound based on the γ-adapted-dimension.

4 A sample complexity upper bound using γ-adapted-dimension

In order to establish an upper bound on the sample complexity, we will bound the fat-shattering dimension of the linear functions over a set in terms of the γ-adapted-dimension of the set. Recall that the fat-shattering dimension is a classic quantity for proving sample complexity upper bounds:

Definition 4.1. *Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \mathbb{R}$, and let $\gamma > 0$. The set $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ is γ-**shattered** by $\mathcal{F}$ if there exist $r_1, \ldots, r_m \in \mathbb{R}$ such that for all $y \in \{\pm 1\}^m$ there is an $f \in \mathcal{F}$ such that $\forall i \in [m],\ y_i(f(x_i) - r_i) \geq \gamma$. The γ-**fat-shattering dimension** of $\mathcal{F}$ is the size of the largest set in $\mathcal{X}$ that is γ-shattered by $\mathcal{F}$.*

The sample complexity of γ-loss minimization is bounded by $\tilde{O}(d_{\gamma/8}/\epsilon^2)$ were $d_{\gamma/8}$ is the $\gamma/8$-fat-shattering dimension of the function class [16, Theorem 13.4]. Let $\mathcal{W}(\mathcal{X})$ be the class of linear functions restricted to the domain $\mathcal{X}$. For any set we show:

Theorem 4.2. *If a set $\mathcal{X}$ is (B^2, k)-limited, then the γ-fat-shattering dimension of $\mathcal{W}(\mathcal{X})$ is at most $\frac{3}{2}(B^2/\gamma^2 + k + 1)$. Consequently, it is also at most $3k_\gamma(\mathcal{X}) + 1$.*

Proof. Let X be a $m \times d$ matrix whose rows are a set of m points in $\mathbb{R}^d$ which is γ-shattered. For any $\epsilon > 0$ we can augment X with an additional column to form the matrix $\tilde{X}$ of dimensions $m \times (d+1)$, such that for all $y \in \{-\gamma, +\gamma\}^m$, there is a $w_y \in \mathbf{B}_{1+\epsilon}^{d+1}$ such that $\tilde{X} w_y = y$ (the details

can be found in the appendix). Since $\mathcal{X}$ is (B^2, k)-limited, there is an orthogonal projection matrix $\tilde{P}$ of size $(d+1) \times (d+1)$ such that $\forall i \in [m], \|\tilde{X}_i' P\|^2 \leq B^2$ where $\tilde{X}_i$ is the vector in row i of $\tilde{X}$. Let $\tilde{V}$ be the sub-space of dimension $d-k$ spanned by the columns of $\tilde{P}$. To bound the size of the shattered set, we show that the projected rows of $\tilde{X}$ on V are 'shattered' using projected labels. We then proceed similarly to the proof of the norm-only fat-shattering bound [17].

We have $\tilde{X} = \tilde{X}\tilde{P} + \tilde{X}(I - \tilde{P})$. In addition, $\tilde{X}w_y = y$. Thus $y - \tilde{X}\tilde{P}w_y = \tilde{X}(I-\tilde{P})w_y$. $I - \tilde{P}$ is a projection onto a $k+1$-dimensional space, thus the rank of $\tilde{X}(I-\tilde{P})$ is at most $k+1$. Let T be an $m \times m$ orthogonal projection matrix onto the subspace orthogonal to the columns of $\tilde{X}(I-\tilde{P})$. This sub-space is of dimension at most $l = m - (k+1)$, thus trace$(T) = l$. $T(y - \tilde{X}\tilde{P}w_y) = T\tilde{X}(I-\tilde{P})w_y = 0_{(d+1)\times 1}$. Thus $Ty = T\tilde{X}\tilde{P}w_y$ for every $y \in \{-\gamma, +\gamma\}^m$.

Denote row i of T by t_i and row i of $T\tilde{X}\tilde{P}$ by z_i. We have $\forall i \leq m$, $\langle z_i, w_y^1 \rangle = t_i y = \sum_{j \leq m} t_i[j] y[j]$. Therefore $\langle \sum_i z_i y[i], w_y^1 \rangle = \sum_{i \leq m} \sum_{j \leq (l+k)} t_i[j] y[i] y[j]$. Since $\|w_y^1\| \leq 1 + \epsilon$, $\forall x \in \mathbb{R}^{d+1}, (1+\epsilon)\|x\| \geq \|x\|\|w_y^1\| \geq \langle x, w_y^1 \rangle$. Thus $\forall y \in \{-\gamma, +\gamma\}^m$, $(1+\epsilon)\|\sum_i z_i y[i]\| \geq \sum_{i \leq m} \sum_{j \leq m} t_i[j] y[i] y[j]$. Taking the expectation of y chosen uniformly at random, we have

$$(1+\epsilon)\mathbb{E}[\|\sum_i z_i y[i]\|] \geq \sum_{i,j} \mathbb{E}[t_i[j] y[i] y[j]] = \gamma^2 \sum_i t_i[i] = \gamma^2 \text{trace}(T) = \gamma^2 l.$$

In addition, $\frac{1}{\gamma^2}\mathbb{E}[\|\sum_i z_i y[i]\|^2] = \sum_{i=1}^l \|z_i\|^2 = \text{trace}(\tilde{P}'\tilde{X}'T^2\tilde{X}\tilde{P}) \leq \text{trace}(\tilde{P}'\tilde{X}'\tilde{X}\tilde{P}) \leq B^2 m$. From the inequality $E[X^2] \leq \mathbb{E}[X]^2$, it follows that $l^2 \leq (1+\epsilon)^2 \frac{B^2}{\gamma^2} m$. Since this holds for any $\epsilon > 0$, we can set $\epsilon = 0$ and solve for m. Thus $m \leq (k+1) + \frac{B^2}{2\gamma^2} + \sqrt{\frac{B^4}{4\gamma^4} + \frac{B^2}{\gamma^2}(k+1)} \leq (k+1) + \frac{B^2}{\gamma^2} + \sqrt{\frac{B^2}{\gamma^2}(k+1)} \leq \frac{3}{2}(\frac{B^2}{\gamma^2} + k + 1)$. □

Corollary 4.3. *Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$, $\mathcal{X} \subseteq \mathbb{R}^d$. Then*

$$m(\epsilon, \gamma, D) \leq \widetilde{O}\left(\frac{k_{\gamma/8}(\mathcal{X})}{\epsilon^2}\right).$$

The corollary above holds only for distributions with bounded support. However, since sub-Gaussian variables have an exponentially decaying tail, we can use this corollary to provide a bound for independently sub-Gaussian distributions as well (see appendix for proof):

Theorem 4.4 (Upper Bound for Distributions in $\mathcal{D}_\rho^{\text{sg}}$)**.** *For any distribution D over $\mathbb{R}^d \times \{\pm 1\}$ such that $D_X \in \mathcal{D}_\rho^{sg}$,*

$$m(\epsilon, \gamma, D) = \tilde{O}(\frac{\rho^2 k_\gamma(D_X)}{\epsilon^2}).$$

This new upper bound is tighter than norm-only and dimension-only upper bounds. But does the γ-adapted-dimension characterize the true sample complexity of the distribution, or is it just another upper bound? To answer this question, we need to be able to derive sample complexity lower bounds as well. We consider this problem in following section.

5 Sample complexity lower bounds using Gram-matrix eigenvalues

We wish to find a distribution-specific lower bound that depends on the γ-adapted-dimension, and matches our upper bound as closely as possible. To do that, we will link the ability to learn with a margin, with properties of the data distribution. The ability to learn is closely related to the probability of a sample to be shattered, as evident from Vapnik's formulations of learnability as a function of the ϵ-entropy. In the preceding section we used the fact that non-shattering (as captured by the fat-shattering dimension) implies learnability. For the lower bound we use the converse fact, presented below in Theorem 5.1: If a sample can be fat-shattered with a reasonably high probability, then learning is impossible. We then relate the fat-shattering of a sample to the minimal eigenvalue of its Gram matrix. This allows us to present a lower-bound on the sample complexity using a lower bound on the smallest eigenvalue of the Gram-matrix of a sample drawn from the data distribution. We use the term 'γ-shattered at the origin' to indicate that a set is γ-shattered by setting the bias $r \in \mathbb{R}^m$ (see Def. 4.1) to the zero vector.

Theorem 5.1. *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. If the probability of a sample of size m drawn from D_X^m to be γ-shattered at the origin is at least η, then there is a margin-error minimization algorithm $\mathcal{A}$, such that $\ell_{m/2}(\mathcal{A}_\gamma, D) \geq \eta/2$.*

Proof. For a given distribution D, let $\mathcal{A}$ be an algorithm which, for every two input samples S and $\tilde{S}_X$, labels $\tilde{S}_X$ using the separator $w \in \operatorname{argmin}_{w \in \mathbf{B}_1^d} \hat{\ell}_\gamma(w, S)$ that maximizes $\mathbb{E}_{\tilde{S}_Y \in D_Y^m}[\hat{\ell}_\gamma(w, \tilde{S})]$. For every $x \in \mathbb{R}^d$ there is a label $y \in \{\pm 1\}$ such that $\mathbb{P}_{(X,Y)\sim D}[Y \neq y \mid X = x] \geq \frac{1}{2}$. If the set of examples in S_X and $\tilde{S}_X$ together is γ-shattered at the origin, then $\mathcal{A}$ chooses a separator with zero margin loss on S, but loss of at least $\frac{1}{2}$ on $\tilde{S}$. Therefore $\ell_{m/2}(\mathcal{A}_\gamma, D) \geq \eta/2$. □

The notion of shattering involves checking the existence of a unit-norm separator w for each label-vector $y \in \{\pm 1\}^m$. In general, there is no closed form for the minimum-norm separator. However, the following Theorem provides an equivalent and simple characterization for fat-shattering:

Theorem 5.2. *Let $S = (X_1, \ldots, X_m)$ be a sample in $\mathbb{R}^d$, denote X the $m \times d$ matrix whose rows are the elements of S. Then S is 1-shattered iff X is invertible and $\forall y \in \{\pm 1\}^m, \quad y'(XX')^{-1}y \leq 1$.*

The proof of this theorem is in the appendix. The main issue in the proof is showing that if a set is shattered, it is also shattered with exact margins, since the set of exact margins $\{\pm 1\}^m$ lies in the convex hull of any set of non-exact margins that correspond to all the possible labelings. We can now use the minimum eigenvalue of the Gram matrix to obtain a sufficient condition for fat-shattering, after which we present the theorem linking eigenvalues and learnability. For a matrix X, $\lambda_n(X)$ denotes the n'th largest eigenvalue of X.

Lemma 5.3. *Let $S = (X_1, \ldots, X_m)$ be a sample in $\mathbb{R}^d$, with X as above. If $\lambda_m(XX') \geq m$ then S is 1-shattered at the origin.*

Proof. If $\lambda_m(XX') \geq m$ then XX' is invertible and $\lambda_1((XX')^{-1}) \leq 1/m$. For any $y \in \{\pm 1\}^m$ we have $\|y\| = \sqrt{m}$ and $y'(XX')^{-1}y \leq \|y\|^2 \lambda_1((XX')^{-1}) \leq m(1/m) = 1$. By Theorem 5.2 the sample is 1-shattered at the origin. □

Theorem 5.4. *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, S be an i.i.d. sample of size m drawn from D, and denote X_S the $m \times d$ matrix whose rows are the points from S. If $\mathbb{P}[\lambda_m(X_S X_S') \geq m\gamma^2] \geq \eta$, then there exists a margin-error minimization algorithm $\mathcal{A}$ such that $\ell_{m/2}(\mathcal{A}_\gamma, D) \geq \eta/2$.*

Theorem 5.4 follows by scaling X_S by γ, applying Lemma 5.3 to establish γ-fat shattering with probability at least η, then applying Theorem 5.1. Lemma 5.3 generalizes the requirement for linear independence when shattering using hyperplanes with no margin (i.e. no regularization). For unregularized (homogeneous) linear separation, a sample is shattered iff it is linearly independent, i.e. if $\lambda_m > 0$. Requiring $\lambda_m > m\gamma^2$ is enough for γ-fat-shattering. Theorem 5.4 then generalizes the simple observation, that if samples of size m are linearly independent with high probability, there is no hope of generalizing from $m/2$ points to the other $m/2$ using unregularized linear predictors. Theorem 5.4 can thus be used to derive a distribution-specific lower bound. Define:

$$\underline{m}_\gamma(D) \triangleq \frac{1}{2} \min \left\{ m \,\middle|\, \mathbb{P}_{S\sim D^m}[\lambda_m(X_S X_S') \geq m\gamma^2] < \frac{1}{2} \right\}$$

Then for any $\epsilon < 1/4 - \ell_\gamma^*(D)$, we can conclude that $m(\epsilon, \gamma, D) \geq \underline{m}_\gamma(D)$, that is, we cannot learn within reasonable error with less than $\underline{m}_\gamma$ examples. Recall that our upper-bound on the sample complexity from Section 4 was $\tilde{O}(k_\gamma)$. The remaining question is whether we can relate $\underline{m}_\gamma$ and k_γ, to establish that the our lower bound and upper bound tightly specify the sample complexity.

6 A lower bound for independently sub-Gaussian distributions

As discussed in the previous section, to obtain sample complexity lower bound we require a bound on the value of the smallest eigenvalue of a random Gram-matrix. The distribution of this eigenvalue has been investigated under various assumptions. The cleanest results are in the case where $m, d \to \infty$ and $\frac{m}{d} \to \beta < 1$, and the coordinates of each example are identically distributed:

Theorem 6.1 (Theorem 5.11 in [18])**.** *Let X_i be a series of $m_i \times d_i$ matrices whose entries are i.i.d. random variables with mean zero, variance σ^2 and finite fourth moments. If $\lim_{i\to\infty} \frac{m_i}{d_i} = \beta < 1$, then $\lim_{i\to\infty} \lambda_m(\frac{1}{d} X_i X_i') = \sigma^2(1-\sqrt{\beta})^2$.*

This asymptotic limit can be used to calculate $\underline{m}_\gamma$ and thus provide a lower bound on the sample complexity: Let the coordinates of $X \in \mathbb{R}^d$ be i.i.d. with variance σ^2 and consider a sample of size m. If d, m are large enough, we have by Theorem 6.1:

$$\lambda_m(XX') \approx d\sigma^2(1-\sqrt{m/d})^2 = \sigma^2(\sqrt{d}-\sqrt{m})^2$$

Solving $\sigma^2(\sqrt{d}-\sqrt{2\underline{m}_\gamma})^2 = 2\underline{m}_\gamma\gamma^2$ we get $\underline{m}_\gamma \approx \frac{1}{2}d/(1+\gamma/\sigma)^2$. We can also calculate the γ-adapted-dimension for this distribution to get $k_\gamma \approx d/(1+\gamma^2/\sigma^2)$, and conclude that $\frac{1}{4}k_\gamma \le \underline{m}_\gamma \le \frac{1}{2}k_\gamma$. In this case, then, we are indeed able to relate the sample complexity lower bound with k_γ, the same quantity that controls our upper bound. This conclusion is easy to derive from known results, however it holds only asymptotically, and only for a highly limited set of distributions. Moreover, since Theorem 6.1 holds asymptotically for each distribution separately, we cannot deduce from it any finite-sample lower bounds for families of distributions.

For our analysis we require *finite-sample* bounds for the smallest eigenvalue of a random Gram-matrix. Rudelson and Vershynin [19, 20] provide such finite-sample lower bounds for distributions with identically distributed sub-Gaussian coordinates. In the following Theorem we generalize results of Rudelson and Vershynin to encompass also non-identically distributed coordinates. The proof of Theorem 6.2 can be found in the appendix. Based on this theorem we conclude with Theorem 6.3, stated below, which constitutes our final sample complexity lower bound.

Theorem 6.2. *Let $B > 0$. There is a constant $\beta > 0$ which depends only on B, such that for any $\delta \in (0,1)$ there exists a number L_0, such that for any independently sub-Gaussian distribution with covariance matrix $\Sigma \le I$ and trace$(\Sigma) \ge L_0$, if each of its independent sub-Gaussian coordinates has moment B, then for any $m \le \beta \cdot$ trace(Σ)*

$$\mathbb{P}[\lambda_m(X_m X_m') \ge m] \ge 1-\delta,$$

Where X_m is an $m \times d$ matrix whose rows are independent draws from D_X.

Theorem 6.3 (Lower bound for distributions in $\mathcal{D}_\rho^{\text{sg}}$)**.** *For any $\rho > 0$, there are a constant $\beta > 0$ and an integer L_0 such that for any D such that $D_X \in \mathcal{D}_\rho^{sg}$ and $k_\gamma(D_X) > L_0$, for any margin $\gamma > 0$ and any $\epsilon < \frac{1}{4} - \ell_\gamma^*(D)$,*

$$m(\epsilon, \gamma, D) \ge \beta k_\gamma(D_X).$$

Proof. The covariance matrix of D_X is clearly diagonal. We assume w.l.o.g. that $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_d)$ where $\lambda_1 \ge \ldots \ge \lambda_d > 0$. Let S be an i.i.d. sample of size m drawn from D. Let X be the $m \times d$ matrix whose rows are the unlabeled examples from S. Let δ be fixed, and set β and L_0 as defined in Theorem 6.2 for δ. Assume $m \le \beta(k_\gamma - 1)$.

We would like to use Theorem 6.2 to bound the smallest eigenvalue of XX' with high probability, so that we can then apply Theorem 5.4 to get the desired lower bound. However, Theorem 6.2 holds only if all the coordinate variances are bounded by 1, and it requires that the moment, and not the relative moment, be bounded. Thus we divide the problem to two cases, based on the value of $\lambda_{k_\gamma+1}$, and apply Theorem 6.2 separately to each case.

Case I: Assume $\lambda_{k_\gamma+1} \ge \gamma^2$. Then $\forall i \in [k_\gamma], \lambda_i \ge \gamma^2$. Let $\Sigma_1 = \text{diag}(1/\lambda_1, \ldots, 1/\lambda_{k_\gamma}, 0, \ldots, 0)$. The random matrix $X\sqrt{\Sigma_1}$ is drawn from an independently sub-Gaussian distribution, such that each of its coordinates has sub-Gaussian moment ρ and covariance matrix $\Sigma \cdot \Sigma_1 \le I_d$. In addition, trace$(\Sigma \cdot \Sigma_1) = k_\gamma \ge L_0$. Therefore Theorem 6.2 holds for $X\sqrt{\Sigma_1}$, and $\mathbb{P}[\lambda_m(X\Sigma_1 X') \ge m] \ge 1-\delta$. Clearly, for any X, $\lambda_m(\frac{1}{\gamma^2}XX') \ge \lambda_m(X\Sigma_1 X')$. Thus $\mathbb{P}[\lambda_m(\frac{1}{\gamma^2}XX') \ge m] \ge 1-\delta$.

Case II: Assume $\lambda_{k_\gamma+1} < \gamma^2$. Then $\lambda_i < \gamma^2$ for all $i \in \{k_\gamma+1, \ldots, d\}$. Let $\Sigma_2 = \text{diag}(0, \ldots, 0, 1/\gamma^2, \ldots, 1/\gamma^2)$, with k_γ zeros on the diagonal. Then the random matrix $X\sqrt{\Sigma_2}$ is drawn from an independently sub-Gaussian distribution with covariance matrix $\Sigma \cdot \Sigma_2 \le I_d$, such that all its coordinates have sub-Gaussian moment ρ. In addition, from the properties of k_γ (see discussion in Section 2), trace$(\Sigma \cdot \Sigma_2) = \frac{1}{\gamma^2}\sum_{i=k_\gamma+1}^{d} \lambda_i \ge k_\gamma - 1 \ge L_0 - 1$. Thus Theorem 6.2 holds for $X\sqrt{\Sigma_2}$, and so $\mathbb{P}[\lambda_m(\frac{1}{\gamma^2}XX') \ge m] \ge \mathbb{P}[\lambda_m(X\Sigma_2 X') \ge m] \ge 1-\delta$.

In both cases $\mathbb{P}[\lambda_m(\frac{1}{\gamma^2}XX') \geq m] \geq 1-\delta$ for any $m \leq \beta(k_\gamma - 1)$. By Theorem 5.4, there exists an algorithm $\mathcal{A}$ such that for any $m \leq \beta(k_\gamma - 1) - 1$, $\ell_m(\mathcal{A}_\gamma, D) \geq \frac{1}{2} - \delta/2$. Therefore, for any $\epsilon < \frac{1}{2} - \delta/2 - \ell^*_\gamma(D)$, we have $m(\epsilon, \gamma, D) \geq \beta(k_\gamma - 1)$. We get the theorem by setting $\delta = \frac{1}{4}$. $\square$

7 Summary and consequences

Theorem 4.4 and Theorem 6.3 provide an upper bound and a lower bound for the sample complexity of any distribution D whose data distribution is in $\mathcal{D}^{\text{sg}}_\rho$ for some fixed $\rho > 0$. We can thus draw the following bound, which holds for any $\gamma > 0$ and $\epsilon \in (0, \frac{1}{4} - \ell^*_\gamma(D))$:

$$\Omega(k_\gamma(D_X)) \leq m(\epsilon, \gamma, D) \leq \tilde{O}(\frac{k_\gamma(D_X)}{\epsilon^2}). \tag{2}$$

In both sides of the bound, the hidden constants depend only on the constant ρ. This result shows that the true sample complexity of learning each of these distributions is characterized by the γ-adapted-dimension. An interesting conclusion can be drawn as to the influence of the conditional distribution of labels $D_{Y|X}$: Since Eq. (2) holds for any $D_{Y|X}$, the effect of the direction of the best separator on the sample complexity is bounded, even for highly non-spherical distributions. We can use Eq. (2) to easily characterize the sample complexity behavior for interesting distributions, and to compare L_2 margin minimization to learning methods.

Gaps between L_1 and L_2 regularization in the presence of irrelevant features. Ng [3] considers learning a single relevant feature in the presence of many irrelevant features, and compares using L_1 regularization and L_2 regularization. When $\|X\|_\infty \leq 1$, upper bounds on learning with L_1 regularization guarantee a sample complexity of $O(\log(d))$ for an L_1-based learning rule [21]. In order to compare this with the sample complexity of L_2 regularized learning and establish a gap, one must use a *lower bound* on the L_2 sample complexity. The argument provided by Ng actually assumes scale-invariance of the learning rule, and is therefore valid only for *unregularized* linear learning. However, using our results we can easily establish a lower bound of $\Omega(d)$ for many specific distributions with $\|X\|_\infty \leq 1$ and $Y = X[1] \in \{\pm 1\}$. For instance, when each coordinate is an independent Bernoulli variable, the distribution is sub-Gaussian with $\rho = 1$, and $k_1 = \lceil d/2 \rceil$.

Gaps between generative and discriminative learning for a Gaussian mixture. Consider two classes, each drawn from a unit-variance spherical Gaussian in a high dimension $\mathbb{R}^d$ and with a large distance $2v >> 1$ between the class means, such that $d >> v^4$. Then $\mathbb{P}_D[X|Y = y] = \mathcal{N}(yv \cdot e_1, I_d)$, where e_1 is a unit vector in $\mathbb{R}^d$. For any v and d, we have $D_X \in \mathcal{D}^{\text{sg}}_1$. For large values of v, we have extremely low margin error at $\gamma = v/2$, and so we can hope to learn the classes by looking for a large-margin separator. Indeed, we can calculate $k_\gamma = \lceil d/(1 + \frac{v^2}{4}) \rceil$, and conclude that the sample complexity required is $\tilde{\Theta}(d/v^2)$. Now consider a generative approach: fitting a spherical Gaussian model for each class. This amounts to estimating each class center as the empirical average of the points in the class, and classifying based on the nearest estimated class center. It is possible to show that for any constant $\epsilon > 0$, and for large enough v and d, $O(d/v^4)$ samples are enough in order to ensure an error of ϵ. This establishes a rather large gap of $\Omega(v^2)$ between the sample complexity of the discriminative approach and that of the generative one.

To summarize, we have shown that the true sample complexity of large-margin learning of a rich family of specific distributions is characterized by the γ-adapted-dimension. This result allows true comparison between this learning algorithm and other algorithms, and has various applications, such as semi-supervised learning and feature construction. The challenge of characterizing true sample complexity extends to any distribution and any learning algorithm. We believe that obtaining answers to these questions is of great importance, both to learning theory and to learning applications.

Acknowledgments

The authors thank Boaz Nadler for many insightful discussions, and Karthik Sridharan for pointing out [14] to us. Sivan Sabato is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. This work was supported by the NATO SfP grant 982480.

References

[1] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

[2] P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *ICML*, 2010.

[3] A.Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *ICML*, 2004.

[4] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Mach. Learn.*, 30(1):31–56, 1998.

[5] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. In *Proceedings of the First Anuual Workshop on Computational Learning Theory*, pages 139–154, August 1988.

[6] C. Gentile and D.P. Helmbold. Improved lower bounds for learning from noisy examples: an information-theoretic approach. In *COLT*, pages 104–115, 1998.

[7] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, September 1991.

[9] S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? In *Proceedings of the Twenty-First Annual Conference on Computational Learning Theory*, pages 33–44, 2008.

[10] N. Vayatis and R. Azencott. Distribution-dependent vapnik-chervonenkis bounds. In *EuroCOLT '99*, pages 230–240, London, UK, 1999. Springer-Verlag.

[11] D.J.H. Garling. *Inequalities: A Journey into Linear Analysis*. Cambrige University Press, 2007.

[12] V.V. Buldygin and Yu. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 1998.

[13] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. In *COLT 2001*, volume 2111, pages 224–240. Springer, Berlin, 2001.

[14] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.

[15] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R.C. Williamson. Generalization bounds via eigenvalues of the gram matrix. Technical Report NC2-TR-1999-035, NeuroCOLT2, 1999.

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

[17] N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[18] Z. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, second edition edition, 2010.

[19] M. Rudelson and R. Vershynin. The smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739, 2009.

[20] M. Rudelson and R. Vershynin. The littlewoodofford problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[21] T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

[22] G. Bennett, V. Goodman, and C. M. Newman. Norms of random matrices. *Pacific J. Math.*, 59(2):359–365, 1975.

[23] F.L. Nazarov and A. Podkorytov. Ball, haagerup, and distribution functions. *Operator Theory: Advances and Applications*, 113 (Complex analysis, operators, and related topics):247–267, 2000.

[24] R.E.A.C. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Proceedings of the Cambridge Philosophical Society*, 28:266272, 1932.