Distribution-Dependent Sample Complexity
of Large Margin Learning

Sivan Sabato   SIVAN.SABATO@MICROSOFT.COM
Microsoft Research New England
1 Memorial Drive
Cambridge, MA 02142, USA

Nathan Srebro   NATI@TTIC.EDU
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave.
Chicago, IL 60637, USA

Naftali Tishby   TISHBY@CS.HUJI.AC.IL
The Rachel and Selim Benin School of Computer Science and Engineering
The Hebrew University
Jerusalem 91904, Israel

Abstract

We obtain a tight distribution-specific characterization of the sample complexity of large-margin classification with L_2 regularization: We introduce the *margin-adapted dimension*, which is a simple function of the second order statistics of the data distribution, and show distribution-specific upper *and* lower bounds on the sample complexity, both governed by the margin-adapted dimension of the data distribution. The upper bounds are universal, and the lower bounds hold for the rich family of sub-Gaussian distributions with independent features. We conclude that this new quantity tightly characterizes the true sample complexity of large-margin classification. To prove the lower bound, we develop several new tools of independent interest. These include new connections between shattering and hardness of learning, new properties of shattering with linear classifiers, and a new lower bound on the smallest eigenvalue of a random Gram matrix generated by sub-Gaussian variables. Our results can be used to quantitatively compare large margin learning to other learning rules, and to improve the effectiveness of methods that use sample complexity bounds, such as active learning.

Keywords: supervised learning, sample complexity, linear classifiers, distribution-dependence

1. Introduction

In this paper we pursue a tight characterization of the sample complexity of learning a classifier, under a particular data distribution, and using a particular learning rule.

Most learning theory work focuses on providing sample-complexity upper bounds which hold for a large class of distributions. For instance, standard distribution-free VC-dimension analysis shows that if one uses the Empirical Risk Minimization (ERM) learning rule, then the sample complexity of learning a classifier from a hypothesis class with VC-dimension d is at most $O\left(\frac{d}{\varepsilon^2}\right)$, where ε is the maximal excess classification error (Vapnik and Chervonenkis, 1971; Anthony and Bartlett, 1999). Such upper bounds can be useful for understanding the positive aspects of a learning rule.

However, it is difficult to understand the deficiencies of a learning rule, or to compare between different rules, based on upper bounds alone. This is because it is possible, and is often the case, that the actual number of samples required to get a low error, for a given data distribution using a given learning rule, is much lower than the sample-complexity upper bound. As a simple example, suppose that the support of a given distribution is restricted to a subset of the domain. If the VC-dimension of the hypothesis class, when restricted to this subset, is smaller than d, then learning with respect to this distribution will require less examples than the upper bound predicts.

Of course, some sample complexity upper bounds are known to be tight or to have an almost-matching lower bound. For instance, the VC-dimension upper bound is tight (Vapnik and Chervonenkis, 1974). This means that there exists *some* data distribution in the class covered by the upper bound, for which this bound cannot be improved. Such a tightness result shows that there cannot be a better upper bound that holds for this entire class of distributions. But it does not imply that the upper bound characterizes the true sample complexity for every *specific* distribution in the class.

The goal of this paper is to identify a simple quantity, which is a function of the distribution, that *does* precisely characterize the sample complexity of learning this distribution under a specific learning rule. We focus on the important hypothesis class of linear classifiers, and on the popular rule of *margin-error-minimization* (MEM). Under this learning rule, a learner must always select a linear classifier that minimizes the margin-error on the input sample.

The VC-dimension of the class of homogeneous linear classifiers in $\mathbb{R}^d$ is d (Dudley, 1978). This implies a sample complexity upper bound of $O\left(\frac{d}{\varepsilon^2}\right)$ using any MEM algorithm, where ε is the excess error relative to the optimal margin error.[1] We also have that the sample complexity of any MEM algorithm is at most $O\left(\frac{B^2}{\gamma^2 \varepsilon^2}\right)$, where B^2 is the average squared norm of the data and γ is the size of the margin (Bartlett and Mendelson, 2002). Both of these upper bounds are tight. For instance, there exists a distribution with an average squared norm of B^2 that requires as many as $C \cdot \frac{B^2}{\gamma^2 \varepsilon^2}$ examples to learn, for some universal constant C (see, e.g., Anthony and Bartlett, 1999). However, the VC-dimension upper bound indicates, for instance, that if a distribution induces a large average norm but is supported by a low-dimensional sub-space, then the true number of examples required to reach a low error is much smaller. Thus, neither of these upper bounds fully describes the sample complexity of MEM for a *specific* distribution.

We obtain a tight distribution-specific characterization of the sample complexity of large-margin learning for a rich class of distributions. We present a new quantity, termed the *margin-adapted dimension*, and use it to provide a tighter distribution-dependent upper bound, and a matching distribution-dependent lower bound for MEM. The upper bound is universal, and the lower bound holds for a rich class of distributions with independent features.

The margin-adapted dimension refines both the dimension and the average norm of the data distribution, and can be easily calculated from the covariance matrix and the mean of the distribution. We denote this quantity, for a margin of γ, by k_γ. Our sample-complexity upper bound shows that $\tilde{O}(\frac{k_\gamma}{\varepsilon^2})$ examples suffice in order to learn any distribution with a margin-adapted dimension of k_γ using a MEM algorithm with margin γ. We further show that for every distribution in a rich family of 'light tailed' distributions—specifically, product distributions of sub-Gaussian random variables—the number of samples required for learning by minimizing the margin error is at least $\Omega(k_\gamma)$.

---

1. This upper bound can be derived analogously to the result for ERM algorithms with ε being the excess classification error. It can also be concluded from our analysis in Theorem 11 below.

Denote by $m(\varepsilon, \gamma, D)$ the number of examples required to achieve an excess error of no more than ε relative to the best possible γ-margin error for a specific distribution D, using a MEM algorithm. Our main result shows the following matching distribution-specific upper and lower bounds on the sample complexity of MEM:

$$\Omega(k_\gamma(D)) \le m(\varepsilon, \gamma, D) \le \tilde{O}\left(\frac{k_\gamma(D)}{\varepsilon^2}\right). \tag{1}$$

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. 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 Ng (2004), and to show a large gap between discriminative and generative learning on a Gaussian-mixture distribution. The tight bounds can also be used for active learning algorithms in which sample-complexity bounds are used to decide on the next label to query.

In this paper we focus only on large margin classification. But in order to obtain the distribution-specific lower bound, we develop new tools that we believe can be useful for obtaining lower bounds also for other learning rules. We provide several new results which we use to derive our main results. These include:

- Linking the fat-shattering of a sample with non-negligible probability to a difficulty of learning using MEM.

- Showing that for a convex hypothesis class, fat-shattering is equivalent to shattering with exact margins.

- Linking the fat-shattering of a set of vectors with the eigenvalues of the Gram matrix of the vectors.

- Providing a new lower bound for the smallest eigenvalue of a random Gram matrix generated by sub-Gaussian variables. This bound extends previous results in analysis of random matrices.

1.1 Paper Structure

We discuss related work on sample-complexity upper bounds in Section 2. We present the problem setting and notation in Section 3, and provide some necessary preliminaries in Section 4. We then introduce the margin-adapted dimension in Section 5. The sample-complexity upper bound is proved in Section 6. We prove the lower bound in Section 7. In Section 8 we show that any non-trivial sample-complexity lower bound for more general distributions must employ properties other than the covariance matrix of the distribution. We summarize and discuss implication in Section 9. Proofs omitted from the text are provided in Appendix A

2. Related Work

As mentioned above, most work on "sample complexity lower bounds" is directed at proving that under some set of assumptions, there exists a data distribution for which one needs at least a certain number of examples to learn with required error and confidence (for instance Antos and Lugosi,

1998; Ehrenfeucht et al., 1988; Gentile and Helmbold, 1998). This type of a lower bound does not, however, indicate much on the sample complexity of other distributions under the same set of assumptions.

For distribution-specific lower bounds, the classical analysis of Vapnik (1995, 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 metric 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 (1991) 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. (2008) 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. Vayatis and Azencott (1999) 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 they also do not fully characterize the distribution-specific sample complexity.

As can be seen in Equation (1), we do not tightly characterize the dependence of the sample complexity on the desired error (as done, for example, in Steinwart and Scovel, 2007), thus our bounds are not tight for asymptotically small error levels. Our results are most significant if the desired error level is a constant well below chance but bounded away from zero. This is in contrast to classical statistical asymptotics that are also typically tight, but are valid only for very small ε. As was recently shown by Liang and Srebro (2010), the sample complexity for very small ε (in the classical statistical asymptotic regime) depends on quantities that can be very different from those that control the sample complexity for moderate error rates, which are more relevant for machine learning.

3. Problem Setting and Definitions

Consider a domain X, and let D be a distribution over $X \times \{\pm 1\}$. We denote by D_X the marginal distribution of D on X. The misclassification error of a classifier $h : X \to \mathbb{R}$ on a distribution D is

$$\ell_0(h, D) \triangleq \mathbb{P}_{(X,Y) \sim D}[Y \cdot h(X) \leq 0].$$

The margin error of a classifier w with respect to a margin $\gamma > 0$ on D is

$$\ell_\gamma(h, D) \triangleq \mathbb{P}_{(X,Y) \sim D}[Y \cdot h(X) \leq \gamma].$$

For a given hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^X$, the best achievable margin error on D is

$$\ell_\gamma^*(\mathcal{H}, D) \triangleq \inf_{h \in \mathcal{H}} \ell_\gamma(h, D).$$

We usually write simply $\ell_\gamma^*(D)$ since $\mathcal{H}$ is clear from context.

A labeled sample is a (multi-)set $S = \{(x_i, y_i)\}_{i=1}^{m} \subseteq X \times \{\pm 1\}$. Given S, we denote the set of its examples without their labels by $S_X \triangleq \{x_1, \ldots, x_m\}$. We use S also to refer to the uniform distribution over the elements in S. Thus the misclassification error of $h : X \rightarrow \{\pm 1\}$ on S is

$$\ell(h, S) \triangleq \frac{1}{m} |\{i \mid y_i \cdot h(x_i) \leq 0\}|,$$

and the γ-margin error on S is

$$\ell_\gamma(h, S) \triangleq \frac{1}{m} |\{i \mid y_i \cdot h(x_i) \leq \gamma\}|.$$

A learning algorithm is a function $\mathcal{A} : \cup_{m=1}^{\infty} (X \times \{\pm 1\})^m \rightarrow \mathbb{R}^X$, that receives a training set as input, and returns a function for classifying objects in X into real values. The high-probability loss of an algorithm $\mathcal{A}$ with respect to samples of size m, a distribution D and a confidence parameter $\delta \in (0, 1)$ is

$$\ell(\mathcal{A}, D, m, \delta) = \inf\{\varepsilon \geq 0 \mid \mathbb{P}_{S \sim D^m}[\ell(\mathcal{A}(S), D) \geq \varepsilon] \leq \delta\}.$$

In this work we investigate the sample complexity of learning using margin-error minimization (MEM). The relevant class of algorithms is defined as follows.

Definition 1 *An margin-error minimization (MEM) algorithm $\mathcal{A}$ maps a margin parameter $\gamma > 0$ to a learning algorithm $\mathcal{A}_\gamma$, such that*

$$\forall S \subseteq X \times \{\pm 1\}, \quad \mathcal{A}_\gamma(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S).$$

The distribution-specific sample complexity for MEM algorithms is the sample size required to guarantee low excess error for the given distribution. Formally, we have the following definition.

Definition 2 (Distribution-specific sample complexity) *Fix a hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^X$. For $\gamma > 0$, $\varepsilon, \delta \in [0, 1]$, and a distribution D, the* distribution-specific sample complexity, *denoted by $m(\varepsilon, \gamma, D, \delta)$, is the minimal sample size such that for any MEM algorithm $\mathcal{A}$, and for any $m \geq m(\varepsilon, \gamma, D, \delta)$,*

$$\ell_0(\mathcal{A}_\gamma, D, m, \delta) - \ell_\gamma^*(D) \leq \varepsilon.$$

Note that we require that *all* possible MEM algorithms do well on the given distribution. This is because we are interested in the MEM strategy in general, and thus we study the guarantees that can be provided regardless of any specific MEM implementation. We sometimes omit δ and write simply $m(\varepsilon, \gamma, D)$, to indicate that δ is assumed to be some fixed small constant.

In this work we focus on linear classifiers. For simplicity of notation, we assume a Euclidean space $\mathbb{R}^d$ for some integer d, although the results can be easily extended to any separable Hilbert space. For a real vector x, $\|x\|$ stands for the Euclidean norm. For a real matrix $\mathbb{X}$, $\|\mathbb{X}\|$ stands for the Euclidean operator norm.

Denote the unit ball in $\mathbb{R}^d$ by $\mathbb{B}_1^d \triangleq \{w \in \mathbb{R}^d \mid \|w\| \leq 1\}$. We consider the hypothesis class of homogeneous linear separators, $\mathcal{W} = \{x \mapsto \langle x, w \rangle \mid w \in \mathbb{B}_1^d\}$. We often slightly abuse notation by using w to denote the mapping $x \mapsto \langle x, w \rangle$.

We often represent sets of vectors in $\mathbb{R}^d$ using matrices. We say that $\mathbb{X} \in \mathbb{R}^{m \times d}$ is the matrix of a set $\{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$ if the rows in the matrix are exactly the vectors in the set. For uniqueness, one may assume that the rows of $\mathbb{X}$ are sorted according to an arbitrary fixed full order on vectors in

$\mathbb{R}^d$. For a PSD matrix $\mathbb{X}$ denote the largest eigenvalue of $\mathbb{X}$ by $\lambda_{\max}(\mathbb{X})$ and the smallest eigenvalue by $\lambda_{\min}(\mathbb{X})$.

We use the O-notation as follows: $O(f(z))$ stands for $C_1 + C_2 f(z)$ for some constants $C_1, C_2 \geq 0$. $\Omega(f(z))$ stands for $C_2 f(z) - C_1$ for some constants $C_1, C_2 \geq 0$. $\widetilde{O}(f(z))$ stands for $f(z)p(\ln(z)) + C$ for some polynomial $p(\cdot)$ and some constant $C > 0$.

4. Preliminaries

As mentioned above, for the hypothesis class of linear classifiers $\mathcal{W}$, one can derive a sample-complexity upper bound of the form $O(B^2/\gamma^2\varepsilon^2)$, where $B^2 = \mathbb{E}_{X \sim D}[\|X\|^2]$ and ε is the excess error relative to the γ-margin loss. This can be achieved as follows (Bartlett and Mendelson, 2002). Let $\mathcal{Z}$ be some domain. The empirical Rademacher complexity of a class of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ with respect to a set $S = \{z_i\}_{i \in [m]} \subseteq \mathcal{Z}$ is

$$\mathcal{R}(\mathcal{F}, S) = \frac{1}{m}\mathbb{E}_\sigma[|\sup_{f \in \mathcal{F}} \sum_{i \in [m]} \sigma_i f(z_i)|],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ are m independent uniform $\{\pm 1\}$-valued variables. The average Rademacher complexity of $\mathcal{F}$ with respect to a distribution D over $\mathcal{Z}$ and a sample size m is

$$\mathcal{R}_m(\mathcal{F}, D) = \mathbb{E}_{S \sim D^m}[\mathcal{R}(\mathcal{F}, S)].$$

Assume a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ and a loss function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$. For a hypothesis $h \in \mathcal{H}$, we introduce the function $h_\ell : \mathcal{X} \times \{\pm 1\} \to \mathbb{R}$, defined by $h_\ell(x, y) = \ell(y, h(x))$. We further define the function class $\mathcal{H}_\ell = \{h_\ell \mid h \in \mathcal{H}\} \subseteq \mathbb{R}^{\mathcal{X} \times \{\pm 1\}}$.

Assume that the range of $\mathcal{H}_\ell$ is in $[0, 1]$. For any $\delta \in (0, 1)$, with probability of $1 - \delta$ over the draw of samples $S \subseteq \mathcal{X} \times \{\pm 1\}$ of size m according to D, every $h \in \mathcal{H}$ satisfies (Bartlett and Mendelson, 2002)

$$\ell(h, D) \leq \ell(h, S) + 2\mathcal{R}_m(\mathcal{H}_\ell, D) + \sqrt{\frac{8\ln(2/\delta)}{m}}. \tag{2}$$

To get the desired upper bound for linear classifiers we use the *ramp loss*, which is defined as follows. For a number r, denote $[\![r]\!] \triangleq \min(\max(r, 0), 1)$. The γ-ramp-loss of a labeled example $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ with respect to a linear classifier $w \in \mathbb{B}_1^d$ is $\mathrm{ramp}_\gamma(w, x, y) = [\![1 - y\langle w, x\rangle/\gamma]\!]$. Let $\mathrm{ramp}_\gamma(w, D) = \mathbb{E}_{(X,Y) \sim D}[\mathrm{ramp}_\gamma(w, X, Y)]$, and denote the class of ramp-loss functions by

$$\mathrm{RAMP}_\gamma = \{(x, y) \mapsto \mathrm{ramp}_\gamma(w, x, y) \mid w \in \mathbb{B}_1^d\}.$$

The ramp-loss is upper-bounded by the margin loss and lower-bounded by the misclassification error. Therefore, the following result can be shown.

Proposition 3 *For any MEM algorithm $\mathcal{A}$, we have*

$$\ell_0(\mathcal{A}_\gamma, D, m, \delta) \leq \ell_\gamma^*(\mathcal{H}, D) + 2\mathcal{R}_m(\mathrm{RAMP}_\gamma, D) + \sqrt{\frac{14\ln(2/\delta)}{m}}. \tag{3}$$

We give the proof in Appendix A.1 for completeness. Since the γ-ramp loss is $1/\gamma$ Lipschitz, it follows from Bartlett and Mendelson (2002) that

$$\mathcal{R}_m(\mathrm{RAMP}_\gamma, D) \leq \sqrt{\frac{B^2}{\gamma^2 m}}.$$

Combining this with Proposition 3 we can conclude a sample complexity upper bound of $O(B^2/\gamma^2\varepsilon^2)$.

In addition to the Rademacher complexity, we will also use the classic notions of *fat-shattering* (Kearns and Schapire, 1994) and *pseudo-shattering* (Pollard, 1984), defined as follows.

Definition 4 *Let $\mathcal{F}$ be a set of functions $f : X \to \mathbb{R}$, and let $\gamma > 0$. The set $\{x_1, \ldots, x_m\} \subseteq X$ is γ-shattered by $\mathcal{F}$ with the witness $r \in \mathbb{R}^m$ if 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-shattering dimension of a hypothesis class is the size of the largest set that is γ-shattered by this class. We say that a set is γ-*shattered at the origin* if it is γ-shattered with the zero vector as a witness.

Definition 5 *Let $\mathcal{F}$ be a set of functions $f : X \to \mathbb{R}$, and let $\gamma > 0$. The set $\{x_1, \ldots, x_m\} \subseteq X$ is pseudo-shattered by $\mathcal{F}$ with the witness $r \in \mathbb{R}^m$ if for all $y \in \{\pm 1\}^m$ there is an $f \in \mathcal{F}$ such that $\forall i \in [m], y[i](f(x_i) - r[i]) > 0$.*

The pseudo-dimension of a hypothesis class is the size of the largest set that is pseudo-shattered by this class.

5. The Margin-Adapted Dimension

When considering learning of linear classifiers using MEM, the dimension-based upper bound and the norm-based upper bound are both tight in the worst-case sense, that is, they are the best bounds that rely only on the dimensionality or only on the norm respectively. Nonetheless, neither is tight in a distribution-specific sense: If the average norm is unbounded while the dimension is small, then there can be an arbitrarily large gap between the true distribution-dependent sample complexity and the bound that depends on the average norm. If the converse holds, that is, the dimension is arbitrarily large while the average-norm is bounded, then the dimensionality bound is loose.

Seeking a tight distribution-specific analysis, one simple approach to tighten these bounds is to consider their minimum, which is proportional to $\min(d, B^2/\gamma^2)$. Trivially, this is an upper bound on the sample complexity as well. 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 (see Figure 1). 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 squared norm in the other directions. This behavior is captured by the *margin-adapted dimension* which we presently define, using the following auxiliary definition.

Definition 6 *Let $b > 0$ and let k be a positive integer. 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 $\mathbb{E}_{X \sim D_X}[\|\mathbb{O}_V \cdot X\|^2] \leq b$, where $\mathbb{O}_V$ is an orthogonal projection onto V.*

Definition 7 (margin-adapted dimension) *The* margin-adapted dimension *of a distribution D_X, denoted by $k_\gamma(D_X)$, is the minimum k such that the distribution is $(\gamma^2 k, k)$-limited.*

We sometimes drop the argument of k_γ when it is clear from context. It is easy to see that for any distribution D_X over $\mathbb{R}^d$, $k_\gamma(D_X) \leq \min(d, \mathbb{E}[\|X\|^2]/\gamma^2)$. Moreover, k_γ can be much smaller than this minimum. For example, consider a random vector $X \in \mathbb{R}^{1001}$ with mean zero and statistically independent coordinates, such that the variance of the first coordinate is 1000, and the variance in each remaining coordinate is 0.001. We have $k_1 = 1$ but $d = \mathbb{E}[\|X\|^2] = 1001$.

$k_\gamma(D_X)$ can be calculated from the uncentered covariance matrix $\mathbb{E}_{X \sim D_X}[XX^T]$ as follows: Let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_d \geq 0$ be the eigenvalues of this matrix. Then

$$k_\gamma = \min\{k \mid \sum_{i=k+1}^{d} \lambda_i \leq \gamma^2 k\}. \tag{4}$$

A quantity similar to this definition of k_γ was studied previously in Bousquet (2002). The eigenvalues of the *empirical* covariance matrix were used to provide sample complexity bounds, for instance in Schölkopf et al. (1999). However, k_γ generates a different type of bound, since it is defined based on the eigenvalues of the distribution and not of the sample. We will see that for small finite samples, the latter can be quite different from the former.

Finally, note that while we define the margin-adapted dimension for a finite-dimensional space for ease of notation, the same definition carries over to an infinite-dimensional Hilbert space. Moreover, k_γ can be finite even if some of the eigenvalues λ_i are infinite, implying a distribution with unbounded covariance.

6. A Distribution-Dependent Upper Bound

In this section we prove an upper bound on the sample complexity of learning with MEM, using the margin-adapted dimension. We do this by providing a tighter upper bound for the Rademacher complexity of RAMP_γ. We bound $\mathcal{R}_m(\text{RAMP}_\gamma, D)$ for any (B^2, k)-limited distribution D_X, using L_2 covering numbers, defined as follows.

Let $(X, \|\cdot\|_\circ)$ be a normed space. An η-covering of a set $\mathcal{F} \subseteq X$ with respect to the norm $\|\cdot\|_\circ$ is a set $C \subseteq X$ such that for any $f \in \mathcal{F}$ there exists a $g \in C$ such that $\|f - g\|_\circ \leq \eta$. The covering-number for given $\eta > 0$, $\mathcal{F}$ and $\circ$ is the size of the smallest such η-covering, and is denoted by $\mathcal{N}(\eta, \mathcal{F}, \circ)$. Let $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, the $L_2(S)$ norm of f is $\|f\|_{L_2(S)} = \sqrt{\mathbb{E}_{X \sim S}[f(X)^2]}$. Thus, we consider covering-numbers of the form $\mathcal{N}(\eta, \text{RAMP}_\gamma, L_2(S))$.
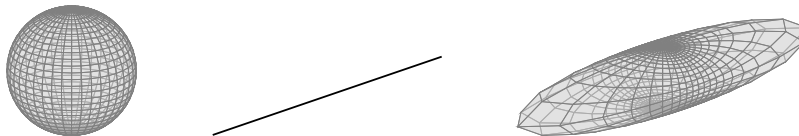

Figure 1: Illustrating covariance matrix ellipsoids. left: norm bound is tight; middle: dimension bound is tight; right: neither bound is tight.

The empirical Rademacher complexity of a function class can be bounded by the L_2 covering numbers of the same function class as follows (Mendelson, 2002, Lemma 3.7): Let $\varepsilon_i = 2^{-i}$. Then

$$\sqrt{m}\mathcal{R}(\text{RAMP}_\gamma, S) \leq C \sum_{i \in [N]} \varepsilon_{i-1} \sqrt{\ln \mathcal{N}(\varepsilon_i, \text{RAMP}_\gamma, L_2(S))} + 2\varepsilon_N \sqrt{m}. \tag{5}$$

To bound the covering number of RAMP_γ, we will restate the functions in RAMP_γ as sums of two functions, each selected from a function class with bounded complexity. The first function class will be bounded because of the norm bound on the subspace V used in Definition 6, and the second function class will have a bounded pseudo-dimension. However, the second function class will depend on the choice of the first function in the sum. Therefore, we require the following lemma, which provides an upper bound on such sums of functions. We use the notion of a *Hausdorff distance* between two sets $G_1, G_2 \subseteq X$, defined as $\Delta_H(G_1, G_2) = \sup_{g_1 \in G_1} \inf_{g_2 \in G_2} \|g_1 - g_2\|_\circ$.

Lemma 8 *Let $(X, \|\cdot\|_\circ)$ be a normed space. Let $\mathcal{F} \subseteq X$ be a set, and let $\mathcal{G} : X \to 2^X$ be a mapping from objects in X to sets of objects in X. Assume that $\mathcal{G}$ is c-Lipschitz with respect to the Hausdorff distance on sets, that is assume that*

$$\forall f_1, f_2 \in X, \Delta_H(\mathcal{G}(f_1), \mathcal{G}(f_2)) \leq c\|f_1 - f_2\|_\circ.$$

Let $\mathcal{F}_G = \{f + g \mid f \in \mathcal{F}, g \in \mathcal{G}(f)\}$. Then

$$\mathcal{N}(\eta, \mathcal{F}_G, \circ) \leq \mathcal{N}(\eta/(2+c), \mathcal{F}, \circ) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta/(2+c), \mathcal{G}(f), \circ).$$

Proof For any set $A \subseteq X$, denote by C_A a minimal η-covering for A with respect to $\|\cdot\|_\circ$, so that $|C_A| = \mathcal{N}(\eta, A, \circ)$. Let $f + g \in \mathcal{F}_G$ such that $f \in \mathcal{F}, g \in \mathcal{G}(f)$. There is a $\hat{f} \in C_{\mathcal{F}}$ such that $\|f - \hat{f}\|_\circ \leq \eta$. In addition, by the Lipschitz assumption there is a $\tilde{g} \in \mathcal{G}(\hat{f})$ such that $\|g - \tilde{g}\|_\circ \leq c\|f - \hat{f}\|_\circ \leq c\eta$. Lastly, there is a $\hat{g} \in C_{\mathcal{G}(\hat{f})}$ such that $\|\tilde{g} - \hat{g}\|_\circ \leq \eta$. Therefore

$$\|f + g - (\hat{f} + \hat{g})\|_\circ \leq \|f - \hat{f}\|_\circ + \|g - \tilde{g}\|_\circ + \|\tilde{g} - \hat{g}\|_\circ \leq (2 + c)\eta.$$

Thus the set $\{f + g \mid f \in C_{\mathcal{F}}, g \in C_{\mathcal{G}(f)}\}$ is a $(2+c)\eta$ cover of $\mathcal{F}_G$. The size of this cover is at most $|C_{\mathcal{F}}| \cdot \sup_{f \in \mathcal{F}} |C_{\mathcal{G}(f)}| \leq \mathcal{N}(\eta, \mathcal{F}, \circ) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta, \mathcal{G}(f), \circ)$. ∎

The following lemma provides us with a useful class of mappings which are 1-Lipschitz with respect to the Hausdorff distance, as required in Lemma 8. The proof is provided in Appendix A.2.

Lemma 9 *Let $f : X \to \mathbb{R}$ be a function and let $Z \subseteq \mathbb{R}^X$ be a function class over some domain X. Let $\mathcal{G} : \mathbb{R}^X \to 2^{\mathbb{R}^X}$ be the mapping defined by*

$$\mathcal{G}(f) \triangleq \{x \mapsto [\![f(x) + z(x)]\!] - f(x) \mid z \in Z\}. \tag{6}$$

Then $\mathcal{G}$ is 1-Lipschitz with respect to the Hausdorff distance.

The function class induced by the mapping above preserves the pseudo-dimension of the original function class, as the following lemma shows. The proof is provided in Appendix A.3.

Lemma 10 *Let $f : X \to \mathbb{R}$ be a function and let $Z \subseteq \mathbb{R}^X$ be a function class over some domain X. Let $\mathcal{G}(f)$ be defined as in Equation (6). Then the pseudo-dimension of $\mathcal{G}(f)$ is at most the pseudo-dimension of Z.*

Equipped with these lemmas, we can now provide the new bound on the Rademacher complexity of RAMP_γ in the following theorem. The subsequent corollary states the resulting sample-complexity upper bound for MEM, which depends on k_γ.

Theorem 11 *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and assume D_X is (B^2, k)-limited. Then*

$$\mathcal{R}(\text{RAMP}_\gamma, D) \leq \sqrt{\frac{O(k + B^2/\gamma^2)\ln(m)}{m}}.$$

Proof In this proof all absolute constants are assumed to be positive and are denoted by C or C_i for some integer i. Their values may change from line to line or even within the same line.

Consider the distribution $\tilde{D}$ which results from drawing $(X, Y) \sim D$ and emitting $(Y \cdot X, 1)$. It too is (B^2, k)-limited, and $\mathcal{R}(\text{RAMP}_\gamma, D) = \mathcal{R}(\text{RAMP}_\gamma, \tilde{D})$. Therefore, we assume without loss of generality that for all (X, Y) drawn from D, $Y = 1$. Accordingly, we henceforth omit the y argument from $\text{ramp}_\gamma(w, x, y)$ and write simply $\text{ramp}_\gamma(w, x) \triangleq \text{ramp}_\gamma(w, x, 1)$.

Following Definition 6, Let $\mathbb{O}_V$ be an orthogonal projection onto a sub-space V of dimension $d - k$ such that $\mathbb{E}_{X \sim D_X}[\|\mathbb{O}_V \cdot X\|^2] \leq B^2$. Let $\bar{V}$ be the complementary sub-space to V. For a set $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$, denote $B(S) = \sqrt{\frac{1}{m}\sum_{i \in [m]} \|\mathbb{O}_V \cdot X\|^2}$.

We would like to use Equation (5) to bound the Rademacher complexity of RAMP_γ. Therefore, we will bound $\mathcal{N}(\eta, \text{RAMP}_\gamma, L_2(S))$ for $\eta > 0$. Note that

$$\text{ramp}_\gamma(w, x) = [\![1 - \langle w, x \rangle / \gamma]\!] = 1 - [\![\langle w, x \rangle / \gamma]\!].$$

Shifting by a constant and negating do not change the covering number of a function class. Therefore, $\mathcal{N}(\eta, \text{RAMP}_\gamma, L_2(S))$ is equal to the covering number of $\{x \mapsto [\![\langle w, x \rangle / \gamma]\!] \mid w \in \mathbb{B}_1^d\}$. Moreover, let

$$\text{RAMP}_\gamma' = \{x \mapsto [\![\langle w_a + w_b, x \rangle / \gamma]\!] \mid w_a \in \mathbb{B}_1^d \cap V, \, w_b \in \bar{V}\}.$$

Then $\{x \mapsto [\![\langle w, x \rangle / \gamma]\!] \mid w \in \mathbb{B}_1^d\} \subseteq \text{RAMP}_\gamma'$, thus it suffices to bound $\mathcal{N}(\eta, \text{RAMP}_\gamma', L_2(S))$. To do that, we show that RAMP_γ' satisfies the assumptions of Lemma 8 for the normed space $(\mathbb{R}^{\mathbb{R}^d}, \|\cdot\|_{L_2(S)})$. Define

$$\mathcal{F} = \{x \mapsto \langle w_a, x \rangle / \gamma \mid w_a \in \mathbb{B}_1^d \cap V\}.$$

Let $\mathcal{G} : \mathbb{R}^{\mathbb{R}^d} \to 2^{\mathbb{R}^{\mathbb{R}^d}}$ be the mapping defined by

$$\mathcal{G}(f) \triangleq \{x \mapsto [\![f(x) + \langle w_b, x \rangle / \gamma]\!] - f(x) \mid w_b \in \bar{V}\}.$$

Clearly, $\mathcal{F}_\mathcal{G} = \{f + g \mid f \in \mathcal{F}, g \in \mathcal{G}(f)\} = \text{RAMP}_\gamma'$. Furthermore, by Lemma 9, $\mathcal{G}$ is 1-Lipschitz with respect to the Hausdorff distance. Thus, by Lemma 8

$$\mathcal{N}(\eta, \text{RAMP}_\gamma', L_2(S)) \leq \mathcal{N}(\eta/3, \mathcal{F}, L_2(S)) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S)). \tag{7}$$

We now proceed to bound the two covering numbers on the right hand side. First, consider $\mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S))$. By Lemma 10, the pseudo-dimension of $\mathcal{G}(f)$ is the same as the pseudo-dimension of $\{x \mapsto \langle w, x \rangle / \gamma \mid w \in \bar{V}\}$, which is exactly k, the dimension of $\bar{V}$. The L_2 covering

number of $\mathcal{G}(f)$ can be bounded by the pseudo-dimension of $\mathcal{G}(f)$ as follows (see, e.g., Bartlett, 2006, Theorem 3.1):

$$\mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S)) \leq C_1 \left(\frac{C_2}{\eta^2}\right)^k. \tag{8}$$

Second, consider $\mathcal{N}(\eta/3, \mathcal{F}, L_2(S))$. Sudakov's minoration theorem (Sudakov 1971, and see also Ledoux and Talagrand, 1991, Theorem 3.18) states that for any $\eta > 0$

$$\ln \mathcal{N}(\eta, \mathcal{F}, L_2(S)) \leq \frac{C}{m\eta^2} \mathbb{E}_s^2 [\sup_{f \in \mathcal{F}} \sum_{i \in [m]} s_i f(x_i)],$$

where $s = (s_1, \ldots, s_m)$ are independent standard normal variables. The right-hand side can be bounded as follows:

$$\gamma \mathbb{E}_s [\sup_{f \in \mathcal{F}} | \sum_{i=1}^{m} s_i f(x_i)|] = \mathbb{E}_s [\sup_{w \in \mathbb{B}_1^d \cap V} |\langle w, \sum_{i=1}^{m} s_i x_i \rangle|]$$

$$\leq \mathbb{E}_s [\| \sum_{i=1}^{m} s_i \mathbb{O}_V x_i \|] \leq \sqrt{\mathbb{E}_s [\| \sum_{i=1}^{m} s_i \mathbb{O}_V x_i \|^2]} = \sqrt{\sum_{i \in [m]} \| \mathbb{O}_V x_i \|^2} = \sqrt{m} B(S).$$

Therefore $\ln \mathcal{N}(\eta, \mathcal{F}, L_2(S)) \leq C \frac{B^2(S)}{\gamma^2 \eta^2}$. Substituting this and Equation (8) for the right-hand side in Equation (7), and adjusting constants, we get

$$\ln \mathcal{N}(\eta, \text{RAMP}_\gamma, L_2(S)) \leq \ln \mathcal{N}(\eta, \text{RAMP}'_\gamma, L_2(S)) \leq C_1 (1 + k \ln(\frac{C_2}{\eta}) + \frac{B^2(S)}{\gamma^2 \eta^2}),$$

To finalize the proof, we plug this inequality into Equation (5) to get

$$\sqrt{m} \mathcal{R}(\text{RAMP}_\gamma, S) \leq C_1 \sum_{i \in [N]} \varepsilon_{i-1} \sqrt{1 + k \ln(C_2/\varepsilon_i) + \frac{B^2(S)}{\gamma^2 \varepsilon_i^2}} + 2\varepsilon_N \sqrt{m}$$

$$\leq C_1 \left(\sum_{i \in [N]} \varepsilon_{i-1} \left(1 + \sqrt{k \ln(C_2/\varepsilon_i)} + \sqrt{\frac{B^2(S)}{\gamma^2 \varepsilon_i^2}} \right) \right) + 2\varepsilon_N \sqrt{m}$$

$$= C_1 \left(\sum_{i \in [N]} 2^{-i+1} + \sqrt{k} \sum_{i \in [N]} 2^{-i+1} \ln(C_2/2^{-i}) + \sum_{i \in [N]} \frac{B(S)}{\gamma} \right) + 2^{-N+1} \sqrt{m}$$

$$\leq C \left(1 + \sqrt{k} + \frac{B(S) \cdot N}{\gamma} \right) + 2^{-N+1} \sqrt{m}.$$

In the last inequality we used the fact that $\sum_i i 2^{-i+1} \leq 4$. Setting $N = \ln(2m)$ we get

$$\mathcal{R}(\text{RAMP}_\gamma, S) \leq \frac{C}{\sqrt{m}} \left(1 + \sqrt{k} + \frac{B(S) \ln(2m)}{\gamma} \right).$$

Taking expectation over both sides, and noting that $\mathbb{E}[B(S)] \leq \sqrt{\mathbb{E}[B^2(S)]} \leq B$, we get

$$\mathcal{R}(\text{RAMP}_\gamma, S) \leq \frac{C}{\sqrt{m}} (1 + \sqrt{k} + \frac{B \ln(2m)}{\gamma}) \leq \sqrt{\frac{O(k + B^2 \ln^2(2m)/\gamma^2)}{m}}.$$

■

Corollary 12 (Sample complexity upper bound) *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Then*

$$m(\varepsilon, \gamma, D) \leq \tilde{O}\left(\frac{k_\gamma(D_X)}{\varepsilon^2}\right).$$

Proof By Proposition 3, we have

$$\ell_0(\mathcal{A}_\gamma, D, m, \delta) \leq \ell_\gamma^*(\mathcal{W}, D) + 2\mathcal{R}_m(\text{RAMP}_\gamma, D) + \sqrt{\frac{14\ln(2/\delta)}{m}}.$$

By definition of $k_\gamma(D_X)$, D_X is $(\gamma^2 k_\gamma, k_\gamma)$-limited. Therefore, by Theorem 11,

$$\mathcal{R}_m(\text{RAMP}_\gamma, D) \leq \sqrt{\frac{O(k_\gamma(D_X))\ln(m)}{m}}.$$

We conclude that

$$\ell_0(\mathcal{A}_\gamma, D, m, \delta) \leq \ell_\gamma^*(\mathcal{W}, D) + \sqrt{\frac{O(k_\gamma(D_X)\ln(m) + \ln(1/\delta))}{m}}.$$

Bounding the second right-hand term by ε, we conclude that $m(\varepsilon, \gamma, D) \leq \tilde{O}(k_\gamma/\varepsilon^2)$. ■

One should note that a similar upper bound can be obtained much more easily under a uniform upper bound on the eigenvalues of the uncentered covariance matrix.[2] However, such an upper bound would not capture the fact that a finite dimension implies a finite sample complexity, regardless of the size of the covariance. If one wants to estimate the sample complexity, then large covariance matrix eigenvalues imply that more examples are required to estimate the covariance matrix from a sample. However, these examples need not be labeled. Moreover, estimating the covariance matrix is not necessary to achieve the sample complexity, since the upper bound holds for any margin-error minimization algorithm.

7. A Distribution-Dependent Lower Bound

The new upper bound presented in Corollary 12 can be tighter than both the norm-only and the dimension-only upper bounds. But does the margin-adapted dimension characterize the true sample complexity of the distribution, or is it just another upper bound? To answer this question, we first need tools for deriving sample complexity lower bounds. Section 7.1 relates fat-shattering with a lower bound on sample complexity. In Section 7.2 we use this result to relate the smallest eigenvalue of a Gram-matrix to a lower bound on sample complexity. In Section 7.3 the family of sub-Gaussian product distributions is presented. We prove a sample-complexity lower bound for this family in Section 7.4.

---

2. This has been pointed out to us by an anonymous reviewer of this manuscript. An upper bound under sub-Gaussianity assumptions can be found in Sabato et al. (2010).

7.1 A Sample Complexity Lower Bound Based on Fat-Shattering

The ability to learn is closely related to the probability of a sample to be shattered, as evident in Vapnik's formulations of learnability as a function of the ε-entropy (Vapnik, 1995). It is well known that the maximal size of a shattered set dictates a sample-complexity upper bound. In the theorem below, we show that for some hypothesis classes it also implies a lower bound. The theorem states that if a sample drawn from a data distribution is fat-shattered with a non-negligible probability, then MEM can fail to learn a good classifier for this distribution.[3] This holds not only for linear classifiers, but more generally for all *symmetric* hypothesis classes. Given a domain X, we say that a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^X$ is symmetric if for all $h \in \mathcal{H}$, we have $-h \in \mathcal{H}$ as well. This clearly holds for the class of linear classifiers $\mathcal{W}$.

Theorem 13 *Let X be some domain, and assume that $\mathcal{H} \subseteq \mathbb{R}^X$ is a symmetric hypothesis class. Let D be a distribution over $X \times \{\pm 1\}$. If the probability of a sample of size m drawn from D_X^m to be γ-shattered at the origin by $\mathcal{W}$ is at least η, then $m(\varepsilon, \gamma, D, \eta/2) \geq \lfloor m/2 \rfloor$ for all $\varepsilon < 1/2 - \ell_\gamma^*(D)$.*

Proof Let $\varepsilon \leq \frac{1}{2} - \ell_\gamma^*(D)$. We show a MEM algorithm $\mathcal{A}$ such that

$$\ell_0(\mathcal{A}_\gamma, D, \lfloor m/2 \rfloor, \eta/2) \geq \frac{1}{2} > \ell_\gamma^*(D) + \varepsilon,$$

thus proving the desired lower bound on $m(\varepsilon, \gamma, D, \eta/2)$.

Assume for simplicity that m is even (otherwise replace m with $m - 1$). Consider two sets $S, \tilde{S} \subseteq X \times \{\pm 1\}$, each of size $m/2$, such that $S_X \cup \tilde{S}_X$ is γ-shattered at the origin by $\mathcal{W}$. Then there exists a hypothesis $h_1 \in \mathcal{H}$ such that the following holds:

- For all $x \in S_X \cup \tilde{S}_X$, $|h_1(x)| \geq \gamma$.

- For all $(x, y) \in S$, $\text{sign}(h_1(x)) = y$.

- For all $(x, y) \in \tilde{S}$, $\text{sign}(h_1(x)) = -y$.

It follows that $\ell_\gamma(h_1, S) = 0$. In addition, let $h_2 = -h_1$. Then $\ell_\gamma(h_2, \tilde{S}) = 0$. Moreover, we have $h_2 \in \mathcal{H}$ due to the symmetry of $\mathcal{H}$. On each point in X, at least one of h_1 and h_2 predict the wrong sign. Thus $\ell_0(h_1, D) + \ell_0(h_2, D) \geq 1$. It follows that for at least one of $i \in \{1, 2\}$, we have $\ell_0(h_i, D) \geq \frac{1}{2}$. Denote the set of hypotheses with a high misclassification error by

$$\mathcal{H}_\otimes = \{h \in \mathcal{H} \mid \ell_0(h, D) \geq \frac{1}{2}\}.$$

We have just shown that if $S_X \cup \tilde{S}_X$ is γ-shattered by $\mathcal{W}$ then at least one of the following holds: (1) $h_1 \in \mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S)$ or (2) $h_2 \in \mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, \tilde{S})$.

Now, consider a MEM algorithm $\mathcal{A}$ such that whenever possible, it returns a hypothesis from $\mathcal{H}_\otimes$. Formally, given the input sample S, if $\mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S) \neq \emptyset$, then

---

3. In contrast, the average Rademacher complexity cannot be used to derive general lower bounds for MEM algorithms, since it is related to the rate of uniform convergence of the entire hypothesis class, while MEM algorithms choose low-error hypotheses (see, e.g., Bartlett et al., 2005).

$\mathcal{A}(S) \in \mathcal{H}_{\otimes} \cap \operatorname{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, S)$. It follows that

$$\mathbb{P}_{S \sim D^{m/2}}[\ell_0(\mathcal{A}(S), D) \geq \tfrac{1}{2}] \geq \mathbb{P}_{S \sim D^{m/2}}[\mathcal{H}_{\otimes} \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, S) \neq \emptyset]$$

$$= \frac{1}{2}(\mathbb{P}_{S \sim D^{m/2}}[\mathcal{H}_{\otimes} \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, S) \neq \emptyset] + \mathbb{P}_{\tilde{S} \sim D^{m/2}}[\mathcal{H}_{\otimes} \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, \tilde{S}) \neq \emptyset])$$

$$\geq \frac{1}{2}(\mathbb{P}_{S, \tilde{S} \sim D^{m/2}}[\mathcal{H}_{\otimes} \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, S) \neq \emptyset \text{ OR } \mathcal{H}_{\otimes} \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_{\gamma}(h, \tilde{S}) \neq \emptyset])$$

$$\geq \frac{1}{2}\mathbb{P}_{S, \tilde{S} \sim D^{m/2}}[S_X \cup \tilde{S}_X \text{ is } \gamma\text{-shattered at the origin}].$$

The last inequality follows from the argument above regarding h_1 and h_2. The last expression is simply half the probability that a sample of size m from D_X is shattered. By assumption, this probability is at least η. Thus we conclude that $\mathbb{P}_{S \sim D^{m/2}}[\ell_0(\mathcal{A}(S), D) \geq \tfrac{1}{2}] \geq \eta/2$. It follows that $\ell_0(\mathcal{A}_{\gamma}, D, m/2, \eta/2) \geq \tfrac{1}{2}$. ∎

As a side note, it is interesting to observe that Theorem 13 does not hold in general for non-symmetric hypothesis classes. For example, assume that the domain is $X = [0, 1]$, and the hypothesis class is the set of all functions that label a finite number of points in $[0, 1]$ by $+1$ and the rest by -1. Consider learning using MEM, when the distribution is uniform over $[0, 1]$, and all the labels are -1. For any $m > 0$ and $\gamma \in (0, 1)$, a sample of size m is γ-shattered at the origin with probability 1. However, any learning algorithm that returns a hypothesis from the hypothesis class will incur zero error on this distribution. Thus, shattering alone does not suffice to ensure that learning is hard.

7.2 A Sample Complexity Lower Bound with Gram-Matrix Eigenvalues

We now return to the case of homogeneous linear classifiers, and link high-probability fat-shattering to properties of the distribution. First, we present an equivalent and simpler characterization of fat-shattering for linear classifiers. We then use it to provide a sufficient condition for the fat-shattering of a sample, based on the smallest eigenvalue of its Gram matrix.

Theorem 14 *Let $\mathbb{X} \in \mathbb{R}^{m \times d}$ be the matrix of a set of size m in $\mathbb{R}^d$. The set is γ-shattered at the origin by $\mathcal{W}$ if and only if $\mathbb{X}\mathbb{X}^T$ is invertible and for all $y \in \{\pm 1\}^m$, $y^T(\mathbb{X}\mathbb{X}^T)^{-1}y \leq \gamma^{-2}$.*

To prove Theorem 14 we require two auxiliary lemmas. The first lemma, stated below, shows that for convex function classes, γ-shattering can be substituted with shattering with exact γ-margins.

Lemma 15 *Let $\mathcal{F} \subseteq \mathbb{R}^X$ be a class of functions, and assume that $\mathcal{F}$ is convex, that is*

$$\forall f_1, f_2 \in \mathcal{F}, \forall \lambda \in [0, 1], \quad \lambda f_1 + (1 - \lambda)f_2 \in \mathcal{F}.$$

If $S = \{x_1, \ldots, x_m\} \subseteq X$ is γ-shattered by $\mathcal{F}$ with witness $r \in \mathbb{R}^m$, then for every $y \in \{\pm 1\}^m$ there is an $f \in \mathcal{F}$ such that for all $i \in [m]$, $y[i](f(x_i) - r[i]) = \gamma$.

The proof of this lemma is provided in Appendix A.4. The second lemma that we use allows converting the representation of the Gram-matrix to a different feature space, while keeping the separation properties intact. For a matrix $\mathbb{M}$, denote its pseudo-inverse by $\mathbb{M}^+$.

Lemma 16 *Let $\mathbb{X} \in \mathbb{R}^{m \times d}$ be a matrix such that $\mathbb{X}\mathbb{X}^T$ is invertible, and let $\mathbb{Y} \in \mathbb{R}^{m \times k}$ such that $\mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. Let $r \in \mathbb{R}^m$ be some real vector. If there exists a vector $\widetilde{w} \in \mathbb{R}^k$ such that $\mathbb{Y}\widetilde{w} = r$, then there exists a vector $w \in \mathbb{R}^d$ such that $\mathbb{X}w = r$ and $\|w\| = \|\mathbb{Y}^T(\mathbb{Y}^T)^+\widetilde{w}\| \le \|\tilde{w}\|$.*

Proof Denote $\mathbb{K} = \mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. Let $\mathbb{S} = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{X}$ and let $w = \mathbb{S}^T\widetilde{w}$. We have $\mathbb{X}w = \mathbb{X}\mathbb{S}^T\widetilde{w} = \mathbb{X}\mathbb{X}^T\mathbb{K}^{-1}\mathbb{Y}\widetilde{w} = \mathbb{Y}\widetilde{w} = r$. In addition, $\|w\|^2 = w^Tw = \widetilde{w}^T\mathbb{S}\mathbb{S}^T\widetilde{w}$. By definition of $\mathbb{S}$,

$$\mathbb{S}\mathbb{S}^T = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{X}\mathbb{X}^T\mathbb{K}^{-1}\mathbb{Y} = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{Y} = \mathbb{Y}^T(\mathbb{Y}\mathbb{Y}^T)^{-1}\mathbb{Y} = \mathbb{Y}^T(\mathbb{Y}^T)^+.$$

Denote $\mathbb{O} = \mathbb{Y}^T(\mathbb{Y}^T)^+$. $\mathbb{O}$ is an orthogonal projection matrix: by the properties of the pseudo-inverse, $\mathbb{O} = \mathbb{O}^T$ and $\mathbb{O}^2 = \mathbb{O}$. Therefore $\|w\|^2 = \widetilde{w}^T\mathbb{S}\mathbb{S}^T\widetilde{w} = \widetilde{w}^T\mathbb{O}\widetilde{w} = \widetilde{w}^T\mathbb{O}\mathbb{O}^T\widetilde{w} = \|\mathbb{O}\widetilde{w}\|^2 \le \|\widetilde{w}\|^2$. ∎

Proof [of Theorem 14] We prove the theorem for 1-shattering. The case of γ-shattering follows by rescaling X appropriately. Let $\mathbb{X}\mathbb{X}^T = \mathbb{U}\Lambda\mathbb{U}^T$ be the SVD of $\mathbb{X}\mathbb{X}^T$, where $\mathbb{U}$ is an orthogonal matrix and Λ is a diagonal matrix. Let $\mathbb{Y} = \mathbb{U}\Lambda^{\frac{1}{2}}$. We have $\mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. We show that the specified conditions are sufficient and necessary for the shattering of the set:

1. Sufficient: If $\mathbb{X}\mathbb{X}^T$ is invertible, then Λ is invertible, thus so is $\mathbb{Y}$. For any $y \in \{\pm1\}^m$, Let $w_y = \mathbb{Y}^{-1}y$. Then $\mathbb{Y}w_y = y$. By Lemma 16, there exists a separator w such that $\mathbb{X}w = y$ and $\|w\| \le \|w_y\| = \sqrt{y^T(\mathbb{Y}\mathbb{Y}^T)^{-1}y} = \sqrt{y^T(\mathbb{X}\mathbb{X}^T)^{-1}y} \le 1$.

2. Necessary: If $\mathbb{X}\mathbb{X}^T$ is not invertible then the vectors in S are linearly dependent, thus S cannot be shattered using linear separators (see, e.g., Vapnik, 1995). The first condition is therefore necessary. Assume S is 1-shattered at the origin and show that the second condition necessarily holds. By Lemma 15, for all $y \in \{\pm1\}^m$ there exists a $w_y \in \mathbb{B}_1^d$ such that $\mathbb{X}w_y = y$. Thus by Lemma 16 there exists a $\widetilde{w}_y$ such that $\mathbb{Y}\widetilde{w}_y = y$ and $\|\widetilde{w}_y\| \le \|w_y\| \le 1$. $\mathbb{X}\mathbb{X}^T$ is invertible, thus so is $\mathbb{Y}$. Therefore $\widetilde{w}_y = \mathbb{Y}^{-1}y$. Thus $y^T(\mathbb{X}\mathbb{X}^T)^{-1}y = y^T(\mathbb{Y}\mathbb{Y}^T)^{-1}y = \|\widetilde{w}_y\| \le 1$.

∎

We are now ready to provide a sufficient condition for fat-shattering based on the smallest eigenvalue of the Gram matrix.

Corollary 17 *Let $\mathbb{X} \in \mathbb{R}^{m \times d}$ be the matrix of a set of size m in $\mathbb{R}^d$. If $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \ge m\gamma^2$ then the set is γ-shattered at the origin by $\mathcal{W}$.*

Proof If $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \ge m\gamma^2$ then $\mathbb{X}\mathbb{X}^T$ is invertible and $\lambda_{\max}((\mathbb{X}\mathbb{X}^T)^{-1}) \le (m\gamma^2)^{-1}$. For any $y \in \{\pm1\}^m$ we have $\|y\| = \sqrt{m}$ and

$$y^T(\mathbb{X}\mathbb{X}^T)^{-1}y \le \|y\|^2\lambda_{\max}((\mathbb{X}\mathbb{X}^T)^{-1}) \le m(m\gamma^2)^{-1} = \gamma^{-2}.$$

By Theorem 14 the sample is γ-shattered at the origin. ∎

Corollary 17 generalizes the requirement of linear independence for shattering with no margin: A set of vectors is shattered with no margin if the vectors are linearly independent, that is if $\lambda_{\min} > 0$. The corollary shows that for γ-fat-shattering, we can require instead $\lambda_{\min} \ge m\gamma^2$. We can now

conclude that if it is highly probable that the smallest eigenvalue of the sample Gram matrix is large, then MEM might fail to learn a good classifier for the given distribution. This is formulated in the following theorem.

Theorem 18 *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Let $m > 0$ and let $\mathbb{X}$ be the matrix of a sample drawn from D_X^m. Let $\eta = \mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2]$. Then for all $\varepsilon < 1/2 - \ell_\gamma^*(D)$, $m(\varepsilon,\gamma,D,\eta/2) \geq \lfloor m/2 \rfloor$.*

The proof of the theorem is immediate by combining Theorem 13 and Corollary 17.

Theorem 18 generalizes the case of learning a linear separator without a margin: If a sample of size m is linearly independent with high probability, then there is no hope of using $m/2$ points to predict the label of the other points. The theorem extends this observation to the case of learning with a margin, by requiring a stronger condition than just linear independence of the points in the sample.

Recall that our upper-bound on the sample complexity from Section 6 is $\tilde{O}(k_\gamma)$. We now define the family of sub-Gaussian product distributions, and show that for this family, the lower bound that can be deduced from Theorem 18 is also linear in k_γ.

7.3 Sub-Gaussian Distributions

In order to derive a lower bound on distribution-specific sample complexity in terms of the covariance of $X \sim D_X$, we must assume that X is not too heavy-tailed. This is because for any data distribution there exists another distribution which is almost identical and has the same sample complexity, but has arbitrarily large covariance values. This can be achieved by mixing the original distribution with a tiny probability for drawing a vector with a huge norm. We thus restrict the discussion to multidimensional sub-Gaussian distributions. This ensures light tails of the distribution in all directions, while still allowing a rich family of distributions, as we presently see. Sub-Gaussianity is defined for scalar random variables as follows (see, e.g., Buldygin and Kozachenko, 1998).

Definition 19 (Sub-Gaussian random variables) *A random variable $X \in \mathbb{R}$ is* sub-Gaussian with moment B, *for $B \geq 0$, if*

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

In this work we further say that X is sub-Gaussian with relative moment $\rho > 0$ *if X is sub-Gaussian with moment $\rho\sqrt{\mathbb{E}[X^2]}$, that is,*

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

Note that a sub-Gaussian variable with moment B and relative moment ρ is also sub-Gaussian with moment B' and relative moment ρ' for any $B' \geq B$ and $\rho' \geq \rho$.

The family of sub-Gaussian distributions is quite extensive: For instance, it includes any bounded, Gaussian, or Gaussian-mixture random variable with mean zero. Specifically, if X is a mean-zero Gaussian random variable, $X \sim N(0, \sigma^2)$, then X is sub-Gaussian with relative moment 1 and the inequalities in the definition above hold with equality. As another example, if X is a uniform random variable over $\{\pm b\}$ for some $b \geq 0$, then X is sub-Gaussian with relative moment 1, since

$$\mathbb{E}[\exp(tX)] = \frac{1}{2}(\exp(tb) + \exp(-tb)) \leq \exp(t^2 b^2/2) = \exp(t^2\mathbb{E}[X^2]/2). \tag{9}$$

Let $\mathbb{B} \in \mathbb{R}^{d \times d}$ be a symmetric PSD matrix. A random vector $X \in \mathbb{R}^d$ is a *sub-Gaussian random vector* with moment matrix $\mathbb{B}$ if for all $u \in \mathbb{R}^d$, $\mathbb{E}[\exp(\langle u, X \rangle)] \leq \exp(\langle \mathbb{B}u, u \rangle / 2)$. The following lemma provides a useful connection between the trace of the sub-Gaussian moment matrix and the moment-generating function of the squared norm of the random vector. The proof is given in Appendix A.5.

Lemma 20 *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with moment matrix $\mathbb{B}$. Then for all $t \in (0, \frac{1}{4\lambda_{\max}(\mathbb{B})}]$, $\mathbb{E}[\exp(t\|X\|^2)] \leq \exp(2t \cdot \mathrm{trace}(\mathbb{B}))$.*

Our lower bound holds for the family of sub-Gaussian product distributions, defined as follows.

Definition 21 (Sub-Gaussian product distributions) *A distribution D_X over $\mathbb{R}^d$ is a* sub-Gaussian product distribution *with moment B and relative moment ρ if there exists some orthonormal basis $a_1, \ldots, a_d \in \mathbb{R}^d$, such that for $X \sim D_X$, $\langle a_i, X \rangle$ are independent sub-Gaussian random variables, each with moment B and relative moment ρ.*

Note that a sub-Gaussian product distribution has mean zero, thus its covariance matrix is equal to its uncentered covariance matrix. For any fixed $\rho \geq 0$, we denote by $\mathcal{D}_\rho^{\mathrm{sg}}$ the family of all sub-Gaussian product distributions with relative moment ρ, in arbitrary dimension. For instance, all multivariate Gaussian distributions and all uniform distributions on the corners of a centered hyper-rectangle are in $\mathcal{D}_1^{\mathrm{sg}}$. All uniform distributions over a full centered hyper-rectangle are in $\mathcal{D}_{3/2}^{\mathrm{sg}}$. Note that if $\rho_1 \leq \rho_2$, $\mathcal{D}_{\rho_1}^{\mathrm{sg}} \subseteq \mathcal{D}_{\rho_2}^{\mathrm{sg}}$.

We will provide a lower bound for all distributions in $\mathcal{D}_\rho^{\mathrm{sg}}$. This lower bound is linear in the margin-adapted dimension of the distribution, thus it matches the upper bound provided in Corollary 12. The constants in the lower bound depend only on the value of ρ, which we regard as a constant.

7.4 A Sample-Complexity Lower Bound for Sub-Gaussian Product Distributions

As shown in Section 7.2, to obtain a sample complexity lower bound it suffices to have a lower bound on the value of the smallest eigenvalue of a random Gram matrix. The distribution of the smallest eigenvalue of a random Gram matrix has been investigated under various assumptions. The cleanest results are in the asymptotic case where the sample size and the dimension approach infinity, the ratio between them approaches a constant, and the coordinates of each example are identically distributed.

Theorem 22 (Bai and Silverstein 2010, Theorem 5.11) *Let $\{\mathbb{X}_i\}_{i=1}^\infty$ be a series of matrices of sizes $m_i \times d_i$, 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 $\lim_{i \to \infty} \lambda_{\min}(\frac{1}{d_i} \mathbb{X}_i \mathbb{X}_i^T) = \sigma^2(1 - \sqrt{\beta})^2$.*

This asymptotic limit can be used to approximate an asymptotic lower bound on $m(\varepsilon, \gamma, D)$, if D_X is a product distribution of i.i.d. random variables with mean zero, variance σ^2, and finite fourth moment. Let $\mathbb{X} \in \mathbb{R}^{m \times d}$ be the matrix of a sample of size m drawn from D_X. We can find $m = m_\circ$ such that $\lambda_{m_\circ}(\mathbb{X}\mathbb{X}^T) \approx \gamma^2 m_\circ$, and use Theorem 18 to conclude that $m(\varepsilon, \gamma, D) \geq m_\circ/2$. If d and m are large enough, we have by Theorem 22 that for $\mathbb{X}$ drawn from D_X^m:

$$\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \approx d\sigma^2(1 - \sqrt{m/d})^2 = \sigma^2(\sqrt{d} - \sqrt{m})^2.$$

Solving the equality $\sigma^2(\sqrt{d} - \sqrt{m_\circ})^2 = m_\circ \gamma^2$ we get $m_\circ = d/(1 + \gamma/\sigma)^2$. The margin-adapted dimension for D_X is $k_\gamma \approx d/(1 + \gamma^2/\sigma^2)$, thus $\frac{1}{2}k_\gamma \leq m_\circ \leq k_\gamma$. In this case, then, the sample complexity lower bound is indeed the same order as k_γ, which controls also the upper bound in Corollary 12. However, this is an asymptotic analysis, which holds for a highly limited set of distributions. Moreover, since Theorem 22 holds asymptotically for each distribution separately, we cannot use it to deduce a uniform finite-sample lower bound for families of distributions.

For our analysis we require *finite-sample* bounds for the smallest eigenvalue of a random Gramm-matrix. Rudelson and Vershynin (2009, 2008) provide such finite-sample lower bounds for distributions which are products of identically distributed sub-Gaussians. In Theorem 23 below we provide a new and more general result, which holds for any sub-Gaussian product distribution. The proof of Theorem 23 is provided in Appendix A.6. Combining Theorem 23 with Theorem 18 above we prove the lower bound, stated in Theorem 24 below.

Theorem 23 *For any $\rho > 0$ and $\delta \in (0,1)$ there are $\beta > 0$ and $C > 0$ such that the following holds. For any $D_X \in \mathcal{D}_\rho^{sg}$ with covariance matrix $\Sigma \leq I$, and for any $m \leq \beta \cdot \text{trace}(\Sigma) - C$, if $\mathbb{X}$ is the $m \times d$ matrix of a sample drawn from D_X^m, then*

$$\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m] \geq \delta.$$

Theorem 24 (Sample complexity lower bound for distributions in $\mathcal{D}_\rho^{\text{sg}}$) *For any $\rho > 0$ there are constants $\beta > 0, C \geq 0$ such that for any D with $D_X \in \mathcal{D}_\rho^{sg}$, for any $\gamma > 0$ and for any $\varepsilon < \frac{1}{2} - \ell_\gamma^*(D)$,*

$$m(\varepsilon, \gamma, D, 1/4) \geq \beta k_\gamma(D_X) - C.$$

Proof Assume w.l.o.g. that the orthonormal basis $a_1, \ldots, a_d$ of independent sub-Gaussian directions of D_X, defined in Definition 21, is the natural basis $e_1, \ldots, e_d$. Define $\lambda_i = \mathbb{E}_{X \sim D_X}[X[i]^2]$, and assume w.l.o.g. $\lambda_1 \geq \ldots \geq \lambda_d > 0$. Let $\mathbb{X}$ be the $m \times d$ matrix of a sample drawn from D_X^m. Fix $\delta \in (0,1)$, and let β and C be the constants for ρ and δ in Theorem 23. Throughout this proof we abbreviate $k_\gamma \triangleq k_\gamma(D_X)$. Let $m \leq \beta(k_\gamma - 1) - C$. We would like to use Theorem 23 to bound $\lambda_{\min}(\mathbb{X}\mathbb{X}^T)$ with high probability, so that Theorem 18 can be applied to get the desired lower bound. However, Theorem 23 holds only if $\Sigma \leq I$. Thus we split to two cases—one in which the dimensionality controls the lower bound, and one in which the norm controls it. The split is based on the value of λ_{k_γ}.

- Case I: Assume $\lambda_{k_\gamma} \geq \gamma^2$. Then $\forall i \in [k_\gamma], \lambda_i \geq \gamma^2$. By our assumptions on D_X, for all $i \in [d]$ the random variable $X[i]$ is sub-Gaussian with relative moment ρ. Consider the random variables $Z[i] = X[i]/\sqrt{\lambda_i}$ for $i \in [k_\gamma]$. $Z[i]$ is also sub-Gaussian with relative moment ρ, and $\mathbb{E}[Z[i]^2] = 1$. Consider the product distribution of $Z[1], \ldots, Z[k_\gamma]$, and let Σ' be its covariance matrix. We have $\Sigma' = I_{k_\gamma}$, and $\text{trace}(\Sigma') = k_\gamma$. Let $\mathbb{Z}$ be the matrix of a sample of size m drawn from this distribution. By Theorem 23, $\mathbb{P}[\lambda_{\min}(\mathbb{Z}\mathbb{Z}^T) \geq m] \geq \delta$, which is equivalent to

$$\mathbb{P}[\lambda_{\min}(\mathbb{X} \cdot \text{diag}(1/\lambda_1, \ldots, 1/\lambda_{k_\gamma}, 0, \ldots, 0) \cdot \mathbb{X}^T) \geq m] \geq \delta.$$

  Since $\forall i \in [k_\gamma], \lambda_i \geq \gamma^2$, we have $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$.

- Case II: Assume $\lambda_{k_\gamma} < \gamma^2$. Then $\lambda_i < \gamma^2$ for all $i \in \{k_\gamma, \ldots, d\}$. Consider the random variables $Z[i] = X[i]/\gamma$ for $i \in \{k_\gamma, \ldots, d\}$. $Z[i]$ is sub-Gaussian with relative moment ρ and $\mathbb{E}[Z[i]^2] \leq 1$.

Consider the product distribution of $Z[k_\gamma], \ldots, Z[d]$, and let Σ' be its covariance matrix. We have $\Sigma' < I_{d-k_\gamma+1}$. By the minimality in Equation (4) we also have $\text{trace}(\Sigma') = \frac{1}{\gamma^2} \sum_{i=k_\gamma}^{d} \lambda_i \geq k_\gamma - 1$. Let $\mathbb{Z}$ be the matrix of a sample of size m drawn from this product distribution. By Theorem 23, $\mathbb{P}[\lambda_{\min}(\mathbb{Z}\mathbb{Z}^T) \geq m] \geq \delta$. Equivalently,

$$\mathbb{P}[\lambda_{\min}(\mathbb{X} \cdot \text{diag}(0, \ldots, 0, 1/\gamma^2, \ldots, 1/\gamma^2) \cdot \mathbb{X}^T) \geq m] \geq \delta,$$

therefore $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$.

In both cases $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$. This holds for any $m \leq \beta(k_\gamma - 1) - C$, thus by Theorem 18 $m(\varepsilon, \gamma, D, \delta/2) \geq \lfloor (\beta(k_\gamma - 1) - C)/2 \rfloor$ for $\varepsilon < 1/2 - \ell_\gamma^*(D)$. We finalize the proof by setting $\delta = \frac{1}{2}$ and adjusting β and C. ∎

8. On the Limitations of the Covariance Matrix

We have shown matching upper and lower bounds for the sample complexity of learning with MEM, for any sub-Gaussian product distribution with a bounded relative moment. This shows that the margin-adapted dimension fully characterizes the sample complexity of learning with MEM for such distributions. What properties of a distribution play a role in determining the sample complexity for general distributions? In the following theorem we show that these properties must include more than the covariance matrix of the distribution, even when assuming sub-Gaussian tails and bounded relative moments.

Theorem 25 *For any integer $d > 1$, there exist two distributions D and P over $\mathbb{R}^d \times \{\pm 1\}$ with identical covariance matrices, such that for any $\varepsilon, \delta \in (0, \frac{1}{4})$, $m(\varepsilon, 1, P, \delta) \geq \Omega(d)$ while $m(\varepsilon, 1, D, \delta) \leq \lceil \log_2(1/\delta) \rceil$. Both D_X and P_X are sub-Gaussian random vectors, with a relative moment of $\sqrt{2}$ in all directions.*

Proof Let D_a and D_b be distributions over $\mathbb{R}^d$ such that D_a is uniform over $\{\pm 1\}^d$ and D_b is uniform over $\{\pm 1\} \times \{0\}^{d-1}$. Let D_X be a balanced mixture of D_a and D_b. Let P_X be uniform over $\{\pm 1\} \times \{\frac{1}{\sqrt{2}}\}^{d-1}$. For both D and P, let $\mathbb{P}[Y = \langle e_1, X \rangle] = 1$. The covariance matrix of D_X and P_X is $\text{diag}(1, \frac{1}{2}, \ldots, \frac{1}{2})$, thus $k_1(D_X) = k_1(P_X) \geq \Omega(d)$.

By Equation (9), P_X, D_a and D_b are all sub-Gaussian product distribution with relative moment 1, thus also with moment $\sqrt{2} > 1$. The projection of D_X along any direction $u \in \mathbb{R}^d$ is sub-Gaussian with relative moment $\sqrt{2}$ as well, since

$$\mathbb{E}_{X \sim D_X}[\exp(\langle u, X \rangle)] = \frac{1}{2}(\mathbb{E}_{X \sim D_a}[\exp(\langle u, X \rangle)] + \mathbb{E}_{X \sim D_b}[\exp(\langle u, X \rangle)])$$

$$= \frac{1}{2}(\prod_{i \in [d]} (\exp(u_i) + \exp(-u_i))/2 + (\exp(u_1) + \exp(-u_1))/2)$$

$$\leq \frac{1}{2}(\prod_{i \in [d]} \exp(u_i^2/2) + \exp(u_1^2/2)) \leq \exp(\|u\|^2/2) \leq \exp((\|u\|^2 + u_1^2)/2)$$

$$= \exp(\mathbb{E}_{X \sim D_X}[\langle u, X \rangle^2]).$$

For P we have by Theorem 24 that for any $\varepsilon \leq \frac{1}{4}$, $m(\varepsilon, 1, P, \frac{1}{4}) \geq \Omega(k_1(P_X)) \geq \Omega(d)$. In contrast, any MEM algorithm $\mathcal{A}_1$ will output the correct separator for D whenever the sample has at least one point drawn from D_b. This is because the separator e_1 is the only $w \in \mathbb{B}_1^d$ that classifies this point with zero 1-margin errors. Such a point exists in a sample of size m with probability $1 - 2^{-m}$. Therefore $\ell_0(\mathcal{A}_1, D, m, 1/2^m) = 0$. It follows that for all $\varepsilon > 0$, $m(\varepsilon, 1, D, \delta) \leq \lceil \log_2(1/\delta) \rceil$. ∎

9. Conclusions

Corollary 12 and Theorem 24 together provide a tight characterization of the sample complexity of any sub-Gaussian product distribution with a bounded relative moment. Formally, fix $\rho > 0$. For any D such that $D_X \in \mathcal{D}_\rho^{\text{sg}}$, and for any $\gamma > 0$ and $\varepsilon \in (0, \frac{1}{2} - \ell_\gamma^*(D))$

$$\Omega(k_\gamma(D_X)) \leq m(\varepsilon, \gamma, D) \leq \tilde{O}\left(\frac{k_\gamma(D_X)}{\varepsilon^2}\right). \tag{10}$$

The upper bound holds uniformly for all distributions, and the constants in the lower bound depend only on ρ. This result shows that the true sample complexity of learning each of these distributions with MEM is characterized by the margin-adapted dimension. An interesting conclusion can be drawn as to the influence of the conditional distribution of labels $D_{Y|X}$: Since Equation (10) 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 note that the upper bound that we have proved involves logarithmic factors which might not be necessary. There are upper bounds that depend on the margin alone and on the dimension alone without logarithmic factors. On the other hand, in our bound, which combines the two quantities, there is a logarithmic dependence which stems from the margin component of the bound. It might be possible to tighten the bound and remove the logarithmic dependence.

Equation (10) can be used to easily characterize the sample complexity behavior for interesting distributions, to compare L_2 margin minimization to other learning methods, and to improve certain active learning strategies. We elaborate on each of these applications in the following examples.

Example 1 (Gaps between L_1 and L_2 regularization in the presence of irrelevant features)
Ng (2004) 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 $O(\ln(d))$ for an L_1-based learning rule (Zhang, 2002). 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. In contrast, using our results we can easily establish a lower bound of $\Omega(d)$ for many specific distributions with a bounded $\|X\|_\infty$ and $Y = sign(X[i])$ for some i. For instance, if each coordinate is a bounded independent sub-Gaussian random variable with a bounded relative moment, we have $k_1 = \lceil d/2 \rceil$ and Theorem 24 implies a lower bound of $\Omega(d)$ on the L_2 sample complexity.*

Example 2 (Gaps between generative and discriminative learning for a Gaussian mixture) *Let there be two classes, each drawn from a unit-variance spherical Gaussian in $\mathbb{R}^d$ with a large distance $2v \gg 1$ between the class means, such that $d \gg v^4$. 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 $D_X \in \mathcal{D}_1^{sg}$. 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 required sample complexity 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 $\varepsilon > 0$, 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.

Example 3 (Active learning) *In active learning, there is an abundance of unlabeled examples, but labels are costly, and the active learning algorithm needs to decide which labels to query based on the labels seen so far. A popular approach to active learning involves estimating the current set of possible classifiers using sample complexity upper bounds (see, e.g., Balcan et al., 2009; Beygelzimer et al., 2010). Without any distribution-specific information, only general distribution-free upper bounds can be used. However, since there is an abundance of unlabeled examples, the active learner can use these to estimate tighter distribution-specific upper bounds. In the case of linear classifiers, the margin-adapted dimension can be calculated from the uncentered covariance matrix of the distribution, which can be easily estimated from unlabeled data. Thus, our sample complexity upper bounds can be used to improve the active learner's label complexity. Moreover, the lower bound suggests that any further improvement of such active learning strategies would require more information other than the distribution's covariance matrix.*

To summarize, we have shown that the true sample complexity of large-margin learning of each of a rich family of distributions is characterized by the margin-adapted dimension. Characterizing the true sample complexity allows a better comparison between this learning approach and other algorithms, and has many potential applications. The challenge of characterizing the true sample complexity extends to any distribution and any learning approach. Theorem 25 shows that other properties but the covariance matrix must be taken into account for general distributions. We believe that obtaining answers to these questions is of great importance, both to learning theory and to learning applications.

Acknowledgments

Appendix A. Proofs Omitted from the Text

In this appendix we give detailed proofs which were omitted from the text.

A.1 Proof of Proposition 3

Proof Let $w^* \in \mathrm{argmin}_{w \in \mathbb{B}_1^d} \ell_\gamma(w, D)$. By Equation (2), with probability $1 - \delta/2$

$$\mathrm{ramp}_\gamma(\mathcal{A}_\gamma(S), D) \leq \mathrm{ramp}_\gamma(\mathcal{A}_\gamma(S), S) + 2\mathcal{R}_m(\mathrm{RAMP}_\gamma, D) + \sqrt{\frac{8\ln(2/\delta)}{m}}.$$

Set $h^* \in \mathcal{H}$ such that $\ell_\gamma(h^*, D) = \ell_\gamma^*(\mathcal{H}, D)$. We have

$$\mathrm{ramp}_\gamma(\mathcal{A}_\gamma(S), S) \leq \ell_\gamma(\mathcal{A}_\gamma(S), S) \leq \ell_\gamma(h^*, S).$$

The first inequality follows since the ramp loss is upper bounded by the margin loss. The second inequality follows since $\mathcal{A}$ is a MEM algorithm. Now, by Hoeffding's inequality, since the range of ramp_γ is in $[0,1]$, with probability at least $1 - \delta/2$

$$\ell_\gamma(h^*, S) \leq \ell_\gamma(h^*, D) + \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

It follows that with probability $1 - \delta$

$$\mathrm{ramp}_\gamma(\mathcal{A}_\gamma(S), D) \leq \ell_\gamma^*(\mathcal{H}, D) + 2\mathcal{R}_m(\mathrm{RAMP}_\gamma, D) + \sqrt{\frac{14\ln(2/\delta)}{m}}. \tag{11}$$

We have $\ell_0 \leq \mathrm{ramp}_\gamma$. Combining this with Equation (11) we conclude Equation (3). ∎

A.2 Proof of Lemma 9

Proof [of Lemma 9] For a function $f : X \to \mathbb{R}$ and a $z \in Z$, define the function $G[f, z]$ by

$$\forall x \in X, \quad G[f, z](x) = [\![f(x) + z(x)]\!] - f(x).$$

Let $f_1, f_2 \in \mathbb{R}^X$ be two functions, and let $g_1 = G[f_1, z] \in \mathcal{G}(f_1)$ for some $w_b \in \bar{V}$. Then, since $G[f_2, z] \in \mathcal{G}(f_2)$, we have

$$\inf_{g_2 \in \mathcal{G}(f_2)} \|g_1 - g_2\|_{L_2(S)} \leq \|G[f_1, z] - G[f_2, z]\|.$$

Now, for all $x \in \mathbb{R}$,

$$\begin{aligned}
|G[f_1, z](x) - G[f_2, z](x)| &= |[\![f_1(x) + z(x)]\!] - f_1(x) - [\![f_2(x) + z(x)]\!] + f_2(x)| \\
&\leq |f_1(x) - f_2(x)|.
\end{aligned}$$

Thus, for any $S \subseteq X$,

$$\begin{aligned}
\|G[f_1, z] - G[f_2, z]\|_{L_2(S)}^2 &= \mathbb{E}_{X \sim S}(G[f_1, z](X) - G[f_2, z](X))^2 \\
&\leq \mathbb{E}_{X \sim S}(f_1(X) - f_2(X))^2 = \|f_1 - f_2\|_{L_2(S)}^2.
\end{aligned}$$

It follows that $\inf_{g_2 \in \mathcal{G}(f_2)} \|g_1 - g_2\|_{L_2(S)} \leq \|f_1 - f_2\|_{L_2(S)}$. This holds for any $g_1 \in \mathcal{G}(f_1)$, thus $\Delta_H(\mathcal{G}(f_1), \mathcal{G}(f_2)) \leq \|f_1 - f_2\|_{L_2(S)}$. ∎

A.3 Proof of Lemma 10

Proof [of Lemma 10] Let k be the pseudo-dimension of $G(f)$, and let $\{x_1,\ldots,x_k\} \subseteq X$ be a set which is pseudo-shattered by $G(f)$. We show that the same set is pseudo-shattered by Z as well, thus proving the lemma. Since $G(f)$ is pseudo-shattered, there exists a vector $r \in \mathbb{R}^k$ such that for all $y \in \{\pm 1\}^k$ there exists a $g_y \in G(f)$ such that $\forall i \in [m], \operatorname{sign}(g_y(x_i) - r[i]) = y[i]$. Therefore for all $y \in \{\pm 1\}^k$ there exists a $z_y \in Z$ such that

$$\forall i \in [k], \operatorname{sign}(\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i]) = y[i].$$

By considering the case $y[i] = 1$, we have

$$0 < \llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] \le 1 - f(x_i) - r[i].$$

By considering the case $y[i] = -1$, we have

$$0 > \llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] \ge -f(x_i) - r[i].$$

Therefore $0 < f(x_i) + r[i] < 1$. Now, let $y \in \{\pm 1\}^k$ and consider any $i \in [k]$. If $y[i] = 1$ then

$$\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] > 0$$

It follows that

$$\llbracket f(x_i) + z_y(x_i) \rrbracket > f(x_i) + r[i] > 0,$$

thus

$$f(x_i) + z_y(x_i) > f(x_i) + r[i].$$

In other words, $\operatorname{sign}(z_y(x_i) - r[i]) = 1 = y[i]$. If $y[i] = -1$ then

$$\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] < 0.$$

It follows that

$$\llbracket f(x_i) + z_y(x_i) \rrbracket < f(x_i) + r[i] < 1,$$

thus

$$f(x_i) + z_y(x_i) < f(x_i) + r[i].$$

in other words, $\operatorname{sign}(z_y(x_i) - r[i]) = -1 = y[i]$. We conclude that Z shatters $\{x_1,\ldots,x_k\}$ as well, using the same vector $r \in \mathbb{R}^k$. Thus the pseudo-dimension of Z is at least k. ∎

A.4 Proof of Lemma 15

To prove Lemma 15, we first prove the following lemma. Denote by $\operatorname{conv}(A)$ the convex hull of a set A.

Lemma 26 *Let $\gamma > 0$. For each $y \in \{\pm 1\}^m$, select $r_y \in \mathbb{R}^m$ such that for all $i \in [m]$, $r_y[i]y[i] \ge \gamma$. Let $R = \{r_y \in \mathbb{R}^m \mid y \in \{\pm 1\}^m\}$. Then $\{\pm \gamma\}^m \subseteq \operatorname{conv}(R)$.*

Proof We will prove the claim by induction on the dimension m. For the base case, if $m = 1$, we have $R = \{a, b\} \subseteq \mathbb{R}$ where $a \leq -\gamma$ and $b \geq \gamma$. Clearly, $\text{conv}(R) = [a, b]$, and $\pm\gamma \in [a, b]$.

For the inductive step, assume the lemma holds for $m - 1$. For a vector $t \in \mathbb{R}^m$, denote by $\bar{t}$ its projection $(t[1], \ldots, t[m-1])$ on $\mathbb{R}^{m-1}$. Similarly, for a set of vectors $S \subseteq \mathbb{R}^m$, let $\bar{S} = \{\bar{s} \mid s \in S\} \subseteq \mathbb{R}^{m-1}$. Define $Y_+ = \{\pm 1\}^{m-1} \times \{+1\}$ and $Y_- = \{\pm 1\}^{m-1} \times \{-1\}$. Let $R_+ = \{r_y \mid y \in Y_+\}$, and similarly for R_-. Then the induction hypothesis holds for $\bar{R}_+$ and $\bar{R}_-$ with dimension $m - 1$. Let $z \in \{\pm\gamma\}^m$. We wish to prove $z \in \text{conv}(R)$. From the induction hypothesis we have $\bar{z} \in \text{conv}(\bar{R}_+)$ and $\bar{z} \in \text{conv}(\bar{R}_-)$. Thus, for all $y \in \{\pm 1\}$ there exist $\alpha_y, \beta_y \geq 0$ such that $\sum_{y \in Y_+} \alpha_y = \sum_{y \in Y_-} \beta_y = 1$, and

$$\bar{z} = \sum_{y \in Y_+} \alpha_y \bar{r}_y = \sum_{y \in Y_-} \beta_y \bar{r}_y.$$

Let $z_a = \sum_{y \in Y_+} \alpha_y r_y$ and $z_b = \sum_{y \in Y_-} \beta_y r_y$ We have that $\forall y \in Y_+, r_y[m] \geq \gamma$, and $\forall y \in Y_-, r_y[m] \leq -\gamma$. Therefore, $z_b[m] \leq -\gamma \leq z[m] \leq \gamma \leq z_a[m]$. In addition, $\bar{z}_a = \bar{z}_b = \bar{z}$. Select $\lambda \in [0, 1]$ such that $z[m] = \lambda z_a[m] + (1 - \lambda) z_b[m]$, then $z = \lambda z_a + (1 - \lambda) z_b$. Since $z_a, z_b \in \text{conv}(R)$, we have $z \in \text{conv}(R)$. ∎

Proof [of Lemma 15] Denote by $f(S)$ the vector $(f(x_1), \ldots, f(x_m))$. Recall that $r \in \mathbb{R}^m$ is the witness for the shattering of S, and let

$$L = \{f(S) - r \mid f \in \mathcal{F}\} \subseteq \mathbb{R}^m.$$

Since S is shattered, for any $y \in \{\pm 1\}^m$ there is an $r_y \in L$ such that $\forall i \in [m], r_y[i]y[i] \geq \gamma$. By Lemma 26, $\{\pm\gamma\}^m \subseteq \text{conv}(L)$. Since $\mathcal{F}$ is convex, L is also convex. Therefore $\{\pm\gamma\}^m \subseteq L$. ∎

A.5 Proof of Lemma 20

Proof [of Lemma 20] It suffices to consider diagonal moment matrices: If $\mathbb{B}$ is not diagonal, let $\mathbb{V} \in \mathbb{R}^{d \times d}$ be an orthogonal matrix such that $\mathbb{V}\mathbb{B}\mathbb{V}^T$ is diagonal, and let $Y = \mathbb{V}X$. We have $\mathbb{E}[\exp(t\|Y\|^2)] = \mathbb{E}[\exp(t\|X\|^2)]$ and $\text{trace}(\mathbb{V}\mathbb{B}\mathbb{V}^T) = \text{trace}(\mathbb{B})$. In addition, for all $u \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(\langle u, Y \rangle)] = \mathbb{E}[\exp(\langle \mathbb{V}^T u, X \rangle)] \leq \exp(\frac{1}{2}\langle \mathbb{B}\mathbb{V}^T u, \mathbb{V}^T u \rangle) = \exp(\frac{1}{2}\langle \mathbb{V}\mathbb{B}\mathbb{V}^T u, u \rangle).$$

Therefore Y is sub-Gaussian with the diagonal moment matrix $\mathbb{V}\mathbb{B}\mathbb{V}^T$. Thus assume w.l.o.g. that $\mathbb{B} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ where $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$.

We have $\exp(t\|X\|^2) = \prod_{i\in[d]}\exp(tX[i]^2)$. In addition, for any $t > 0$ and $x \in \mathbb{R}$, $2\sqrt{\Pi t}\cdot\exp(tx^2) = \int_{-\infty}^{\infty}\exp(sx - \frac{s^2}{4t})ds$. Therefore, for any $u \in \mathbb{R}^d$,

$$(2\sqrt{\Pi t})^d \cdot \mathbb{E}[\exp(t\|X\|^2)] = \mathbb{E}\left[\prod_{i\in[d]}\int_{-\infty}^{\infty}\exp(u[i]X[i] - \frac{u[i]^2}{4t})du[i]\right]$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i\in[d]}\exp(u[i]X[i] - \frac{u[i]^2}{4t})du[i]\right]$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(\langle u,X\rangle - \frac{\|u\|^2}{4t})\prod_{i\in[d]}du[i]\right]$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathbb{E}[\exp(\langle u,X\rangle)]\exp(-\frac{\|u\|^2}{4t})\prod_{i\in[d]}du[i]$$

By the sub-Gaussianity of X, the last expression is bounded by

$$\leq \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(\frac{1}{2}\langle \mathbb{B}u,u\rangle - \frac{\|u\|^2}{4t})\prod_{i\in[d]}du[i]$$

$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i\in[d]}\exp(\frac{\lambda_i u[i]^2}{2} - \frac{u[i]^2}{4t})du[i]$$

$$= \prod_{i\in[d]}\int_{-\infty}^{\infty}\exp(u[i]^2(\frac{\lambda_i}{2} - \frac{1}{4t}))du[i] = \Pi^{d/2}(\prod_{i\in[d]}(\frac{1}{4t} - \frac{\lambda_i}{2}))^{-\frac{1}{2}}.$$

The last equality follows from the fact that for any $a > 0$, $\int_{-\infty}^{\infty}\exp(-a\cdot s^2)ds = \sqrt{\Pi/a}$, and from the assumption $t \leq \frac{1}{4\lambda_1}$. We conclude that

$$\mathbb{E}[\exp(t\|X\|^2)] \leq (\prod_{i\in[d]}(1 - 2\lambda_i t))^{-\frac{1}{2}} \leq \exp(2t\cdot\sum_{i=1}^{d}\lambda_i) = \exp(2t\cdot\text{trace}(\mathbb{B})),$$

where the second inequality holds since $\forall x \in [0,1]$, $(1 - x/2)^{-1} \leq \exp(x)$. ∎

A.6 Proof of Theorem 23

In the proof of Theorem 23 we use the fact $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) = \inf_{\|x\|_2=1}\|\mathbb{X}^T x\|^2$ and bound the right-hand side via an ε-net of the unit sphere in $\mathbb{R}^m$, denoted by $S^{m-1} \triangleq \{x \in \mathbb{R}^m \mid \|x\|_2 = 1\}$. An ε-net of the unit sphere is a set $C \subseteq S^{m-1}$ such that $\forall x \in S^{m-1}, \exists x' \in C, \|x - x'\| \leq \varepsilon$. Denote the minimal size of an ε-net for S^{m-1} by $\mathcal{N}_m(\varepsilon)$, and by $C_m(\varepsilon)$ a minimal ε-net of S^{m-1}, so that $C_m(\varepsilon) \subseteq S^{m-1}$ and $|C_m(\varepsilon)| = \mathcal{N}_m(\varepsilon)$. The proof of Theorem 23 requires several lemmas. First we prove a concentration result for the norm of a matrix defined by sub-Gaussian variables. Then we bound the probability that the squared norm of a vector is small.

Lemma 27 *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that $\mathbb{Y}_{ij}$ are independent sub-Gaussian variables with moment B. Let Σ be a diagonal $d \times d$ PSD matrix such that $\Sigma \leq I$. Then for all $t \geq 0$*

and $\varepsilon \in (0,1)$,

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \mathcal{N}_m(\varepsilon) \exp(\frac{\mathrm{trace}(\Sigma)}{2} - \frac{t^2(1-\varepsilon)^2}{4B^2}).$$

Proof We have $\|\sqrt{\Sigma}\mathbb{Y}\| \leq \max_{x \in \mathcal{C}_m(\varepsilon)} \|\sqrt{\Sigma}\mathbb{Y}x\|/(1-\varepsilon)$, see for instance in Bennett et al. (1975). Therefore,

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \sum_{x \in \mathcal{C}_m(\varepsilon)} \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \geq (1-\varepsilon)t]. \tag{12}$$

Fix $x \in \mathcal{C}_m(\varepsilon)$. Let $V = \sqrt{\Sigma}\mathbb{Y}x$, and assume $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. For $u \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(\langle u, V \rangle)] = \mathbb{E}[\exp(\sum_{i \in [d]} u_i \sqrt{\lambda_i} \sum_{j \in [m]} \mathbb{Y}_{ij} x_j)] = \prod_{j,i} \mathbb{E}[\exp(u_i \sqrt{\lambda_i} \mathbb{Y}_{ij} x_j)]$$

$$\leq \prod_{j,i} \exp(u_i^2 \lambda_i B^2 x_j^2/2) = \exp(\frac{B^2}{2} \sum_{i \in [d]} u_i^2 \lambda_i \sum_{j \in [m]} x_j^2)$$

$$= \exp(\frac{B^2}{2} \sum_{i \in [d]} u_i^2 \lambda_i) = \exp(\langle B^2 \Sigma u, u \rangle/2).$$

Thus V is a sub-Gaussian vector with moment matrix $B^2\Sigma$. Let $s = 1/(4B^2)$. Since $\Sigma \leq I$, we have $s \leq 1/(4B^2 \max_{i \in [d]} \lambda_i)$. Therefore, by Lemma 20,

$$\mathbb{E}[\exp(s\|V\|^2)] \leq \exp(2sB^2 \mathrm{trace}(\Sigma)).$$

By Chernoff's method, $\mathbb{P}[\|V\|^2 \geq z^2] \leq \mathbb{E}[\exp(s\|V\|^2)]/\exp(sz^2)$. Thus

$$\mathbb{P}[\|V\|^2 \geq z^2] \leq \exp(2sB^2 \mathrm{trace}(\Sigma) - sz^2) = \exp(\frac{\mathrm{trace}(\Sigma)}{2} - \frac{z^2}{4B^2}).$$

Set $z = t(1-\varepsilon)$. Then for all $x \in S^{m-1}$

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \geq t(1-\varepsilon)] = \mathbb{P}[\|V\| \geq t(1-\varepsilon)] \leq \exp(\frac{\mathrm{trace}(\Sigma)}{2} - \frac{t^2(1-\varepsilon)^2}{4B^2}).$$

Therefore, by Equation (12),

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \mathcal{N}_m(\varepsilon) \exp(\frac{\mathrm{trace}(\Sigma)}{2} - \frac{t^2(1-\varepsilon)^2}{4B^2}).$$

∎

Lemma 28 *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that $\mathbb{Y}_{ij}$ are independent centered random variables with variance 1 and fourth moments at most B. Let Σ be a diagonal $d \times d$ PSD matrix such that $\Sigma \leq I$. There exist $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that for any $x \in S^{m-1}$*

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 \leq \alpha \cdot (\mathrm{trace}(\Sigma) - 1)] \leq \eta^{\mathrm{trace}(\Sigma)}.$$

To prove Lemma 28 we require Lemma 29 (Rudelson and Vershynin, 2008, Lemma 2.2) and Lemma 30, which extends Lemma 2.6 in the same work.

Lemma 29 *Let $T_1, \ldots, T_n$ be independent non-negative random variables. Assume that there are $\theta > 0$ and $\mu \in (0,1)$ such that for any i, $\mathbb{P}[T_i \leq \theta] \leq \mu$. There are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on θ and μ such that*

$$\mathbb{P}[\sum_{i=1}^{n} T_i < \alpha n] \leq \eta^n.$$

Lemma 30 *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that the columns of $\mathbb{Y}$ are i.i.d. random vectors. Assume further that $\mathbb{Y}_{ij}$ are centered, and have a variance of 1 and a fourth moment at most B. Let Σ be a diagonal $d \times d$ PSD matrix. Then for all $x \in S^{m-1}$,*

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \leq \sqrt{\text{trace}(\Sigma)/2}] \leq 1 - 1/(196B).$$

Proof Let $x \in S^{m-1}$, and $T_i = (\sum_{j=1}^{m} \mathbb{Y}_{ij}x_j)^2$. Let $\lambda_1, \ldots, \lambda_d$ be the values on the diagonal of Σ, and let $T_\Sigma = \|\sqrt{\Sigma}\mathbb{Y}x\|^2 = \sum_{i=1}^{d} \lambda_i T_i$. First, since $\mathbb{E}[\mathbb{Y}_{ij}] = 0$ and $\mathbb{E}[\mathbb{Y}_{ij}] = 1$ for all i, j, we have

$$\mathbb{E}[T_i] = \sum_{i \in [m]} x_j^2 \mathbb{E}[\mathbb{Y}_{ij}^2] = \|x\|^2 = 1.$$

Therefore $\mathbb{E}[T_\Sigma] = \text{trace}(\Sigma)$. Second, since $\mathbb{Y}_{i1}, \ldots, \mathbb{Y}_{im}$ are independent and centered, we have (Ledoux and Talagrand, 1991, Lemma 6.3)

$$\mathbb{E}[T_i^2] = \mathbb{E}[(\sum_{j \in [m]} \mathbb{Y}_{ij}x_j)^4] \leq 16\mathbb{E}_\sigma[(\sum_{j \in [m]} \sigma_j \mathbb{Y}_{ij}x_j)^4],$$

where $\sigma_1, \ldots, \sigma_m$ are independent uniform $\{\pm 1\}$ variables. Now, by Khinchine's inequality (Nazarov and Podkorytov, 2000),

$$\mathbb{E}_\sigma[(\sum_{j \in [m]} \sigma_j \mathbb{Y}_{ij}x_j)^4] \leq 3\mathbb{E}[(\sum_{j \in [m]} \mathbb{Y}_{ij}^2 x_j^2)^2] = 3 \sum_{j,k \in [m]} x_j^2 x_k^2 \mathbb{E}[\mathbb{Y}_{ij}^2]\mathbb{E}[\mathbb{Y}_{ik}^2].$$

Now $\mathbb{E}[\mathbb{Y}_{ij}^2]\mathbb{E}[\mathbb{Y}_{ik}^2] \leq \sqrt{\mathbb{E}[\mathbb{Y}_{ij}^4]\mathbb{E}[\mathbb{Y}_{ik}^4]} \leq B$. Thus $\mathbb{E}[T_i^2] \leq 48B\sum_{j,k \in [m]} x_j^2 x_k^2 = 48B\|x\|^4 = 48B$. Thus,

$$\mathbb{E}[T_\Sigma^2] = \mathbb{E}[(\sum_{i=1}^{d} \lambda_i T_i)^2] = \sum_{i,j=1}^{d} \lambda_i \lambda_j \mathbb{E}[T_i T_j]$$

$$\leq \sum_{i,j=1}^{d} \lambda_i \lambda_j \sqrt{\mathbb{E}[T_i^2]\mathbb{E}[T_j^2]} \leq 48B(\sum_{i=1}^{d} \lambda_i)^2 = 48B \cdot \text{trace}(\Sigma)^2.$$

By the Paley-Zigmund inequality (Paley and Zygmund, 1932), for $\theta \in [0, 1]$

$$\mathbb{P}[T_\Sigma \geq \theta\mathbb{E}[T_\Sigma]] \geq (1-\theta)^2 \frac{\mathbb{E}[T_\Sigma]^2}{\mathbb{E}[T_\Sigma^2]} \geq \frac{(1-\theta)^2}{48B}.$$

Therefore, setting $\theta = 1/2$, we get $\mathbb{P}[T_\Sigma \leq \text{trace}(\Sigma)/2] \leq 1 - 1/(196B)$. ∎

Proof [of Lemma 28] Let $\lambda_1, \ldots, \lambda_d \in [0, 1]$ be the values on the diagonal of Σ. Consider a partition $Z_1, \ldots, Z_k$ of $[d]$, and denote $L_j = \sum_{i \in Z_j} \lambda_i$. There exists such a partition such that for all $j \in [k]$,

$L_j \leq 1$, and for all $j \in [k-1]$, $L_j > \frac{1}{2}$. Let $\Sigma[j]$ be the sub-matrix of Σ that includes the rows and columns whose indexes are in Z_j. Let $\mathbb{Y}[j]$ be the sub-matrix of $\mathbb{Y}$ that includes the rows in Z_j. Denote $T_j = \|\sqrt{\Sigma[j]}\mathbb{Y}[j]x\|^2$. Then

$$\|\sqrt{\Sigma}\mathbb{Y}x\|^2 = \sum_{j \in [k]} \sum_{i \in Z_j} \lambda_i (\sum_{j=1}^{m} \mathbb{Y}_{ij}x_j)^2 = \sum_{j \in [k]} T_j.$$

We have $\mathrm{trace}(\Sigma) = \sum_{i=1}^{d} \lambda_i \geq \sum_{j \in [k-1]} L_j \geq \frac{1}{2}(k-1)$. In addition, $L_j \leq 1$ for all $j \in [k]$. Thus $\mathrm{trace}(\Sigma) \leq k \leq 2\,\mathrm{trace}(\Sigma) + 1$. For all $j \in [k-1]$, $L_j \geq \frac{1}{2}$, thus by Lemma 30, $\mathbb{P}[T_j \leq 1/4] \leq 1 - 1/(196B)$. Therefore, by Lemma 29 there are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 < \alpha \cdot (\mathrm{trace}(\Sigma) - 1)] \leq \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 < \alpha(k-1)]$$
$$= \mathbb{P}[\sum_{j \in [k]} T_j < \alpha(k-1)] \leq \mathbb{P}[\sum_{j \in [k-1]} T_j < \alpha(k-1)] \leq \eta^{k-1} \leq \eta^{2\,\mathrm{trace}(\Sigma)}.$$

The lemma follows by substituting η for η^2. ∎

Proof [of Theorem 23] We have

$$\sqrt{\lambda_{\min}(\mathbb{X}\mathbb{X}^T)} = \inf_{x \in S^{m-1}} \|\mathbb{X}^T x\| \geq \min_{x \in C_m(\varepsilon)} \|\mathbb{X}^T x\| - \varepsilon\|\mathbb{X}^T\|. \tag{13}$$

For brevity, denote $L = \mathrm{trace}(\Sigma)$. Assume $L \geq 2$. Let $m \leq L \cdot \min(1, (c - K\varepsilon)^2)$ where c, K, ε are constants that will be set later such that $c - K\varepsilon > 0$. By Equation (13)

$$\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \leq m] \leq \mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \leq (c - K\varepsilon)^2 L]$$
$$\leq \mathbb{P}[\min_{x \in C_m(\varepsilon)} \|\mathbb{X}^T x\| - \varepsilon\|\mathbb{X}^T\| \leq (c - K\varepsilon)\sqrt{L}] \tag{14}$$
$$\leq \mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] + \mathbb{P}[\min_{x \in C_m(\varepsilon)} \|\mathbb{X}^T x\| \leq c\sqrt{L}]. \tag{15}$$

The last inequality holds since the inequality in line (14) implies at least one of the inequalities in line (15). We will now upper-bound each of the terms in line (15). We assume w.l.o.g. that Σ is not singular (since zero rows and columns can be removed from $\mathbb{X}$ without changing $\lambda_{\min}(\mathbb{X}\mathbb{X}^T)$). Define $\mathbb{Y} \triangleq \sqrt{\Sigma^{-1}}\mathbb{X}^T$. Note that $\mathbb{Y}_{ij}$ are independent sub-Gaussian variables with (absolute) moment ρ. To bound the first term in line (15), note that by Lemma 27, for any $K > 0$,

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] = \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq K\sqrt{L}] \leq \mathcal{N}_m(\frac{1}{2})\exp(L(\frac{1}{2} - \frac{K^2}{16\rho^2})).$$

By Rudelson and Vershynin (2009), Proposition 2.1, for all $\varepsilon \in [0,1]$, $\mathcal{N}_m(\varepsilon) \leq 2m(1 + \frac{2}{\varepsilon})^{m-1}$. Therefore

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] \leq 2m5^{m-1}\exp(L(\frac{1}{2} - \frac{K^2}{16\rho^2})).$$

Let $K^2 = 16\rho^2(\frac{3}{2} + \ln(5) + \ln(2/\delta))$. Recall that by assumption $m \leq L$, and $L \geq 2$. Therefore

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] \leq 2m5^{m-1}\exp(-L(1 + \ln(5) + \ln(2/\delta)))$$
$$\leq 2L5^{L-1}\exp(-L(1 + \ln(5) + \ln(2/\delta))).$$

Since $L \geq 2$, we have $2L \exp(-L) \leq 1$. Therefore

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] \leq 2L \exp(-L - \ln(2/\delta)) \leq \exp(-\ln(2/\delta)) = \frac{\delta}{2}. \qquad (16)$$

To bound the second term in line (15), since $\mathbb{Y}_{ij}$ are sub-Gaussian with moment ρ, $\mathbb{E}[\mathbb{Y}_{ij}^4] \leq 5\rho^4$ (Buldygin and Kozachenko, 1998, Lemma 1.4). Thus, by Lemma 28, there are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on ρ such that for all $x \in S^{m-1}$, $\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 \leq \alpha(L-1)] \leq \eta^L$. Set $c = \sqrt{\alpha/2}$. Since $L \geq 2$, we have $c\sqrt{L} \leq \sqrt{\alpha(L-1)}$. Thus

$$\mathbb{P}[\min_{x \in C_m(\varepsilon)} \|\mathbb{X}^T x\| \leq c\sqrt{L}] \leq \sum_{x \in C_m(\varepsilon)} \mathbb{P}[\|\mathbb{X}^T x\| \leq c\sqrt{L}]$$

$$\leq \sum_{x \in C_m(\varepsilon)} \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \leq \sqrt{\alpha(L-1)}] \leq \mathcal{N}_m(\varepsilon)\eta^L.$$

Let $\varepsilon = c/(2K)$, so that $c - K\varepsilon > 0$. Let $\theta = \min(\frac{1}{2}, \frac{\ln(1/\eta)}{2\ln(1+2/\varepsilon)})$. Set $L_\circ$ such that $\forall L \geq L_\circ$, $L \geq \frac{2\ln(2/\delta)+2\ln(L)}{\ln(1/\eta)}$. For $L \geq L_\circ$ and $m \leq \theta L \leq L/2$,

$$\mathcal{N}_m(\varepsilon)\eta^L \leq 2m(1+2/\varepsilon)^{m-1}\eta^L$$

$$\leq L \exp(L(\theta\ln(1+2/\varepsilon) - \ln(1/\eta)))$$

$$= \exp(\ln(L) + L(\theta\ln(1+2/\varepsilon) - \ln(1/\eta)/2) - L\ln(1/\eta)/2)$$

$$\leq \exp(L(\theta\ln(1+2/\varepsilon) - \ln(1/\eta)/2) + \ln(\delta/2)) \qquad (17)$$

$$\leq \exp(\ln(\delta/2)) = \frac{\delta}{2}. \qquad (18)$$

Line (17) follows from $L \geq L_\circ$, and line (18) follows from $\theta\ln(1+2/\varepsilon) - \ln(1/\eta)/2 \leq 0$. Set $\beta = \min\{(c-K\varepsilon)^2, 1, \theta\}$. Combining Equation (15), Equation (16) and Equation (18) we have that if $L \geq \bar{L} \triangleq \max(L_\circ, 2)$, then $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \leq m] \leq \delta$ for all $m \leq \beta L$. Specifically, this holds for all $L \geq 0$ and for all $m \leq \beta(L - \bar{L})$. Letting $C = \beta\bar{L}$ and substituting δ for $1 - \delta$ we get the statement of the theorem. ∎

References

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

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Mach. Learn.*, 30(1):31–56, 1998. ISSN 0885-6125.

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

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

P. Bartlett. Lecture notes. `http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/lecture25.ps`, 2006. unpublished.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

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.

G. M. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, Sep 1991.

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

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, pages 199–207. 2010.

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

V. V. Buldygin and Y. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 1998.

R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

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, Aug 1988.

C. Gentile and D. P. Helmbold. Improved lower bounds for learning from noisy examples: an information-theoretic approach. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 104–115, 1998.

M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *Proceedings of the Twenty-Seventh International Conference on Machine learning (ICML)*, 2010.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

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.

A. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine learning (ICML)*, 2004.

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

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

M. Rudelson and R. Vershynin. The Littlewood-Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

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

S. Sabato, N. Srebro, and N. Tishby. Tight sample complexity of large-margin learning. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 2038–2046, 2010.

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

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

V. N. Sudakov. Gaussian processes and measures of solid angles in hilbert space. *Sov. Math.Dokl.*, 12:412–415, 1971.

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

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

V. N. Vapnik and A. Y. Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974. (In Russian).

N. Vayatis and R. Azencott. Distribution-dependent vapnik-chervonenkis bounds. In *EuroCOLT '99*, pages 230–240, London, UK, 1999. Springer-Verlag. ISBN 3-540-65701-0.

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