Learning with Spectral Kernels and Heavy-Tailed Data

Michael W. Mahoney
Department of Mathematics,
Stanford University,
Stanford, CA 94305,
mmahoney@cs.stanford.edu

Hariharan Narayanan
Department of Computer Science,
University of Chicago,
Chicago, IL 60637,
hari@cs.uchicago.edu.

Abstract

Heavy-tailed data, *e.g.*, graphs in which the degree sequence decays according to a power law, are ubiquitous in applications. In many of those applications, spectral kernels, *e.g.*, Laplacian Eigenmaps and Diffusion Maps, are commonly-used analytic tools. We establish learnability results applicable in both settings. Our first result is an exact learning bound for learning a classification hyperplane when the components of the feature vector decay according to a power law. Thus, although the distribution of data is infinite dimensional and unbounded, a nearly optimal linear classification hyperplane is learnable due to the polynomial decay in the probability that the i^{th} feature of a random data point is non-zero. Our second result is a "gap-tolerant" learning bound for learning a nearly-optimal Δ-margin classification hyperplane when the kernel is constructed according to the Diffusion Maps procedure. Each proof bounds the annealed entropy and thus makes important use of distribution-dependent information. The proof of our first result is direct, while the proof of our second result uses as an intermediate step a commonly-accepted but not yet rigorously-proved bound for the VC dimension of gap-tolerant classifiers. We offer a rigorous proof of this result for the usual case where the margin is measured in the ℓ_2 norm, and we prove a generalization of this result to the case where the data need not have compact support and where the margin may be measured with respect to the more general ℓ_p norm.

1 Introduction

In this paper, we prove bounds on the sample complexity of classification algorithms in two situations—when spectral kernels are used to describe the data and when the data have heavy-tail properties—where distribution-independent techniques based on the Vapnik-Chervonenkis (VC) dimension fail to provide nontrivial bounds. In each case, we bound the annealed entropy of the classifier, making important use of distribution-dependent information, thereby providing dimension independent sample complexity bounds.

Heavy-tailed distributions are probability distributions whose tails are not exponentially bounded [14]. Such distributions can arise via several mechanisms, and they are ubiquitous in applications. Recall that graphs in which the degree sequence decays according to a power law have received a great deal of attention recently. Recall also that such diverse phenomenon as the degree distribution of "protein networks," the distribution of packet transmission rates over the internet, and the frequency of word use in common text have heavy-tailed behavior. See, *e.g.*, [5] and references therein for more details. When calculating the sample complexity of a classification task for data from such sources, bounds that do not take into account the distribution are likely to be very weak. In this paper, we develop distribution-dependent bounds to classify data whose magnitude decays in a heavy-tailed manner. In particular, in Theorem 4 in Section 3 below, we show that if the probability of the i^{th} coordinate of random data point being non-zero is less than $Ci^{-\alpha}$ for some $C > 0, \alpha > 1$, then the sample complexity for finding a nearly optimal linear hyperplane classifier via Empirical Risk Minimization is independent of the number of features. We prove this result by providing a dimension-independent upper bound on the annealed entropy of the class of linear classifiers in $\mathbb{R}^d$.

The distribution-dependent ideas we develop are applicable more generally. In particular, they can be used to bound the sample complexity of a classification task under the assumption that the expected value of a norm of the data is bounded, *i.e.*, when the magnitude of the feature vector of the data in some norm has a finite moment. As an application of this idea, we present distribution-dependent bounds to classify data using spectral kernels that have received attention recently. Spectral kernels such as Laplacian Eigenmaps [2] and Diffusion Maps [13], have received a great deal of attention recently for dimensionality reduction and related learning tasks [15]. Let $f_0, f_1, \ldots, f_n$ be the eigenfunctions of the normalized Laplacian on a graph G and let $\lambda_0, \lambda_1, \ldots$ be the corresponding eigenvalues. The Diffusion Map is the following feature map $\Phi : v \mapsto (\lambda_0^k f_0(v), \ldots, \lambda_n^k f_n(v))$, and Laplacian Eigenmaps is the special case when $k = 0$. In this case, the support of the data distribution is unbounded as the size of the graph increases and, although the norm of the Diffusion Map feature vector is bounded, individual elements of the feature vector may fluctuate wildly. Thus, existing results do not give dimension-independent sample complexity bounds for classification via Empirical Risk Minimization. In Theorem 6 in Section 4 below, we give dimension-independent upper bounds on the sample complexity of learning a gap-tolerant classifier, using crucially the fact that $\mathbb{E}_v\|\Phi(v)\|^2 = 1$ even if $\sup_v \|\Phi(v)\|$ is unbounded as $n \to \infty$. As with the proof of our main heavy-tailed learning result, the proof of our main spectral learning result bounds an annealed entropy.

The bound we provide on the annealed entropy of gap-tolerant classifiers in our second learning result is of more general interest. Thus, in Theorem 7 in Section 5, we prove an upper bound on the annealed entropy of gap-tolerant classifiers in ℓ_2, and more generally in a Banach space of type $p \in (1, 2]$, under the assumption that the expectation of some moment of the norm of the feature vector is bounded. To establish this result, we show that the VC dimension of gap-tolerant classifiers in a Hilbert space when the margin is Δ over a bounded domain such as a ball of radius R is bounded above by $\lfloor R^2/\Delta^2 \rfloor + 1$. Such bounds have been stated previously by Vapnik [16]. Recall, however, that in the course of his proof bounding the VC dimension of a gap-tolerant classifier whose margin is Δ over a ball of radius R (See [16], page 353.), Vapnik states, without further justification, that due to symmetry the set of points in a ball that is extremal in the sense of being the hardest to shatter with gap-tolerant classifiers is the regular simplex. Attention has been drawn to this fact by Burges ([4], footnote 20), who mentions that a rigorous proof of this fact seems to be absent. Thus, we provide a new proof of the upper bound on the VC dimension of such classifiers without making this assumption. (See Lemma 8 and its proof.) Moreover, the idea underlying our new proof of Lemma 8 generalizes to the case when the the data may be unbounded and when the gap is measured with respect to more general Banach space norms. In particular, we show that the VC dimension of gap-tolerant classifiers with margin Δ in a ball of radius R in a Banach space of Rademacher type $p \in (1, 2]$ and type constant T (See Definition 3 below.) is bounded below by $(R/\Delta)^{\frac{p}{p-1}}$ and above by $\sim (3TR/\Delta)^{\frac{p}{p-1}}$. (See Lemma 9 and Lemma 11 below.) After this paper was written, the authors learnt that Rademacher complexities have been used by L. Gurvits in [8] to prove upper bounds for the sample complexity of learning bounded linear functionals on ℓ_p balls.

2 Background and Preliminaries

In this paper, we consider the supervised learning problem of binary classification, *i.e.*, we consider an input space $\mathcal{X}$ (*e.g.*, a Euclidean space, or Hilbert space, or Banach space—see below) and an output space $\mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$, and where the data consist of pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ that are random variables distributed according to an unknown distribution. We shall assume that for any X, there is at most one pair (X, Y) that is observed.

We observe ℓ i.i.d. pairs $(X_i, Y_i), i = 1, \ldots, \ell$ sampled according to this unknown distribution, and the goal is to construct a classification function $\alpha : \mathcal{X} \to \mathcal{Y}$ which predicts $\mathcal{Y}$ from $\mathcal{X}$ with low probability of error.

We will be interested in the following two classification concepts. First, an *ordinary linear hyperplane classifier* consists of an oriented hyperplane, and points are labeled ± 1, depending on which side of the hyperplane they lie. Second, a *gap-tolerant classifier* consists of an oriented hyperplane and a margin of thickness Δ in some norm. Any point outside the margin is labeled ± 1, depending on which side of the hyperplane it falls on, and all points within the margin are declared "correct," without receiving a ± 1 label. This latter setting has been considered in [16, 4] (as a way of implementing structural risk minimization—apply empirical risk minimization to a succession of problems, and choose where the gap Δ that gives the minimum risk bound).

The *risk* $R(\alpha)$ of a linear hyperplane classifier α is the probability that α misclassifies a random data point (x, y) drawn from $\mathcal{P}$; more formally, $R(\alpha) := \mathbb{E}_{\mathcal{P}}[\alpha(x) \neq y]$. Given a set of ℓ labeled data points $(x_1, y_1), \ldots, (x_\ell, y_\ell)$, the *empirical risk* $R_{emp}(\alpha, \ell)$ of a linear hyperplane classifier α is the frequency of misclassification on the empirical data; more formally, $R_{emp}(\alpha, \ell) := \frac{1}{\ell}\sum_{i=1}^{\ell} \mathcal{I}[x_i \neq y_i]$, where $\mathcal{I}[\cdot]$ denotes the indicator of the respective event. The risk and empirical risk for gap-tolerant classifiers are defined in the same manner. Note, in particular, that data points labeled as "correct" do not contribute to the risk for a gap-tolerant classifier, *i.e.*, data points that are on the "wrong" side of the hyperplane but that are within the Δ margin are not considered as incorrect and do not contribute to the risk.

In problems of classification, the ultimate goal is to find a classifier that minimizes the true risk, *i.e.*, $\arg\min_{\alpha \in \Lambda} R(\alpha)$. Since the true risk of a classifier α, $R(\alpha)$, is unknown, an empirical surrogate is often used. In particular, *Empirical Risk Minimization (ERM)* is the procedure of choosing a classifier α from a set of classifiers Λ by minimizing the empirical risk $\arg\min_{\alpha \in \Lambda} R_{emp}(\alpha, \ell)$. The consistency and rate of convergence of ERM—see [16] for precise definitions—can be related to uniform bounds on the difference between the empirical risk and the true risk over all $\alpha \in \Lambda$. There is a large body of literature on sufficient conditions for this kind of uniform convergence. In this paper, our main emphasis is on the annealed entropy:

Definition 1 (Annealed Entropy) *Let $\mathcal{P}$ be a probability measure supported on a vector space $\mathcal{H}$. Given a set Λ of decision rules and a set of points $Z = \{z_1, \ldots, z_\ell\} \subset \mathcal{H}$, let $N^\Lambda(z_1, \ldots, z_\ell)$ be the number of ways of labeling $\{z_1, \ldots, z_\ell\}$ into positive and negative samples such that there exists a gap-tolerant classifier that predicts incorrectly the label of each z_i. Given the above notation,*

$$H^\Lambda_{ann}(k) := \ln \mathbb{E}_{\mathcal{P}^{\times k}} N^\Lambda(z_1, \ldots, z_k)$$

is the annealed entropy of the classifier Λ with respect to $\mathcal{P}$.

Note that the annealed entropy is a distribution-specific measure, *i.e.*, the same family of classifiers can have different annealed entropies when measured with respect to different distributions. This definition of annealed entropy for gap-tolerant classifiers also holds for ordinary linear hyperplane classifiers. Moreover, for ordinary linear hyperplane classifiers, this definition is identical to the definition obtained by replacing the word "incorrectly" above "correctly," whereas for gap-tolerant classifiers, this is not true. For a more detailed exposition of this issue, we refer the reader to ([4], Appendix A.2).

As the following theorem states, the annealed entropy of a classifier can be used to get an upper bound on the generalization error.

Theorem 2 (Vapnik, [16]) *Given the above notation, the inequality*

$$\mathbb{P}\left[\sup_{\alpha\in\Lambda}\frac{R(\alpha)-R_{emp}(\alpha,\ell)}{\sqrt{R(\alpha)}}>\epsilon\right]$$
$$<4\exp\left(\left(\frac{H^{\Lambda}_{ann}(2\ell)}{\ell}-\frac{\epsilon^2}{4}\right)\ell\right)$$

holds true, for any number of samples ℓ and for any error parameter ϵ.

Normed vector spaces (such as Hilbert spaces and Banach spaces) are relevant to learning theory for the following reason. Data are often accompanied with an underlying metric which carries information about how likely it is that two data points have the same label. This makes concrete the intuition that points with the same class label are clustered together. Many algorithms cannot be implemented over an arbitrary metric space, but require a linear structure. If the original metric space does not have such a structure, as is the case when classifying for example, biological data or decision trees, it is customary to construct a feature space representation, which embeds data into a vector space. We will be interested in the commonly-used Hilbert spaces, in which distances in the feature space are measure with respect to the ℓ_2 distance, as well as more general Banach spaces:

Definition 3 (Banach space, type, and type constant) *A Banach space is a complete normed vector space. A Banach space $\mathcal{B}$ is said to have (Rademacher) type p if there exists $T<\infty$ such that for all n and $x_1,\dots,x_n\in\mathcal{B}$*

$$\mathbb{E}_\epsilon[\|\sum_{i=1}^{n}\epsilon_i x_i\|^p_{\mathcal{B}}]\le T^p\sum_{i=1}^{n}\|x_i\|^p_{\mathcal{B}}.$$

The smallest T for which the above holds with p equal to the type, is called the type constant of $\mathcal{B}$.

Our results, where the margin is measured in ℓ_2 can be transferred to a setting with kernels. Given a kernel $k(\cdot,\cdot)$, it is well known that linear classification using a kernel $k(\cdot,\cdot)$ is equivalent to mapping x onto the functional $k(x,\cdot)$ and then finding a separating halfspace in the Reproducing Kernel Hilbert Space (RKHS) which is the Hilbert Space generated by the functionals of the form $k(x,\cdot)$. Since the span of any finite set of points in a Hilbert Space can be isometrically embedded in ℓ_2, our results hold in the setting of kernel-based learning as well, when one first uses the feature map $x\mapsto k(x,\cdot)$ and works in the RKHS.

3 Learning with heavy-tailed data

In this section, we state and prove Theorem 4, our main result for dimension-independent learning from data in which the feature map exhibits a heavy-tailed decay.

Consider the following toy model for classifying web pages using keywords. One approach to this problem could be to associate with each web page the indicator vector corresponding to all keywords that it contains. The dimension of this feature space is the number of possible keywords, which is typically very large, and empirical evidence indicates that the frequency of words decays in a heavy-tailed manner. Thus the VC dimension of the feature space is very large, and in a distribution-free setting it is not possible to classify data in such a feature space unless the number of samples is of the order of the VC dimension. We show that if the probability that the i^{th} keyword is present is heavy-tailed as a function of i, then the sample complexity of the binary classification problem is dimension-independent.

More precisely, the following theorem provides a dimension-independent upper bound on the number of samples needed to learn by ERM, with a given accuracy and confidence, a linear hyperplane that classifies heavy-tailed data into positive and negative labels, under the assumption that the probability of the i^{th} coordinate of a random data point being non-zero is less than $Ci^{-\alpha}$ for some $C>0,\alpha>1$. The proof of this result proceeds by providing providing a dimension-independent upper bound on the annealed entropy of the class of linear classifiers in $\mathbb{R}^d$, and then appealing to Theorem 2 relating the annealed entropy to the generalization error.

Theorem 4 (Bounds for Heavy-Tailed Data) *Let $\mathcal{P}$ be a probability distribution in $\mathbb{R}^d$. Suppose $\mathcal{P}[x_i\neq 0]\le Ci^{-\alpha}$ for some absolute constant $C>0$, with $\alpha>1$. Then, the annealed entropy of ordinary linear hyperplane classifiers is*

$$H^{\Lambda}_{ann}(\ell)\quad\le\quad\left(\frac{C}{\alpha-1}\ell^{\frac{1}{\alpha}}+1\right)\ln\ell \tag{1}$$

Consequently, the minimum number of random samples $\ell=\ell(\epsilon,\delta)$ needed to learn, by ERM, a classifier whose risk differs from the minimum risk $R(\alpha)$ by $<\epsilon\sqrt{R(\alpha)}$ with probability $>1-\delta$ is less than or equal to

$$2\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1}+\ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1}+\ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}}\right).$$

Proof: Let the event that a sample $z_i=(z_{i1},z_{i2},\dots)$ has a non-zero coordinate $z_{ik'}$ for some $k'>\ell^{1/\alpha}$ be denoted E_i. The probability of this event can be bounded as follows. If $\alpha\neq 1$, and $k=\ell^{1/\alpha}$

$$\begin{aligned}\mathbb{P}[E_i] &= \mathbb{P}[\exists k'>\ell^{1/\alpha},\text{ such that } z_{ik'}\neq 0]\\ &\le C\sum_{i=k+1}^{\infty}i^{-\alpha}\\ &\le \frac{Ck^{-\alpha+1}}{\alpha-1}\end{aligned}$$

We partition the z_i into two classes :

$$X=\{x_1,\dots,x_{\ell-m}\}:=\{z_i\text{ such that }E_i\text{ holds }\}$$

and

$$Y=\{y_1,\dots,y_m\}:=\{z_i\text{ such that }E_i\text{ does not hold }\}.$$

N^Λ is sub-multiplicative by Lemma 5. Taking an expectation over ℓ i.i.d samples from $\mathcal{P}$,

$$\mathbb{E}[N^\Lambda(z_1,\dots,z_\ell)]\quad\le\quad\mathbb{E}[N^\Lambda(x_1,\dots,x_{\ell-m})N^\Lambda(y_1,\dots,y_m)]$$

The dimension of the span of $\{y_1, \dots, y_m\}$ is at most k, and by a result from VC theory ([16], page 159) we have

$$N^\Lambda(y_1, \dots, y_m) \leq \exp(k \ln(\frac{m}{k}) + 1).$$

Then,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-m}) e m^k].$$

Moving em^k outside this expression,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq$$

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})] e m^k.$$

Note that $N^\Lambda(x_1, \dots, x_{\ell-k})$ is always bounded above by $2^{\ell-k}$. The events $E_1, E_2, \dots$ are independent and identically distributed. Let $\mathbb{P}[E_i] = p$. $\ell - k$ is the sum of ℓ independent p-Bernoulli variables.

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})] \leq \mathbb{E}[2^{\ell-k}].$$

$\mathbb{E}[2^{\ell-k}]$ can be written as

$$\begin{aligned}
\prod_{i=1}^{\ell} (1 + \mathbb{P}[E_i]) &= (1+p)^\ell && (2)\\
&\leq e^{p\ell} && (3)\\
&= e^{\ell\,(\frac{C\,k^{-\alpha+1}}{\alpha-1})}. && (4)
\end{aligned}$$

Putting everything together, we see that

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq$$

$$e(\ell)^k e^{\frac{C\ell\,k^{-a+1}}{\alpha-1}}.$$

Since $k = \ell^{\frac{1}{\alpha}}$, we see that

$$\begin{aligned}
H^\Lambda_{ann}(\ell) &= \ln \mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] && (5)\\
&\leq \left(\frac{C}{\alpha-1}\ell^{\frac{1}{\alpha}} + 1\right) \ln(\ell). && (6)
\end{aligned}$$

In order to obtain sample complexity bounds, we need to apply Theorem 2 and substitute the above expression for annealed entropy. For the probability that the error of ERM exceeds $\epsilon\sqrt{R(\alpha)}$ to be less than δ (where α is the optimal classifier), it is sufficient that ℓ satisfy

$$4 \exp\left(\frac{C2^{1/\alpha}}{\alpha-1}\ell^{\frac{1-\alpha}{\alpha}} \ln(2\ell) - \epsilon^2/4\right)\ell \leq \delta.$$

For this to be true, it is enough that

$$\frac{\epsilon^2 \ell^{1-\frac{1}{\alpha}}}{4} \geq \frac{C2^{\frac{1}{\alpha}} \ln(2\ell)}{\alpha-1} + \ln(4/\delta).$$

A calculation shows that

$$\frac{2\alpha\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)^{\frac{\alpha}{\alpha-1}} \ln\left(\frac{4}{\epsilon^2}\left(\frac{C2^{\frac{1}{\alpha}}}{\alpha-1} + \ln\frac{4}{\delta}\right)\right)}{\alpha-1}$$

is a value of ℓ that satisfies the previous expression. ■

For ease of reference, we note the following easily established fact about N^Λ. This lemma is used in the proof of Theorem 4 above and Theorem 6 below.

Lemma 5 *Let $\{x_1, \dots, x_\ell\} \cup \{y_1, \dots, y_k\}$ be a partition of the data Z into two parts. Then, N^Λ is submultiplicative in the following sense:*

$$N^\Lambda(x_1, \dots, x_\ell, y_1, \dots y_k) \leq N^\Lambda(x_1, \dots, x_\ell) N^\Lambda(y_1, \dots, y_k).$$

Proof: This holds because any partition of $Z := \{x_1, \dots, x_\ell, y_1, \dots, y_k\}$ into two parts by an element $\mathcal{I} \in \Lambda$ induces such a partition for the sets $\{x_1, \dots, x_\ell\}$ and $\{y_1, \dots, y_\ell\}$, and for any pair of partitions of $\{x_1, \dots, x_\ell\}$ and $\{y_1, \dots, y_k\}$, there is at most one partition of Z that induces them. ■

4 Learning with spectral kernels

In this section, we state and prove Theorem 6, our main result for dimension-independent learning from data in which the feature map is constructed from a spectral kernel.

Spectral kernels have received a great deal of attention recently for data classification, regression, and dimensionality reduction [15]. Consider, for example, Laplacian Eigenmaps [2] and Diffusion Maps [13]. Given a neighborhood graph $G = (V, E)$ constructed from the data, let $f_1, f_2, \dots, f_n$ be the eigenfunctions of the normalized Laplacian of G and let $\lambda_1, \lambda_2, \dots$ be the corresponding eigenvalues. The Diffusion Map is the following feature map

$$\Phi : v \mapsto (\lambda_1^k f_1(v), \dots, \lambda_n^k f_n(v)),$$

and Laplacian Eigenmaps is the special case when $k = 0$. In this case, the support of the data distribution is unbounded as the size of the graph increases, the VC dimension is $O(n)$, and the existing results do not give dimension-independent sample complexity bounds for classification by ERM. Even for gap-tolerant classifiers, which are easier to learn than ordinary linear hyperplane classifiers, the existing bounds of Vapnik are not independent of the number n of nodes. It is possible that on some vertices v the eigenfunctions fluctuate wildly. Moreover, even on special classes of graphs, such as random graphs $G(n, p)$, a non-trivial uniform upper bound stronger than $O(n)$ on $\|\Phi(v)\|$ over all vertices v does not appear to be known. Thus, VC theory provides an upper bound of $O\left((n/\Delta)^2\right)$ on the VC dimension of gap-tolerant classifiers applied to the Diffusion feature space corresponding to a graph with n nodes. (Recall that by Lemma 8 below, the VC dimension of the space of "gap-tolerant" classifiers corresponding to a margin Δ, applied to a ball of radius R is $\sim (R/\Delta)^2$.) Of course, although this bound is quadratic in the number of nodes, VC theory for ordinary linear classifiers gives an $O(n)$ bound.

Using the tools developed in this paper, we give dimension-independent upper bounds on the sample complexity of learning a gap-tolerant classifier when vertices and their labels are observed i.i.d form the uniform distribution, We crucially use the fact that $\mathbb{E}_v\|\Phi(v)\|^2 = 1$ even if $\sup_v \|\Phi(v)\|$ is unbounded as $n \to \infty$. More precisely, the following theorem provides a dimension-independent (*i.e.*, independent of the size n of the graph and the dimension of the feature space) upper bound on the number of samples needed to learn by ERM, with a given accuracy and confidence, a gap-tolerant hyperplane classifier, under the assumption that the Diffusion Map of some scale is used as the feature map. The proof of this result proceeds by providing providing a dimension-independent upper bound on the annealed entropy of gap-tolerant classifiers in the feature space of the Diffusion Maps, and then appealing to Theorem 2 relating the annealed entropy to the generalization error. The bound on the annealed entropy follows from Theorem 7 in Section 5 below, by noting that, although bounds on the individual entries of the feature map do not appear to be known, there exist nontrivial bounds on the magnitude of the feature vectors.

Theorem 6 (Bounds for Spectral Kernels) *Let the following Diffusion map be given.*

$$\Phi : v \mapsto (\lambda_1^k f_1(v), \dots, \lambda_n^k f_n(v)),$$

where f_i are normalized eigenfunctions (whose $\ell_2(\mu)$) norm is 1, μ being the uniform distribution), λ_i are the eigenvalues of the corresponding Markov Chain and $k \geq 0$. Then, the annealed entropy of a gap-tolerant classifier in the feature space of the Diffusion Maps is

$$H^{\Lambda}_{ann}(\ell) \leq (\frac{\ell^{\frac{1}{2}}}{\Delta} + 1)(1 + \ln(\ell + 1)). \tag{7}$$

Consequently, the minimum number of random samples $\ell = \ell(\epsilon, \delta)$ needed from the uniform distribution to learn, by ERM, a gap-tolerant classifier for a labeled graph via a Diffusion Map, whose risk differs from the minimum risk $R(\alpha)$ by $< \epsilon\sqrt{R(\alpha)}$ with probability $> 1 - \delta$ is less or equal to

$$200\left(\left(\frac{1}{\Delta^2\epsilon^4} + \ln(1/\delta)\right)\ln^2\left(\frac{\ln(1/\delta)}{\Delta\epsilon}\right)\right). \tag{8}$$

Proof:A diffusion map for the graph $(V, E) = G$ is the feature map that associates with a vertex x, the feature vector $\mathbf{x} = (\lambda_1^\alpha f_1(x), \dots, \lambda_m^\alpha f_m(x))$, when the eigenfunctions corresponding to the top m eigenvalues are chosen. Let μ be the uniform distribution on V and $|V| = n$. We note that if the f_j are normalized eigenfunctions, *i.e.*, $\forall j, \sum_{x \in V} f_j(x)^2 = 1$,

$$\mathbb{E}\|\mathbf{x}\|^2 \;=\; \frac{\sum_{i=1}^n \lambda_i^{2\alpha}}{n} \leq 1. \tag{9}$$

The above inequality holds because the eigenvalues have magnitudes that are less or equal to 1:

$$1 = \lambda_1 \geq \cdots \geq \lambda_n \geq -1.$$

An application of Theorem 7 tells us the following. The annealed entropy of learning a gap-tolerant classifier in the feature space of the Diffusion Maps is

$$H^{\Lambda}_{ann}(\ell) = (\frac{\ell^{\frac{1}{2}}}{\Delta} + 1)(1 + \ln(\ell + 1)).$$

For any ℓ the inequality

$$\mathbb{P}\left[\sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{emp}(\alpha, \ell)}{\sqrt{R(\alpha)}} > \epsilon\right] <$$

$$4\exp\left(\left(\frac{(\frac{\sqrt{2}}{\Delta} + 1)(\ln(2\ell + 1) + 1)}{\ell^{\frac{1}{2}}} - \frac{\epsilon^2}{4}\right)\ell\right)$$

holds true. It follows that $\ell(\epsilon, \delta)$ is bounded above by any ℓ that satisfies

$$4\exp\left(\left(\frac{(\frac{\sqrt{2}}{\Delta} + 1)(\ln(2\ell + 1) + 1)}{\ell^{\frac{1}{2}}} - \frac{\epsilon^2}{4}\right)\ell\right) < \delta.$$

A tedious but elementary calculation shows that

$$\frac{\ell(\epsilon, \delta)}{(\ln(2\ell(\epsilon, \delta) + 1) + 1)^2} \leq \frac{\left(\frac{8\sqrt{2}}{\Delta} + 4\right)}{\epsilon^2} + 2\ln\frac{4}{\delta} + \frac{\left(\frac{4\sqrt{2}}{\Delta} + 2\right)^2}{\epsilon^4}.$$

Therefore,

$$\ell(\epsilon, \delta) < 200\left(\frac{1}{\Delta^2\epsilon^4} + \ln(1/\delta)\right)\ln^2\left(\frac{\ln(1/\delta)}{\Delta\epsilon}\right).$$

■

5 Gap-tolerant classifiers in Hilbert spaces and in Banach spaces

In this section, we state and prove Theorem 7, our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers in ℓ_2 and, more generally, in a Banach space of type p with type constant T. After stating and proving Theorem 7 and stating several intermediate lemmas in the next subsection, we devote the subsequent subsections to proving the intermediate lemmas.

5.1 Summary of results for gap-tolerant classification

The following theorem is our main result regarding an upper bound for the annealed entropy of gap-tolerant classifiers. The ℓ_2 bound for this theorem will use Lemma 8, while the Banach space bound will use Lemma 9, both of which are stated and proved below. We state the ℓ_2 bound explicitly since the bound provided by Lemma 8 is slightly better than the corresponding bound for implied by the more general Lemma 9. Note that it may seem counter-intuitive that in the case of ℓ_2 (when we set $\gamma = 2$ below) the dependence of Δ is Δ^{-1}, which is weaker than in the VC bound where it is Δ^{-2}. The explanation is that the bound on annealed entropy here depends on the number of samples ℓ, while the VC dimension does not. Therefore, the weaker dependence on Δ is compensated for by a term that in fact tends to ∞ as the number of samples $\ell \to \infty$.

Theorem 7 (Annealed entropy) *Let $\mathcal{P}$ be a probability measure on a Hilbert space $\mathcal{H}$, let $\Delta > 0$, and let $\mathbb{E}_{\mathcal{P}}\|x\|^2 = r^2 < \infty$. Then the annealed entropy of gap-tolerant classifiers in $\mathcal{H}$, where the gap is Δ, is*

$$H^{\Lambda}_{ann}(\ell) \leq$$

$$\left(\ell^{\frac{1}{2}}\left(\frac{r}{\Delta}\right) + 1\right)(1 + \ln(\ell + 1)).$$

More generally, let $\mathcal{P}$ be a probability measure on a Banach space $\mathcal{B}$ of type p and type constant T. Let $\gamma, \Delta > 0$, and let $\eta = \frac{p}{p+\gamma(p-1)}$. If $\mathbb{E}_{\mathcal{P}}\|x\|^\gamma = r^\gamma < \infty$, then the annealed entropy of gap-tolerant classifiers in $\mathcal{B}$, where the gap is Δ, is

$$H^{\Lambda}_{ann}(\ell) \leq$$

$$\left(\eta^{-\eta}(1-\eta)^{-1+\eta}\left(\frac{\ell}{\ln(\ell + 1)}\left(\frac{3Tr}{\Delta}\right)^\gamma\right)^\eta + 64\right)\ln(\ell{+}1).$$

Proof: We prove the Banach space result first. Let ℓ independent, identically distributed (i.i.d) samples $z_1, \dots, z_\ell$ be chosen from $\mathcal{P}$. We partition them into two classes :

$$X = \{x_1, \dots, x_{\ell-k}\} := \{z_i \mid \|z_i\| > R\},$$

and

$$Y = \{y_1, \dots, y_k\} := \{z_i \mid \|z_i\| \leq R\}.$$

Our objective is to bound from above the annealed entropy $H^\Lambda_{ann}(\ell) = \ln \mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)]$. By Lemma 5 N^Λ is submultiplicative. Therefore,

$$N^\Lambda(z_1, \dots, z_\ell) \leq N^\Lambda(x_1, \dots, x_{\ell-k}) N^\Lambda(y_1, \dots, y_k).$$

Taking an expectation over ℓ i.i.d samples from $\mathcal{P}$,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k}) N^\Lambda(y_1, \dots, y_k)].$$

Now applying Lemma 9, we see that

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq$$

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})(k+1)^{(3TR/\Delta)^{\frac{p}{p-1}}+64}].$$

Moving $(k+1)^{((2+o(1)TR/\Delta)^{\frac{p}{p-1}})}$ outside this expression,

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq$$

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})](k+1)^{(3TR/\Delta)^{\frac{p}{p-1}}+64}.$$

Note that $N^\Lambda(x_1, \dots, x_{\ell-k})$ is always bounded above by $2^{\ell-k}$. The random variables $\mathbb{I}[E_i[\|x_i\| > R]]$ are i.i.d. Let $\mathbb{P}[\|x_i\| > R] = \rho$.
$\ell - k$ is the sum of ℓ independent Bernoulli variables. Moreover, by Markov's inequality,

$$\mathbb{P}[\|x_i\| > R] \;\leq\; \frac{\mathbb{E}[\|x_i\|^\gamma]}{R^\gamma},$$

and therefore $\rho \leq (\frac{r}{R})^\gamma$.

$$\mathbb{E}[N^\Lambda(x_1, \dots, x_{\ell-k})] \leq \mathbb{E}[2^{\ell-k}].$$

Let $I[\cdot]$ denote an indicator variable. $\mathbb{E}[2^{\ell-k}]$ can be written as

$$\prod_{i=1}^{\ell} \mathbb{E}[2^{I[\|x_i\|>R]}] = (1+\rho)^\ell \leq e^{\rho\ell}.$$

Putting everything together, we see that

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \qquad (10)$$

$$\exp\left(\ell\left(\frac{r}{R}\right)^\gamma + \ln(k+1)\left(64 + \frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}\right). \qquad (11)$$

By setting $\eta := \frac{p}{\gamma(p-1)+p}$, and adjusting R so that

$$\ell\left(\frac{r}{R}\right)^\gamma \eta^{-1} = (1-\eta)^{-1}\ln(\ell+1)\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}.$$

We see that

$$\ell\left(\frac{r}{R}\right)^\gamma + \left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} =$$

$$\left(\ell\left(\frac{r}{R}\right)^\gamma \eta^{-1}\right)^\eta \left((1-\eta)^{-1}\ln(\ell+1)\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}}\right)^{1-\eta} =$$

$$\eta^{-\eta}(1-\eta)^{-1+\eta}\left(\ell\left(\frac{3Tr}{\Delta}\right)^\gamma\right)^\eta$$

Thus, it follows that

$$H^\Lambda_{ann}(\ell) = \log \mathbb{E}\left[N^\Lambda(z_1, \dots, z_\ell)\right]$$

$$\leq \left(\eta^{-\eta}(1-\eta)^{-1+\eta}\left(\frac{\ell}{\ln(\ell+1)}\left(\frac{3Tr}{\Delta}\right)^\gamma\right)^\eta + 64\right)\ln(\ell+1).$$

We next prove the special case when the Banach space is a Hilbert space. In this case, we have $\gamma = 2$. Then, using Lemma 8 instead of the VC bound for general Banach spaces, the analogue of Equation 10 that we obtain is

$$\mathbb{E}[N^\Lambda(z_1, \dots, z_\ell)] \leq \qquad (12)$$

$$\exp\left(\ell\left(\frac{r}{R}\right)^2 + \ln(\ell+1)\left(\frac{R^2}{\Delta^2}+1\right)\right). \qquad (13)$$

If we substitute $R = (\ell r^2 \Delta^2)^{\frac{1}{4}}$, it follows that

$$H^\Lambda_{ann}(\ell) = \log \mathbb{E}\left[N^\Lambda(z_1, \dots, z_\ell)\right]$$

$$\leq \left(\ell^{\frac{1}{2}}\left(\frac{r}{\Delta}\right) + 1\right)(1 + \ln(\ell+1)).$$

■

The proof of Theorem 7 Thus, as an intermediate step, we will need a bound on the VC dimension of a gap-tolerant classifier. The following lemma is due to Vapnik [16].

Lemma 8 (Upper bound; Hilbert Space) *In a Hilbert-space, the VC dimension of a gap-tolerant classifier whose margin is Δ over a ball of radius R is $\leq \lfloor \frac{R^2}{\Delta^2} \rfloor + 1$.*

Note that in the course of his proof (See [16], page 353.), Vapnik states, without further justification, that due to symmetry the set of points that is extremal in the sense of being the hardest to shatter with gap-tolerant classifiers is the regular simplex. Attention has also been drawn to this fact by Burges ([4], footnote 20), who mentions that a rigorous proof of this fact seems to be absent. Thus, we provide a new proof of Lemma 8 in Section 5.2.

The idea underlying our new proof of Lemma 8 generalizes to the case when the the gap is measured in more general Banach spaces. We state the following lemma for a Banach space of type p with type constant T. Recall, *e.g.*, that ℓ_p for $p \geq 1$ is a Banach space of type $\min(2, p)$ and type constant 1.

Lemma 9 (Upper bound; Banach Space) *In a Banach Space of type p and type constant T the VC dimension n of gap-tolerant classifiers can by bounded above by $\left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} + 64$*

The proof of Lemma 9 may be found in Section 5.3. It will employ the following form of the Chernoff bound, which we state for ease of reference.

Lemma 10 (Chernoff Bound) *Let $X_1, \dots, X_n$ be discrete independent random variables such that $\mathbb{E}[X_i] = 0$ for all i and $|X_i| \leq 1$ for all i. Let $X = \sum_{i=1}^n X_i$ for all i and σ^2 be the variance of X. Then*

$$\mathbb{P}[|X| \geq \lambda\sigma] \leq 2e^{-\lambda^2/4}$$

for any $0 \leq \lambda \leq 2\sigma$.

Finally, for completness, we next state a lower bound for VC dimension of gap-tolerant classifiers when the margin is measured in a norm that is associated with a Banach space of type $p \in (1, 2]$. Since we are interested only in a lower bound, we consider the special case of ℓ_p^n. Note that this argument does not immediately generalize to Banach spaces of higher type because for $p > 2$, ℓ_p has type 2.

Lemma 11 (Lower Bound) *For each $p \in (1, 2]$, there exists a Banach space of type p such that the VC dimension of gap-tolerant classifiers with gap Δ over a ball of radius R is greater or equal to*

$$\left(\frac{R}{\Delta}\right)^{\frac{p}{p-1}}.$$

Further, this bound is achieved when the space is ℓ_p.

5.2 Proof of Lemma 8

Proof: Suppose the VC dimension is n. Then there exists a set of n points $X = \{x_1, \dots, x_n\}$ in $B(R)$ that can be completely shattered using gap-tolerant classifiers. We will consider two cases, first that n is even, and then that n is odd.

First, assume that n is even, i.e., that $n = 2k$ for some positive integer k. We apply the probabilistic method to obtain a upper bound on n. Note that for every set $S \subseteq [n]$, the set $X_S := \{x_i | i \in S\}$ can be separated from $X - X_S$ using a gap-tolerant classifier. Therefore the distance between the centroids (respective centers of mass) of these two sets is greater or equal to 2Δ. In particular, for each S having $k = n/2$ elements,

$$\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \notin S} x_i}{k}\| \geq 2\Delta.$$

Suppose now that S is chosen uniformly at random from the $\binom{n}{k}$ sets of size k. Then,

$$4\Delta^2 \leq \mathbb{E}\left[\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \notin S} x_i}{k}\|^2\right]$$

$$= k^{-2}\left\{\frac{2k+1}{2k}\sum_{i=1}^{n}\|x_i\|^2 - \frac{\|\sum_1^n x_i\|^2}{2k}\right\}$$

$$\leq \frac{4(n+1)}{n^2}R^2$$

Therefore,

$$\Delta^2 \leq \frac{n+1}{n^2}R^2$$

$$< \frac{R^2}{n-1}$$

and so

$$n < \frac{R^2}{\Delta^2} + 1.$$

Next, assume that n is odd. We perform a similar calculation for $n = 2k + 1$. As before, we average over all sets S of cardinality k the squared distance between the centroid of X_S and the centroid (center of mass) of $X - X_S$. Proceeding as before,

$$\begin{aligned}
4\Delta^2 &\leq \mathbb{E}\left[\|\frac{\sum_{i \in S} x_i}{k} - \frac{\sum_{i \notin S} x_i}{k+1}\|^2\right] \\
&= \frac{\sum_{i=1}^n \|x_i\|^2(1 + \frac{1}{2n}) - \frac{1}{2n}\|\sum_{1 \leq i \leq n} x_i\|^2}{k(k+1)} \\
&\leq \frac{\sum_{i=1}^n \|x_i\|^2(1 + \frac{1}{2n})}{k(k+1)} \\
&= \frac{4k+3}{2k(2k+1)(k+1)}\{(2k+1)R^2\} \\
&< \frac{4R^2}{n-1}
\end{aligned}$$

Therefore, $n < \frac{R^2}{\Delta^2} + 1$. ■

5.3 Proof of Lemma 9

Proof: We use the inequality 15 determining the Rademacher type of $\mathcal{B}$. This, while permitting greater generality, provides weaker bounds than previously obtained in the Euclidean case. When necessary, we use Lemmas proved later. Note that if $\mu := \sum_{i=1}^n x_i$, then by repeated application of the Triangle Inequality,

$$\begin{aligned}
\|x_i - \mu\| &\leq (1 - \frac{1}{n})\|x_i\| + \sum_{j \neq i}\frac{\|x_j\|}{n} \\
&< 2\sup_i \|x_i\|.
\end{aligned}$$

This shows that if we start with $x_1, \dots, x_n$ having norm $\leq R$, $\|x_i - \mu\| \leq 2R$ for all i. The property of being shattered by gap-tolerant classifiers is translation invariant. Then, for $\emptyset \subsetneq S \subsetneq [n]$, it can be verified that

$$2\Delta \leq \frac{\sum_{i \in S}(x_i - \mu)}{|S|} - \frac{\sum_{i \notin S}(x_i - \mu)}{n - |S|}$$

$$= \frac{n}{2|S|(n - |S|)}\left(\sum_{i \in S}(x_i - \mu) - \sum_{i \notin S}(x_i - \mu)\right) \quad (14)$$

The Rademacher Inequality states that

$$\mathbb{E}_\epsilon[\|\sum_{i=1}^n \epsilon_i x_i\|^p] \leq T^p \sum_{i=1}^n \|x_i\|^p. \quad (15)$$

Using the version of Chernoff's bound in Lemma 10

$$\mathbb{P}[|\sum_{i=1}^n \epsilon_i| \leq \lambda\sqrt{n}] \geq 1 - 2e^{-\lambda^2/4}. \quad (16)$$

We shall denote the above event by E_λ. Now, let $x_1, \dots, x_n$ be n points in $\mathcal{B}$ with a norm less or equal to R. Let $\mu =$

$\frac{\sum_{i=1}^n x_i}{n}$ as before.

$$\begin{aligned} 2^pT^pnR^p &\geq 2^pT^p \sum_{i=1}^n \|x_i\|^p \\ &\geq T^p \sum_{i=1}^n \|x_i - \mu\|^p \\ &\geq \mathbb{E}_\epsilon[\|\epsilon_i(x_i - \mu)\|^p] \\ &\geq \mathbb{E}_\epsilon[\|\epsilon_i(x_i - \mu)\|^p | E_\lambda]\ \mathbb{P}[E_\lambda] \\ &\geq \mathbb{E}_\epsilon[(n - \lambda^2)^p (2\Delta)^p (1 - 2e^{-\lambda^2/4})] \end{aligned}$$

The last inequality follows from 14 and 16). We infer from the preceding sequence of inequalities that

$$n^{p-1} \leq 2^pT^p \left(\frac{R}{\Delta}\right)^p \left\{(1 - \frac{\lambda^2}{n})^p (1 - 2e^{-\lambda^2/4})\right\}^{-1}.$$

The above is true for any $\lambda \in (0, 2\sqrt{n})$, by the conditions in the Chernoff bound stated in Lemma 10. If $n \geq 64$, choosing λ equal to 8 gives us $n^{p-1} \leq 3^pT^p \left(\frac{R}{\Delta}\right)^p$. Therefore, it is always true that $n \leq \left(\frac{3TR}{\Delta}\right)^{\frac{p}{p-1}} + 64$. ■

5.4 Proof of Lemma 11

Proof: We shall show that the first n unit norm basis vectors in the canonical basis can be shattered using gap-tolerant classifiers, where $\Delta = n^{\frac{1-p}{p}}$. Therefore in this case, the VC dimension is $\geq (\frac{R}{\Delta})^{\frac{p}{p-1}}$. Let e_j be the j^{th} basis vector. In order to prove that the set $\{e_1, \dots, e_n\}$ is shattered, due to symmetry under permutations, it suffices to prove that for each k, $\{e_1, \dots, e_k\}$ can be separated from $\{e_{k+1}, \dots, e_n\}$ using a gap-tolerant classifier. Points in ℓ_p are infinite sequences $(x_1, \dots)$ of finite ℓ_p norm. Consider the hyperplane H defined by $\sum_{i=1}^k x_i - \sum_{i=k+1}^n x_i = 0$. Clearly, it separates the sets in question. We may assume e_j to be e_1, replacing if necessary, k by $n-k$. Let $x = \inf_{y \in H} \|e_1 - y\|_p$. Clearly, all coordinates $x_{n+1}, \dots$ of x are 0. In order to get a lower bound on the ℓ_p distance, we use the power-mean inequality:
If $p \geq 1$, and $x_1, \dots, x_n \in \mathbb{R}$,

$$\left(\frac{\sum_{i=1}^n |x_i|^p}{n}\right)^{\frac{1}{p}} \geq \frac{\sum_{i=1}^n |x_i|}{n}.$$

This implies that

$$\begin{aligned} \|e_1 - x\|_p &\geq n^{\frac{1-p}{p}} \|e_1 - x\|_1 \\ &= n^{\frac{1-p}{p}} \left(|1 - x_1| + \sum_{i=2}^n |x_i|\right) \\ &\geq n^{\frac{1-p}{p}} \left(1 - \sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i\right) \\ &= n^{\frac{1-p}{p}}. \end{aligned}$$

For $p > 2$, the type of ℓ_p is 2 [12]. Since $\frac{p}{p-1}$ is a decreasing function of p in this regime, we do not recover any useful bounds. ■

6 Conclusion

The distribution-dependent bounds we have presented are restricted to the classification problem, but we expect that a similar analysis will yield corresponding results for other statistical learning problems such as regression and density estimation. More interesting for applications would be the following two extensions. First, due to the extreme sparsity of typical heavy-tailed data, it would be of interest to establish analogous results for Structural Risk Minimization, rather than simply Empirical Risk Minimization. Second, rather than define a "diffusion map" that depends on global eigenfunctions, it would be interesting to define and use one based on recently-developed "local" spectral methods [17, 1].

More generally, however, one might view our results as providing weak bounds when compared with traditional results in learning theory. Such a view is consistent with the recent empirical results of Leskovec, Lang, Dasgupta, and Mahoney (LLDM) [10, 11], who examined the clustering and community structure in over 70 large social and information networks taken from a wide range of application domains. LLDM present empirical evidence that indicates that commonly-studied heavy-tailed graphs such as large informatics graphs are not consistent with the hypothesis that the data are drawn from a low-dimensional structure such as a low-dimensional manifold. Coupled with the empirical results of LLDM, our theoretical results presented in this paper suggest that many of the existing spectral-based techniques in machine learning will need to be modified to perform learning on "real world" heavy-tailed data.

References

[1] R. Andersen and F.R.K. Chung and K. Lang, *Local Graph Partitioning using PageRank Vectors*, FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science pages, 475–486

[2] M. Belkin and P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*, Neural Computation, June 2003.

[3] B. Bollobás and O. M. Riordan, *Mathematical results on scale-free random graphs*, Handbook of Graphs and Networks, Wiley 2004.

[4] C.J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery, Vol. 2, Number 2, p. 121-167, Kluwer Academic Publishers, 1998

[5] F.R.K. Chung and L. Lu. *Complex Graphs and Networks*, volume 107 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 2006.

[6] R. Der and D. Lee, *Large-Margin Classification in Banach Spaces*, AISTAT. vol. 2; 2007. p. 91-98

[7] Y. Freund and R. Schapire, *Large Margin Classification Using the Perceptron Algorithm*, Machine Learning, Volume 37 , Issue 3 December 1999

[8] L. Gurvits, A Note on a Scale-Sensitive Dimension of Linear Bounded Functionals in Banach Spaces,*Proceedings of the 8th International Conference on Algorithmic Learning Theory* 352 - 363, 1997

[9] M. Hein, O. Bousquet, and B. Schölkopf, *Maximal margin classification for metric spaces*, Journal of Computer and System Sciences, Volume 71 , Issue 3 (October 2005)

[10] J. Leskovec and K.J. Lang and A. Dasgupta and M.W. Mahoney, *Statistical Properties of Community Structure in Large*

Social and Information Networks, WWW '08: Proceedings of the 17th International Conference on World Wide Web, 2008, 695–704,

[11] J. Leskovec and K.J. Lang and A. Dasgupta and M.W. Mahoney, *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters*, October, 2008, arXiv:0810.1355,

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

[13] R.R. Coifman and S. Lafon, *Diffusion maps*, Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets, Vol 21, July 2006, pp 5-30.

[14] S.I. Resnick, *Heavy Tailed Phenomena*, Springer 2007.

[15] L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee (2006). *Spectral methods for dimensionality reduction.* In O. Chapelle, B. Schoelkopf, and A. Zien (eds.), Semisupervised Learning. MIT Press: Cambridge, MA.

[16] V. Vapnik, *Statistical Learning Theory*, Wiley 1998.

[17] D.A. Spielman and S.-H. Teng, *Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems*, March 2004, arXiv:cs/0310051v9,