Efficient classification for metric data

Lee-Ad Gottlieb[*]
Weizmann Institute of Science
lee-ad.gottlieb@weizmann.ac.il

Aryeh (Leonid) Kontorovich
Ben Gurion University
karyeh@cs.bgu.ac.il

Robert Krauthgamer[*]
Weizmann Institute of Science
robert.krauthgamer@weizmann.ac.il

Abstract

Recent advances in large-margin classification of data residing in general metric spaces (rather than Hilbert spaces) enable classification under various natural metrics, such as edit and earthmover distance. The general framework developed for this purpose by von Luxburg and Bousquet [JMLR, 2004] left open the question of computational efficiency and providing direct bounds on classification error.

We design a new algorithm for classification in general metric spaces, whose runtime and accuracy depend on the doubling dimension of the data points. It thus achieves superior classification performance in many common scenarios. The algorithmic core of our approach is an approximate (rather than exact) solution to the classical problems of Lipschitz extension and of Nearest Neighbor Search. The algorithm's generalization performance is established via the fat-shattering dimension of Lipschitz classifiers.

1 Introduction

A recent line of work extends the large-margin classification paradigm from Hilbert spaces to less structured ones, such as Banach or even metric spaces [HBS05, vLB04, DL07]. In this metric approach, data is presented as points with distances but without requiring the additional structure of inner products. The potentially significant advantage is that the metric can be carefully suited to the type of data, e.g. earthmover distance for images, or edit distance for sequences.

However, much of the existing machinery of generalization bounds [CV95, SS02] depends strongly on the inner-product structure of the Hilbert space. von Luxburg and Bousquet [vLB04] developed a powerful framework of large-margin classification for a general metric space $\mathcal{X}$. First, they show that the natural hypotheses (classifiers) to consider in this context are maximally smooth Lipschitz functions; indeed, they reduce classification (of points in a metric space $\mathcal{X}$) to finding a Lipschitz function ($f : \mathcal{X} \to \mathbb{R}$) consistent with the data, which is a classic problem in Analysis, known as Lipschitz extension. Next, they establish error bounds in the form of expected-loss. Finally, the computational problem of evaluating the classification function is reduced, assuming zero training error, to exact 1-nearest neighbor search. This matches a common classification heuristic, see e.g. [CH67], and the analysis of [vLB04] may be viewed as a rigorous explanation for the empirical success of this heuristic.

An important question left open by the work of [vLB04] is the efficient computation of the classifier. Specifically, exact nearest neighbor search in general metrics might require time that is linear in the sample size, and it is algorithmically nontrivial to deal with training error. In particular, the task of choosing which points will be misclassified by the hypothesis (i.e. optimizing the bias-variance tradeoff) remains to be addressed.

Our contribution. We solve the problems delineated above by showing that data with a low doubling dimension admits accurate and computationally efficient classification. In fact, this is the first time in which the doubling dimension of the data points is tied to either classification error or algorithmic runtime. (Previously, the doubling dimension of the space of classifiers was controlled by the VC dimension of the classifier space [BLL09].) We first give an alternate generalization bound for Lipschitz classifiers, which directly bounds the classification error, rather than expected loss. (A similar bound can in fact be derived from the analysis of [vLB04].) Our bound is based on an elementary analysis of the fat-shattering dimension, see Section 3.

[*] This work was supported in part by The Israel Science Foundation (grant #452/08), and by a Minerva grant.

We then present our main contribution, and give an efficient computational implementation of the Lipschitz classifier. In Section 4 we prove that once a Lipschitz classifier has been chosen, the classifier can be computed (evaluated) quickly on any new point $x \in \mathcal{X}$, by utilizing approximate nearest neighbor search (which is known to be fast when points have a low doubling dimension). In Section 5 we further show how to quickly compute a near-optimal classifier (in terms of classification error bound), even when the training error is nonzero. In particular, this necessitates the optimization of the number of incorrectly labeled examples – and moreover, their identity – as part of the bias-variance tradeoff. In Section 6 we give an example to illustrate the potential power of our approach.

2 Definitions and notation

We use standard notation and definitions throughout.

Metric spaces. A *metric* ρ on a set $\mathcal{X}$ is a positive symmetric function satisfying the triangle inequality $\rho(x,y) \le \rho(x,z) + \rho(z,y)$; together the two comprise the metric space $(\mathcal{X}, \rho)$. The diameter of a set $A \subseteq \mathcal{X}$, is defined by $\mathrm{diam}(A) = \sup_{x,y \in A} \rho(x,y)$. The Lipschitz constant of a function $f : \mathcal{X} \to \mathbb{R}$, denoted by $\|f\|_{\mathrm{Lip}}$, is defined to be the smallest $L > 0$ that satisfies $|f(x) - f(y)| \le L\rho(x,y)$ for all $x, y \in \mathcal{X}$.

Doubling dimension. For a metric $(\mathcal{X}, \rho)$, let λ be the smallest value such that every ball in $\mathcal{X}$ can be covered by λ balls of half the radius. The *doubling dimension* of $\mathcal{X}$ is $\mathrm{ddim}(\mathcal{X}) = \log_2 \lambda$. A metric is *doubling* when its doubling dimension is bounded. Note that while a low Euclidean dimension implies a low doubling dimension (Euclidean metrics of dimension d have doubling dimension $O(d)$ [GKL03]), low doubling dimension is strictly more general than low Euclidean dimension.

The following packing property can be demonstrated via a repetitive application of the doubling property: For set S with doubling dimension $\mathrm{ddim}(\mathcal{X})$, if the minimum interpoint distance in S is at least α, and $\mathrm{diam}(S) \le \beta$, then $|S| \le \lceil \beta/\alpha \rceil^{\mathrm{ddim}(\mathcal{X})+1}$ (see, for example [KL04]).

Learning. Our setting in this paper is a generalization of PAC known as *probabilistic concept learning* [KS94]. In this model, examples are drawn independently from $\mathcal{X} \times \{-1, 1\}$ according to some unknown probability distribution P, and the learner, having observed n such pairs (x, y) produces a hypothesis $h : \mathcal{X} \to \{-1, 1\}$. The *generalization error* is the probability of misclassifying a new point drawn from P:

$$P\{(x,y) : h(x) \neq y\}.$$

The quantity above is random (since it depends on a random sequence) and we wish to upper-bound it in probability. Most bounds of this sort contains a *sample error* term (corresponding in statistics to bias), which is the fraction of observed examples misclassified by h and a *hypothesis complexity* term (corresponding to variance in statistics) which measures the richness of the class of all admissible hypotheses [Was06]. Keeping in line with the literature, we ignore the measure-theoretic technicalities associated with taking suprema over uncountable function classes.

3 Generalization bounds

In this section, we take a preliminary step towards our efficient classification algorithm by deriving generalization bounds for Lipschitz classifiers over doubling spaces. As noted by [vLB04] Lipschitz functions are the natural object to consider in an optimization/regularization framework. The basic intuition behind our proofs is that the Lipschitz constant plays the role of the inverse margin in the confidence of the classifier. As in [vLB04], small Lipschitz constant corresponds to large margin, which in turn yields low hypothesis complexity and variance. In retrospect, our generalization bound (Corollary 5 below) can be derived as a consequence of [vLB04, Theorem 18] in conjunction with [BM02, Theorem 5(b)].

We apply tools from generalized Vapnik-Chervonenkis theory to the case of Lipschitz classifiers. Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$ and recall the definition of the fat-shattering dimension [ABCH97, BS99]: a set $X \subset \mathcal{X}$ is said to be γ-shattered by $\mathcal{F}$ if there exists some function $r : X \to \mathbb{R}$ such that for each label assignment $y \in \{-1, 1\}^X$ there is an $f \in \mathcal{F}$ satisfying $y(x)(f(x) - r(x)) \ge \gamma > 0$ for all $x \in X$. The γ-fat-shattering dimension of $\mathcal{F}$, denoted by $\mathrm{fat}_\gamma(\mathcal{F})$, is the cardinality of the largest set γ-shattered by $\mathcal{F}$.

For the case of Lipschitz functions, we will show that the notion of fat-shattering dimension may be somewhat simplified. We say that a set $X \subset \mathcal{X}$ is γ-shattered *at zero* by a collection of functions $\mathcal{F}$ if for each $y \in \{-1, 1\}^X$ there is an $f \in \mathcal{F}$ satisfying $y(x)f(x) \ge \gamma$ for all $x \in X$. (This is the definition above with $r \equiv 0$.) We write $\mathrm{fat}_\gamma^0(\mathcal{F})$ to denote the cardinality of the largest set γ-shattered at zero by $\mathcal{F}$ and show that for Lipschitz function classes the two complexity measures are the same.

Lemma 1 *Let $\mathcal{F}$ be the collection of all $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \le L$. Then* $\mathrm{fat}_\gamma(\mathcal{F}) = \mathrm{fat}_\gamma^0(\mathcal{F})$.

Proof: We begin by recalling the classic Lipschitz extension result, essentially due to McShane and Whitney [McS34, Whi34]. Any real-valued function f defined on a subset X of a metric space $\mathcal{X}$ has an extension f^* to all of $\mathcal{X}$ satisfying $\|f^*\|_{\text{Lip}} = \|f\|_{\text{Lip}}$. Thus, in what follows we will assume that any function f defined on $X \subset \mathcal{X}$ is also defined on all of $\mathcal{X}$ via some Lipschitz extension (in particular, to bound $\|f\|_{\text{Lip}}$ it suffices to bound the restricted $\| f|_X \|_{\text{Lip}}$).

Consider some finite $X \subset \mathcal{X}$. If X is γ-shattered at zero by $\mathcal{F}$ then by definition it is also γ-shattered. Now assume that X is γ-shattered by $\mathcal{F}$. Thus, there is some function $r : X \to \mathbb{R}$ such that for each $y \in \{-1,1\}^X$ there is an $f = f_{r,y} \in \mathcal{F}$ such that $f_{r,y}(x) \geq r(x) + \gamma$ if $y(x) = +1$ and $f_{r,y}(x) \leq r(x) - \gamma$ if $y(x) = -1$. Let us define the function $\tilde{f}_y$ on X and as per above, on all of $\mathcal{X}$, by $\tilde{f}_y(x) = \gamma y(x)$. It is clear that the collection $\left\{\tilde{f}_y : y \in \{-1,1\}^X\right\}$ γ-fat-shatters X at zero; it only remains to verify that $\tilde{f}_y \in \mathcal{F}$, i.e.,

$$\sup_{y \in \{-1,1\}^X} \left\|\tilde{f}_y\right\|_{\text{Lip}} \leq \sup_{y \in \{-1,1\}^X} \|f_{r,y}\|_{\text{Lip}}.$$

Indeed,

$$\sup_{y \in \{-1,1\}^X, x, x' \in X} \frac{f_{r,y}(x) - f_{r,y}(x')}{\rho(x,x')} \geq \sup_{x,x' \in X} \frac{r(x) - r(x') + 2\gamma}{\rho(x,x')} \geq \sup_{x,x' \in X} \frac{2\gamma}{\rho(x,x')} = \sup_{y \in \{-1,1\}^X} \left\|\tilde{f}_y\right\|_{\text{Lip}}.$$

■

A consequence of Lemma 1 is that in considering the generalization properties of Lipschitz functions we need only bound the γ-fat-shattering dimension at zero. The latter follows from the observation that the packing number of a metric space controls the fat-shattering dimension of Lipschitz functions defined over the metric space. Let $M(\mathcal{X}, \rho, \varepsilon)$ be defined as the ε-packing number of $\mathcal{X}$, the cardinality of the largest ε-separated subset of $\mathcal{X}$.

Theorem 2 *Let $(\mathcal{X}, \rho)$ be a metric space. Fix some $L > 0$, and let $\mathcal{F}$ be the collection of all $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{\text{Lip}} \leq L$. Then for all $\gamma > 0$,*

$$\text{fat}_\gamma(\mathcal{F}) = \text{fat}_\gamma^0(\mathcal{F}) \leq M(\mathcal{X}, \rho, 2\gamma/L).$$

Proof: Suppose that $S \subseteq \mathcal{X}$ is fat γ-shattered at zero. The case $|S| = 1$ is trivial, so we assume the existence of $x \neq x' \in S$ and $f \in \mathcal{F}$ such that $f(x) \geq \gamma > -\gamma \geq f(x')$. The Lipschitz property then implies that $\rho(x,x') \geq 2\gamma/L$, and the claim follows. ■

Corollary 3 *Let metric space $\mathcal{X}$ have doubling dimension $\text{ddim}(\mathcal{X})$, and let $\mathcal{F}$ be the collection of real-valued functions over $\mathcal{X}$ with Lipschitz constant at most L. Then for all $\gamma > 0$,*

$$\text{fat}_\gamma(\mathcal{F}) \leq \left\lceil \frac{L\text{diam}(\mathcal{X})}{2\gamma} \right\rceil^{\text{ddim}(\mathcal{X})+1}.$$

Proof: The claim follows immediately from Theorem 2 and the packing property of doubling spaces. ■

Equipped with these estimates for the fat-shattering dimension of Lipschitz classifiers, we can invoke a standard generalization bound stated in terms of this quantity. For the remainder of this section, we take $\gamma = 1$ and say that a function f classifies an example (x_i, y_i) correctly if

$$y_i f(x_i) \geq 1. \tag{1}$$

The following generalization bounds appear in [BS99]:

Theorem 4 *Let $\mathcal{F}$ be a collection of real-valued functions over some set $\mathcal{X}$, define $d = \text{fat}_{1/16}(\mathcal{F})$ and let and P be some probability distribution on $\mathcal{X} \times \{-1,1\}$. Suppose that (x_i, y_i), $i = 1, \ldots, n$ are drawn from $\mathcal{X} \times \{-1,1\}$ independently according to P and that some $f \in \mathcal{F}$ classifies the n examples correctly, in the sense of (1). Then with probability at least $1 - \delta$*

$$P\{(x,y) : \text{sgn}(f(x)) \neq y\} \leq \frac{2}{n}\left(d \log_2(34en/d) \log_2(578n) + \log_2(4/\delta)\right).$$

Furthermore, if $f \in \mathcal{F}$ is correct on all but k examples, we have with probability at least $1 - \delta$

$$P\{(x,y) : \text{sgn}(f(x)) \neq y\} \leq k/n + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}.$$

Applying Corollary 3, we obtain the following consequence of Theorem 4:

Corollary 5 *Let metric space $\mathcal{X}$ have doubling dimension* $\mathrm{ddim}(\mathcal{X})$*, and let $\mathcal{F}$ be the collection of real-valued functions over $\mathcal{X}$ with Lipschitz constant at most L. Then for any $f \in \mathcal{F}$ that classifies a sample of size n correctly, we have with probability at least* $1 - \delta$

$$P\{(x,y) : \mathrm{sgn}(f(x)) \neq y\} \quad \leq \quad \frac{2}{n}\left(d \log_2(34en/d) \log_2(578n) + \log_2(4/\delta)\right).$$

Likewise, if f is correct on all but k examples, we have with probability at least $1 - \delta$

$$P\{(x,y) : \mathrm{sgn}(f(x)) \neq y\} \quad \leq \quad k/n + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}. \qquad (2)$$

In both cases, $d = \mathrm{fat}_{1/16}(\mathcal{F}) \leq \lceil 8L\mathrm{diam}(\mathcal{X})\rceil^{\mathrm{ddim}(\mathcal{X})+1}$.

4 Lipschitz extension classifier

Given a labeled set $(X, Y) \subset \mathcal{X} \times \{-1, 1\}$, we construct our classifier in a similar manner to [vLB04, Lemma 12], via a Lipschitz extension of the labels Y to all of $\mathcal{X}$. Let $S^+, S^- \subset X$ be the sets of positive and negative labeled points that the classifier correctly labels. Our starting point is the same extension function used in [vLB04], namely, for all $\alpha \in [0, 1]$

$$f_\alpha = \alpha \min_i \left(y_i + 2\frac{d(x, x_i)}{d(S^+, S^-)}\right) + (1 - \alpha) \max_j \left(y_j - 2\frac{d(x, x_j)}{d(S^+, S^-)}\right).$$

However, evaluating the exact value of $f_\alpha(x)$ for each point $x \subset \mathcal{X}$ (or even the sign of $f_\alpha(x)$ at this point) requires an exact nearest neighbor search, and in arbitrary metric space nearest neighbor search may require $\Theta(|X|)$ time.

In this section, we give a classifier whose sign can be evaluated using a $(1 + \varepsilon)$-approximate nearest neighbor search. There exists a search structure for an n point set that can be built in $2^{O(\mathrm{ddim}(\mathcal{X}))} n \log n$ time and supports approximate nearest neighbor searches in time $2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \varepsilon^{-O(\mathrm{ddim}(\mathcal{X}))}$ [CG06, HM06] (see also [KL04, BKL06]). In constructing the classifier, we assume that the sample points have already been partitioned in a manner that yields a favorable bias-variance tradeoff, as in Section 5 below. Therefore, the algorithm below takes as input a set of point $S_1 \subset X$ that must be correctly classified, and a set of error points $S_0 = X - S_1$ that may be ignored in the classifier construction (but which affect the resulting generalization bound).

Theorem 6 *Let $\mathcal{X}$ be a metric space, and fix $0 < \varepsilon \leq \frac{1}{2}$. Given a labeled sample $S = (x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \ldots, n$, let S be partitioned into S_0 and S_1, of sizes k and $n - k$, where S_0 contains points that may be misclassified, and S_1 contains points that may not be misclassified. Define $S_1^+, S_1^- \subset S_1$ according to their labels and define $L = 2/d(S_1^+, S_1^-)$. Then there exists a binary classification function $h : \mathcal{X} \to \{-1, 1\}$ satisfying the following:*

(a) $h(x)$ can be evaluated at each $x \in \mathcal{X}$ via a single $(1 + \varepsilon)$-nearest neighbor query. In particular, $h(x)$ can be evaluated in time $2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \varepsilon^{-O(\mathrm{ddim}(\mathcal{X}))}$, after an initial computation of $(2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \varepsilon^{-O(\mathrm{ddim}(\mathcal{X}))})n$ time.

(b) With probability at least $1 - \delta$

$$P\{(x,y) : h(x) \neq y\} \quad \leq \quad 2\left(\frac{k}{n} + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}\right)$$

where $d = \lceil 8(1 + \varepsilon)L\mathrm{diam}(\mathcal{X})\rceil^{\mathrm{ddim}(\mathcal{X})+1}$.

Proof: Let the distance function $\tilde{d}(\cdot, \cdot)$ be the approximate distance between a point and a set (or between two sets), as determined by a fixed $(1 + \frac{\varepsilon}{4})$-nearest neighbor search structure. Let

$$\tilde{f}_1(x) := \min_i \left(y_i + 2\frac{\tilde{d}(x, x_i)}{\tilde{d}(S_1^+, S_1^-)}\right),$$

and let the classifier be $h(x) := \mathrm{sgn}(\tilde{f}_1(x))$. $h(x)$ can be evaluated via an approximate nearest neighbor query in time $2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \varepsilon^{-O(\mathrm{ddim}(\mathcal{X}))}$, assuming that a search structure has been precomputed in time $2^{O(\mathrm{ddim}(\mathcal{X}))} n \log n$, and $\tilde{d}(S_1^+, S_1^-)$ has been precomputed via $O(n)$ nearest neighbor searches in time $(2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \varepsilon^{-O(\mathrm{ddim}(\mathcal{X}))})n$.

It remains to bound the generalization error of h. To this end, define

$$f_1^+(x) = (1+\varepsilon)f_1(x) + \varepsilon = (1+\varepsilon)\min_i \left(y_i + 2\frac{d(x,x_i)}{d(S_1^+, S_1^-)} \right) + \varepsilon,$$

$$f_1^-(x) = (1+\varepsilon)f_1(x) - \varepsilon = (1+\varepsilon)\min_i \left(y_i + 2\frac{d(x,x_i)}{d(S_1^+, S_1^-)} \right) - \varepsilon.$$

Note that $f_1^+(x) > f_1^-(x)$. Both $f_1^+(x)$ and $f_1^-(x)$ correctly classify all labeled points of S_1 and have Lipschitz constant $(1+\varepsilon)L$, so their classification bounds are given by Corollary 5 with this Lipschitz constant.

We claim that $h(x)$ always agrees with the sign of at least one of $f_1^+(x)$ and $f_1^-(x)$: If $f_1^+(x)$ and $f_1^-(x)$ disagree in their sign, then the claim follows trivially. Assume then that the signs of $f_1^+(x)$ and $f_1^-(x)$ agree. Suppose that $f_1^+(x)$ and $f_1^-(x)$ are positive, which implies that $y_j + 2\frac{d(x,x_j)}{d(S_1^+,S_1^-)} > \frac{\varepsilon}{1+\varepsilon}$ for all j. Now recall that $\tilde{f}_1(x) = \min_i \left(y_i + 2\frac{\tilde{d}(x,x_i)}{\tilde{d}(S^+,S^-)} \right) \geq \min_i \left(y_i + \frac{2}{(1+\varepsilon/4)^2} \frac{d(x,x_i)}{d(S^+,S^-)} \right)$. If $y_i = +1$, then trivially $h(x)$ is positive. If $y_i = -1$, we have that $2\frac{d(x,x_i)}{d(S^+,S^-)} > \frac{\varepsilon}{1+\varepsilon} + 1 = \frac{1+2\varepsilon}{1+\varepsilon}$, and so $\tilde{f}_1(x) \geq \min_i \left(y_i + \frac{2}{(1+\varepsilon/4)^2} \frac{d(x,x_i)}{d(S^+,S^-)} \right) > -1 + \frac{1}{(1+\varepsilon/4)^2} \left(\frac{1+2\varepsilon}{1+\varepsilon} \right) > 0$, and we are done. Suppose then that $f_1^+(x)$ and $f_1^-(x)$ are negative, which implies that $y_j + 2\frac{d(x,x_j)}{d(S_1^+,S_1^-)} < -\frac{\varepsilon}{1+\varepsilon}$ for some fixed j. Now it must be that $y_j = -1$, and so $2\frac{d(x,x_j)}{d(S^+,S^-)} < -\frac{\varepsilon}{1+\varepsilon} + 1 = \frac{1}{1+\varepsilon}$. Now recall that $\tilde{f}_1(x) = \min_i \left(y_i + 2\frac{\tilde{d}(x,x_i)}{\tilde{d}(S^+,S^-)} \right) \leq \left(y_j + 2(1+\varepsilon/4)^2 \frac{d(x,x_j)}{d(S^+,S^-)} \right) < -1 + (1+\varepsilon/4)^2 \left(\frac{1}{1+\varepsilon} \right) < 0$, and we are done.

It follows that if $h(x)$ misclassifies x, then x must be misclassified by at least one of $f_1^+(x)$ and $f_1^-(x)$. Hence, the generalization bound of $h(x)$ is not greater than the sum of the generalization bounds of $f_1^+(x)$ and $f_1^-(x)$. ■

5 Bias-variance tradeoffs

In this section, we show how to efficiently construct a classifier that optimizes the bias-variance tradeoff implicit in Corollary 5, equation (2). Let $\mathcal{X}$ be a metric space, and assume we are given a labeled sample $S = (x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$. For any Lipschitz constant L, let $k(L)$ be the minimal sample error of S over all classifiers with Lipschitz constant L. We rewrite the generalization bound as follows:

$$G(L) = P\left\{(x,y) : \mathrm{sgn}(f(x)) \neq y\right\} \leq k(L)/n + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}$$

where $d = \lceil 8L\mathrm{diam}(\mathcal{X}) \rceil^{\mathrm{ddim}(\mathcal{X})+1}$. This bound contains a free parameter, L, which may be tuned to optimize the bias-variance tradeoff. More precisely, decreasing L drives the bias term (number of mistakes) up and the variance term (fat-shattering dimension) down. For some optimal values of L, $G(L)$ achieves a minimum value. The following theorem gives our bias-variance tradeoff.

Theorem 7 *Let $\mathcal{X}$ be a metric space. Given a labeled sample $S = (x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \ldots, n$, there exists a binary classification function $h : \mathcal{X} \to \{-1, 1\}$ satisfying the following properties:*

(a) $h(x)$ can be evaluated at each $x \in \mathcal{X}$ in time $2^{O(\mathrm{ddim}(\mathcal{X}))} \log n$, after an initial computation of $O(n^2 \log n)$ time.

(b) The generalization error of h is bound by

$$P\left\{(x,y) : \mathrm{sgn}(f(x)) \neq y\right\} \quad \leq \quad c \cdot \inf_{L>0} \left(k(L)/n + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)} \right).$$

for some constant c, and where $d = d(L) = \lceil 8L\mathrm{diam}(\mathcal{X}) \rceil^{\mathrm{ddim}(\mathcal{X})+1}$.

We proceed with a description of the algorithm. We will first give an algorithm with runtime $O(n^{4.376})$, and then improve the runtime to $O(n^2 \log n)$.

Algorithm description. Here we give a randomized algorithm that finds an optimal value L^*, that is $G(L^*) = \inf_{L>0} G(L)$. The runtime of this algorithm is $O(n^{4.376})$ with high probability.

First note the behavior of $k(L)$ as L increases. $k(L)$ may decrease when the value of L crosses some critical value: This critical value is determined by point pairs $x_i \in S^+, x_j \in S^-$ (that is, $L = \frac{2}{d(x_i,x_j)}$) and implies that the classification function can correctly classify both these points. There are $O(n^2)$ critical values of L, and these can be determined by enumerating all interpoint distances between sets $S^+, S^- \subset S$.

Below, we will show that for any given L, the value $k(L)$ can be computed in randomized time $O(n^{2.376})$. More precisely, we will show how to compute a partition of S into sets S_1 (with Lipschitz constant L) and S_0 (of size $k(L)$) in this time. Given sets $S_0, S_1 \subset S$, we can construct the classifier of Corollary 5. Since there are $O(n^2)$ critical values of L, we can calculate $k(L)$ for each critical value in $O(n^{4.376})$ total time, and thereby determine L^*. Then by Corollary 5, we may compute a classifier with a bias-variance tradeoff arbitrarily close to optimal.

It is left to describe how value $k(L)$ is computed for any L in randomized time $O(n^{2.376})$. Consider the following algorithm: Construct a bipartite graph $G = (V^+, V^-, E)$. The vertex sets V^+, V^- correspond to the labeled sets $S^+, S^- \in S$, respectively. The length of edge $e = (u, v)$ connecting vertices $u \in V^+$ and $v \in V^-$ is equal to the distance between the points, and E includes all edges of length less than $2/L$. (E can be computed in $O(n^2 \log n)$ time.) Now, for all edges $e \in E$, a classifier with Lipschitz constant L necessarily misclassifies at least one endpoint of e. Hence, the problem of finding a classifier with Lipschitz constant L that misclassifies a minimum number of points in S is equivalent to finding a minimum vertex cover for bipartite graph G. By König's theorem, minimum bipartite vertex cover is itself equivalent to the maximum matching problem on bipartite graphs. An exact solution to the bipartite matching problem may be computed in randomized time $O(n^{2.376})$ [MS04]. This solution immediately identifies sets S_0, S_1, which allows us to compute a classifier with a bias-variance tradeoff arbitrarily close to optimal.

Improved algorithmic runtime. The runtime given above can be reduced from randomized $O(n^{4.376})$ to deterministic $O(n^2 \log n)$, if we are willing to settle for a generalization bound $G(L)$ within a constant factor of the optimal $G(L^*)$.

The first improvement is in the runtime of the vertex cover algorithm. It is well known that a 2-approximation to the minimum vertex cover on an arbitrary graph can be computed by a greedy algorithm in time $O(|V^+ + V^-| + |E|) = O(n^2)$ [GJ77]. Hence, we may evaluate in $O(n^2)$ time a function $k'(L)$ which satisfies $k(L) \le k'(L) \le 2k(L)$.

The second improvement uses a binary search over the values of L, which allows us to evaluate $k'(L)$ for only $O(\log n)$ values of L, as opposed to all $\Theta(n^2)$ values above. Now, we seek the a value of L for which

$$G'(L) = k'(L)/n + \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}$$

is minimal. Call this value L'. Also note that for all L, $G'(L) \le 2G(L)$, from which it follows that $G'(L') \le 2G(L^*)$. While we cannot efficiently find L', we are able to use a binary search to find a value L for which $G'(L) \le 2G'(L') \le 4G(L^*)$. In particular we seek the minimum value of L for which

$$k'(L)/n \le \sqrt{\frac{2}{n}\left(d \ln(34en/d) \log_2(578n) + \ln(4/\delta)\right)}.$$

Now, decreasing L can only increase $k'(L)$, so the solution to the inequality above necessarily yields an L for which $G'(L) \le 2G'(L') \le 4G(L^*)$. The solution to the inequality can be computed through a binary search on all values of L. By Corollary 5, we can construct a classifier with a bias-variance tradeoff within a factor $4(1+\varepsilon)$ of optimal. The total runtime is $O(n^2 \log n)$.

6 Example: Earthmover metric

To illustrate the potential power of our approach, we now analyze the doubling dimension of an earthmover metric $\mathcal{X}_k$ that is often used in computer vision applications. ($k \ge 2$ is a parameter.) Each point in $\mathcal{X}_k$ is a multiset of size k in the unit square in the Euclidean plane, formally $S \subset [0,1]^2$ and $|S| = k$ (allowing and counting multiplicities). The distance between such sets S, T (i.e. two points in $\mathcal{X}_k$) is given by

$$\mathrm{EMD}(S,T) = \min_{\pi : S \to T} \Big\{ \tfrac{1}{k} \sum_{s \in S} \|s - \pi(s)\|_2 \Big\},$$

where the minimum is over all one-to-one mappings $\pi : S \to T$. In other words, $\mathrm{EMD}(S,T)$ is the minimum-cost matching between the two sets S, T, where costs correspond to Euclidean distance.

Lemma 8 *The earthmover metric X above satisfies* $\mathrm{diam}(\mathcal{X}_k) \le \sqrt{2}$, *and* $\mathrm{ddim}(\mathcal{X}_k) \le O(k \log k)$.

Proof: For the rest of this proof, a point refers to the unit square, not $\mathcal{X}_k$. Fix $r > 0$ and consider a ball (in $\mathcal{X}_k$) of radius r around some S. Let N be an $r/2$-net of the unit square $[0,1]^2$. Now consider all multisets T of size k of the unit square which satisfy the following condition: every point in T belongs to the net N and is within (Euclidean) distance $(k + 1/2)r$ from at least one point of S. Points in such a multiset T are chosen from a collection of size at most $k \cdot \left\lceil \frac{(k+1/2)r}{r/2} \right\rceil^{O(1)} \leq k^{O(1)}$ (by the packing property in the Euclidean plane). Thus, the number of such multisets T is at most $\lambda \leq (k^{O(1)})^k = k^{O(k)}$.

We complete the proof of the lemma, by showing that the r-ball (in $\mathcal{X}_k$) around S is covered by the λ balls of radius $r/2$ whose centers are given by the above multisets T. To see this, consider a multiset S' such that $\mathrm{EMD}(S, S') \leq r$, and let us show that S' is contained in an $r/2$-ball around one of the above multisets T. Observe that every point in S' is within distance at most kr from at least one point of S. By "mapping" each point in S' to its nearest point in the net N, we get a multiset T as above with $\mathrm{EMD}(S', T) \leq r/2$. ■

References

[ABCH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[BKL06] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 97–104, New York, NY, USA, 2006. ACM.

[BLL09] Nader H. Bshouty, Yi Li, and Philip M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6):323 – 335, 2009.

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BS99] Peter Bartlett and John Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In *Advances in kernel methods: support vector learning*, pages 43–54, Cambridge, MA, USA, 1999. MIT Press.

[CG06] Richard Cole and Lee-Ad Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *STOC*, pages 574–583, 2006.

[CH67] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[DL07] Ricky Der and Daniel Lee. Large-Margin Classification in Banach Spaces. In *JMLR Workshop and Conference Proceedings Volume 2: AISTATS 2007*, pages 91–98, 2007.

[GJ77] M. R. Garey and D. S. Johnson. The rectilinear steiner tree problem is np-complete. *SIAM Journal on Applied Mathematics*, 32(4):826–834, 1977.

[GKL03] Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.

[HBS05] Matthias Hein, Olivier Bousquet, and Bernhard Schölkopf. Maximal margin classification for metric spaces. *J. Comput. Syst. Sci.*, 71(3):333–359, 2005.

[HM06] Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.*, 35(5):1148–1184, 2006.

[KL04] Robert Krauthgamer and James R. Lee. Navigating nets: simple algorithms for proximity search. In *SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 798–807, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

[KS94] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994.

[McS34] E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934.

[MS04] Marcin Mucha and Piotr Sankowski. Maximum matchings via gaussian elimination. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 248–255, Washington, DC, USA, 2004. IEEE Computer Society.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2002.

[vLB04] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

[Was06] Larry Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

[Whi34] Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934.