Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Adaptive metric dimensionality reduction

Lee-Ad Gottlieb [a,*], Aryeh Kontorovich [b,1], Robert Krauthgamer [c,2]

[a] Ariel University, Ariel, Israel
[b] Ben-Gurion University of the Negev, Beer Sheva, Israel
[c] Weizmann Institute of Science, Rehovot, Israel

A R T I C L E I N F O

Article history:
Available online xxxx

Keywords:
Metric space
PCA
Dimensionality reduction
Doubling dimension
Rademacher complexity

A B S T R A C T

We study adaptive data-dependent dimensionality reduction in the context of supervised learning in general metric spaces. Our main statistical contribution is a generalization bound for Lipschitz functions in metric spaces that are doubling, or nearly doubling. On the algorithmic front, we describe an analogue of PCA for metric spaces: namely an efficient procedure that approximates the data's intrinsic dimension, which is often much lower than the ambient dimension. Our approach thus leverages the dual benefits of low dimensionality: (1) more efficient algorithms, e.g., for proximity search, and (2) more optimistic generalization bounds.

1. Introduction

Linear classifiers play a central role in supervised learning, with a rich and elegant theory. This setting assumes data is represented as points in a Hilbert space, either explicitly as feature vectors or implicitly via a kernel. A significant strength of the Hilbert-space model is its inner-product structure, which has been exploited statistically and algorithmically by sophisticated techniques from geometric and functional analysis, placing the celebrated hyperplane methods on a solid foundation. However, the success of the Hilbert-space model obscures its limitations—perhaps the most significant of which is that it cannot represent many norms and distance functions that arise naturally in applications. Formally, metrics such as L_1, earthmover, and edit distance cannot be embedded into a Hilbert space without distorting distances by a large factor [1–3]. Indeed, the last decade has seen a growing interest and success in extending the theory of linear classifiers to Banach spaces and even to general metric spaces, see e.g. [4–8].

A key factor in the performance of learning is the dimensionality of the data, which is known to control the learner's efficiency, both statistically, i.e. sample complexity [9], and algorithmically, i.e. computational runtime [10]. This dependence on dimension is true not only for Hilbertian spaces, but also for general metric spaces, where both the sample complexity and the algorithmic runtime can be bounded in terms of the covering number or the doubling dimension [5,11,12].

In this paper, we demonstrate that the learner's statistical and algorithmic efficiency can be controlled by the data's *intrinsic dimensionality*, rather than its *ambient dimension* (e.g., the representation dimension). This provides rigorous confirmation for the informal insight that real-life data (e.g., visual or acoustic signals) can often be learned efficiently because it tends to lie close to low-dimensional manifolds, even when represented in a high-dimensional feature space. Our simple

* Corresponding author.
E-mail addresses: leead@ariel.ac.il (L.-A. Gottlieb), karyeh@cs.bgu.ac.il (A. Kontorovich), robert.krauthgamer@weizmann.ac.il (R. Krauthgamer).
1 This research was partially supported by the Israel Science Foundation (grant #1141/12) and a Yahoo Faculty award.
2 This work was supported in part by a US–Israel BSF grant #2010418, an Israel Science Foundation grant #897/13, and by the Citi Foundation.

http://dx.doi.org/10.1016/j.tcs.2015.10.040
0304-3975/© 2015 Elsevier B.V. All rights reserved.

Please cite this article in press as: L.-A. Gottlieb et al., Adaptive metric dimensionality reduction, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.10.040

and general framework quantifies what it means for data to be *approximately* low-dimensional, and shows how to leverage this for computational and statistical gain.

Most generalization bounds depend on the intrinsic dimension, rather than the ambient one, when the training sample lies *exactly* on a low-dimensional subspace. This phenomenon is indeed immediate in generalization bounds obtained via the empirical Rademacher complexity [13,14], but we are not aware of rigorous analysis that specifies such bounds to the case where the sample is "close" to a low-dimensional subspace.

Our contribution We present learning algorithms and generalization bounds that adapt to the *intrinsic* dimensionality of the data, and can exploit a training set that is *close* to being low-dimensional for improved runtime complexity and statistical accuracy.

We start with the scenario of a Hilbertian space, which is technically simpler. Let the observed sample be $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^N \times \{-1, 1\}$, and suppose that $\{x_1, \dots, x_n\}$ is close to a low-dimensional linear subspace $T \subset \mathbb{R}^N$, in the sense that its distortion $\eta = \frac{1}{n}\sum_i \|x_i - P_T(x_i)\|_2^2$ is small, where $P_T : \mathbb{R}^N \to T$ denotes orthogonal projection onto T. We prove in Section 3 that when $\dim(T)$ and the distortion η are small, a linear classifier generalizes well regardless of the ambient dimension N or the separation margin. Implicit in our result is a statistical tradeoff between the reduced dimension and the distortion, which can be optimized (for the sample at-hand) by performing PCA; see Corollary 3.2.

Our approach quantifies the statistical effect of using PCA, which is commonly done to denoise or improve algorithmic runtime, prior to constructing a linear classifier. We show that for low intrinsic dimension and when the PCA cutoff value is chosen properly, a linear classifier trained on points produced by PCA incurs only a small loss in generalization bounds; see Equation (6).[3]

We then develop this approach significantly beyond the Euclidean case, to the much richer setting of general metric spaces. A completely new challenge that arises here is the algorithmic part, because no metric analogue to dimension reduction via PCA is known. Let the observed sample be $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$, where $(\mathcal{X}, \rho)$ is some metric space. The statistical framework proposed by von Luxburg and Bousquet [5], where classifiers are realized by Lipschitz functions, was extended by Gottlieb et al. [11] to obtain generalization bounds and algorithmic runtime that depend on the metric's doubling dimension, denoted $\mathrm{ddim}(\mathcal{X})$ (see Section 2 for definitions). The present work makes a considerably less restrictive assumption—that the sample points lie *close* to some low-dimensional set.

We establish in Section 4 new generalization bounds for the scenario where there exists some multiset $\tilde{S} = \{\tilde{x}_1, \dots, \tilde{x}_n\}$ of low doubling dimension, whose distortion $\eta = \frac{1}{n}\sum_i \rho(x_i, \tilde{x}_i)$ is small. In this case, the Lipschitz-extension classifier will generalize well, regardless of the ambient dimension $\mathrm{ddim}(\mathcal{X})$; see Theorem 4.5. These generalization bounds feature a tradeoff between the intrinsic dimension and the distortion, which is separate from the tradeoff between the classifier's empirical loss and its smoothness (Lipschitz constant). Here, the first tradeoff must be optimized before addressing the second one.

Hence, we address in Section 5 the computational problem of finding (in polynomial time) a near-optimal point set $\tilde{S}$, given a bound on η. Formally, we devise an algorithm that achieves a bicriteria approximation, meaning that $\mathrm{ddim}(\tilde{S})$ and η of the reported solution exceed the values of a target low-dimensional solution by at most a constant factor; see Theorem 5.1. Using this algorithm, one can optimize the dimension-distortion tradeoff (for the sample at-hand) within a constant factor. Having determined these values, one can compute a Lipschitz classifier that optimizes the tradeoff between empirical loss and smoothness (in the generalization bound), by running the algorithm of Gottlieb et al. [11] on the sample. One may instead run this algorithm on the modified training set (low-dimensional $\tilde{S}$), and then the runtime of constructing the classifier and evaluating it on new points will depend on the intrinsic dimension instead of the ambient dimension. We show that this incurs only a small loss in the generalization bound; see Equation (8).

Related work The topic of dimensionality reduction, even restricted to learning theory, is far too vast to adequately survey within the scope of this paper (for recent surveys, see e.g. [15,16]). We are only aware of dimensionality results with provable guarantees for the Euclidean case—mainly in the context of improving algorithmic runtimes—achieved by projecting data onto a random low-dimensional subspace, see e.g. [17–20]. On the other hand, data-dependent dimensionality reduction techniques have been observed empirically to improve and speed up classification performance. For instance, PCA may be applied as a preprocessing step before learning algorithms such as SVM, or the two can be put together into a combined algorithm, see e.g. [21–24].

There is little existing rigorous work on dimension reduction in general metric spaces. One heuristic approach is Multi-Dimensional Scaling (MDS), which may be viewed as a generalization of PCA [25]. Given a finite metric space $\mathcal{X}$ and a target dimension d, MDS attempts to find a low-distortion embedding of $\mathcal{X}$ into $\mathbb{R}^d$. Thus, its usefulness is limited to metrics that are "nearly" Euclidean, and it will not be effective in general on inherently non-Euclidean metrics such as L_1, earthmover, and edit distance. Furthermore, MDS optimizes a highly non-convex function, and we are not aware of any performance guarantees for this algorithm—even for input metrics that are nearly Euclidean. A different metric dimension reduction problem was considered in [26]: *removing* from an input set S as few points as possible, with the goal of retaining a large

[3] However, our bounds do not explain some reports of empirical *improvements* achieved by PCA preprocessing.

Please cite this article in press as: L.-A. Gottlieb et al., Adaptive metric dimensionality reduction, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.10.040

subset of low doubling dimension. While close in spirit, their objective is technically different from ours, and the problem seems to require rather different techniques.

Algorithmic questions aside, there is the issue *statistical* benefits gained by reducing the dimensionality—or merely by realizing that the data is effectively low-dimensional, or nearly so. Here again, previous work has mainly addressed statistical efficiency in Hilbertian spaces. For SVM classification, a large body of work derives generalization bounds from the empirical properties of the kernel integral operator—its spectrum, entropy numbers of its range on the unit ball, and other measures of "effective dimension" [27–31]. Recently, various geometric notions of "low intrinsic dimensionality" were proposed by Sabato et al. [32] for the purpose of providing improved bounds on the sample complexity of large-margin learning. In addition to "passively" benefiting from low intrinsic dimensionality, one could actively seek to reduce the data-dimension, and some attempts to quantify the resulting statistical benefits have been made [33,34]. In the context of regression, [35] and the series of works cited therein observe that "identifying intrinsic low dimensional structure from a seemingly high dimensional source" yields more optimistic minimax rates.

In general metric spaces, Dasgupta, Kpotufe and coauthors have observed in a series of works that various k-NN methods automatically adapt to low data dimensionality [36–39]. The analysis assumes that the data lies exactly (or almost so) on a low-dimensional subset, and is not amenable to the sort of dimension-distortion tradeoff we study in this paper. Furthermore, the classifier we employ here is based on Lipschitz-extension, which eventually boils down to 1-NN, and enjoys several advantages over k-NN: the effect of approximate proximity search on generalization performance may be quantified [11] and near-optimal sample compression may be achieved [40], without sacrificing the Bayes consistency of k-NN [41].

2. Definitions and notation

We use standard notation and definitions throughout, and assume a familiarity with the basic notions of Euclidean and normed spaces. We write $\mathbb{1}_{\{\cdot\}}$ for the indicator function of the relevant predicate and $\mathrm{sgn}(x) := 2 \cdot \mathbb{1}_{\{x \geq 0\}} - 1$.

Metric spaces A *metric* ρ on a set $\mathcal{X}$ is a positive symmetric function satisfying the triangle inequality $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$; together the two comprise the metric space $(\mathcal{X}, \rho)$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be L-Lipschitz if $|f(x) - f(y)| \leq L \cdot \rho(x, y)$ for all $x, y \in \mathcal{X}$. The *diameter*, denoted by $\mathrm{diam}(\mathcal{X})$, is defined to be $\sup_{x,x' \in \mathcal{X}} \rho(x, x')$.

Doubling dimension For a metric $(\mathcal{X}, \rho)$, let $\lambda_{\mathcal{X}} > 0$ be the smallest value such that every ball in $\mathcal{X}$ can be covered by $\lambda_{\mathcal{X}}$ balls of half the radius. Then $\lambda_{\mathcal{X}}$ is the *doubling constant* of $\mathcal{X}$, and the *doubling dimension* of $\mathcal{X}$ is defined as $\mathrm{ddim}(\mathcal{X}) := \log_2(\lambda_{\mathcal{X}})$. It is well-known that while a d-dimensional Euclidean space, or any subset of it, has doubling dimension $O(d)$; however, low doubling dimension is strictly more general than low Euclidean dimension, see e.g. [42].

We will use $|\cdot|$ to denote the cardinality of finite metric spaces.

Covering numbers The ε-covering number of a metric space $(\mathcal{X}, \rho)$, denoted $\mathcal{N}(\varepsilon, \mathcal{X}, \rho)$, is defined as the smallest number of balls of radius ε that suffices to cover $\mathcal{X}$. The balls may be centered at points of $\mathcal{X}$, or of points in an ambient space including $\mathcal{X}$. Covering numbers may be estimated by repeatedly invoking the doubling property (see e.g. [10]):

Lemma 2.1. *If* $(\mathcal{X}, \rho)$ *is a metric space with* $\mathrm{ddim}(\mathcal{X}) < \infty$ *and* $\mathrm{diam}(\mathcal{X}) < \infty$*, then for all* $0 < \varepsilon < \mathrm{diam}(\mathcal{X})$*,*

$$\mathcal{N}(\varepsilon, \mathcal{X}, \rho) \leq \left(\frac{2\mathrm{diam}(\mathcal{X})}{\varepsilon}\right)^{\mathrm{ddim}(\mathcal{X})}.$$

Learning Our setting in this paper is the *agnostic PAC* learning model, see e.g. [43], where labeled examples (X_i, Y_i) are drawn independently from $\mathcal{X} \times \{-1, 1\}$ according to some unknown probability distribution $\mathbb{P}$. The learner, having observed n labeled examples, produces a hypothesis $h : \mathcal{X} \to \{-1, 1\}$. The *generalization error* $\mathbb{P}(h(X) \neq Y)$ is the probability of misclassifying a new point, although sometimes the 0–1 penalty here is replaced with a loss function. Most generalization bounds consist of a *sample error* term (approximately corresponding to *bias* in statistics), which is the fraction of observed examples misclassified by h and a *hypothesis complexity* term (a rough analogue of *variance* in statistics) which measures the richness of the class of all admissible hypotheses [44]. A data-driven procedure for selecting the correct hypothesis complexity is known as *model selection* and is typically performed by some variant of Structural Risk Minimization [45]—an analogue of the bias-variance tradeoff in statistics. The measure-theoretic technicalities associated with taking suprema over uncountable function classes are typically glossed over in learning literature. However, we note in passing that if $\mathcal{X}$ is a compact metric space and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is a collection of Lipschitz functions, then $\mathcal{F}$ contains a countable $\mathcal{F}' \subset \mathcal{F}$ such that every member of $\mathcal{F}$ is a pointwise limit of a sequence in $\mathcal{F}'$, which more than suffices to guarantee the measurability of the supremum [46].

Rademacher complexity For any n points $z_1, \ldots, z_n$ in some set $\mathcal{Z}$ and any collection of functions F mapping $\mathcal{Z}$ to a bounded range, define the Rademacher complexity of F evaluated at the n points as

Please cite this article in press as: L.-A. Gottlieb et al., Adaptive metric dimensionality reduction, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.10.040

$$\hat{R}_n(F;\{z_i\}) = \frac{1}{n}\mathbb{E}\sup_{g\in F}\sum_{i=1}^{n}\sigma_i g(z_i),$$

where σ_i are i.i.d. random variables that take on ± 1 with probability $1/2$. The seminal work of [13] and [14] established the central role of Rademacher complexities in generalization bounds.

The Rademacher complexity of a binary function class F may be controlled by the VC-dimension d of F through an application of Massart's and Sauer's lemmas:

$$\hat{R}_n(F;\{z_i\}) \le \sqrt{\frac{2d\log(en/d)}{n}}. \tag{1}$$

Considerably more delicate bounds may be obtained by estimating the covering numbers and using Dudley's chaining integral:

$$\hat{R}_n(F;\{z_i\}) \le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_{\alpha}^{\infty}\sqrt{\frac{\log\mathcal{N}(t,F,\|\cdot\|_2)}{n}}dt\right). \tag{2}$$

A proof of (1) may be found in [43], and (2) is (essentially) contained in [47].

3. Adaptive dimensionality reduction: Euclidean case

In this section, we illustrate our statistical approach in the familiar setting of Euclidean spaces, where the algorithmic problem of finding a low-distortion subspace is solved by PCA. Consider the problem of supervised classification in $\mathbb{R}^N$ by linear hyperplanes, where $N \gg 1$. The training sample is (X_i, Y_i), $i = 1, \dots, n$, with $(X_i, Y_i) \in \mathbb{R}^N \times \{-1, 1\}$. Since this set is finite, there is no loss of generality in normalizing $\|X_i\|_2 \le 1$. We consider the hypothesis class $H = \{x \mapsto \mathrm{sgn}(w\cdot x) : \|w\|_2 \le 1\}$. Absent additional assumptions on the data, this is a high-dimensional learning problem with a costly sample complexity. Indeed, the VC-dimension of linear hyperplanes in N dimensions is N. If, however, it turns out that the data actually lies on a k-dimensional subspace of $\mathbb{R}^N$, Eq. (1) implies that $\hat{R}_n(H;\{X_i\}) \le \sqrt{2k\log(en/k)/n}$, and hence a much better generalization for $k \ll N$. A more common distributional assumption is that of large-margin separability. In fact, the main insight articulated in [48] is that data separable by margin γ effectively lies in an $\tilde{O}(1/\gamma^2)$-dimensional space.

Suppose now that the data lies "close" to a low-dimensional subspace. Formally, we say that the data $\{X_i\}$ is *η-close* to a subspace $T \subset \mathbb{R}^N$ if $\frac{1}{n}\sum_{i=1}^{n}\|X_i - P_T(X_i)\|_2^2 \le \eta$ (recall that $P_T(\cdot)$ denotes the orthogonal projection onto the subspace T). Whenever this holds, the Rademacher complexity can be bounded in terms of $\dim(T)$ and η alone (Theorem 3.1). As a consequence, we obtain a bound on the expected hinge-loss (Corollary 3.2).

Theorem 3.1. *Let $X_1, \dots, X_n$ lie in $\mathbb{R}^N$ with $\|X_i\|_2 \le 1$ and define the function class $F = \{x \mapsto w\cdot x : \|w\|_2 \le 1\}$. Suppose that the data $\{X_i\}$ is η-close to some subspace $T \subset \mathbb{R}^N$ and $\eta \ge 0$. Then $\hat{R}_n(F;\{X_i\}) \le 17\sqrt{\frac{\dim(T)}{n}} + \sqrt{\frac{\eta}{n}}$.*

Remark. Notice that our Rademacher complexity bound is independent of the ambient dimension N. Our bound exhibits a tension between $\dim(T)$ and η,—as we seek a lower-dimensional approximation, we are liable to incur a larger distortion—and optimizing the tradeoff between them amounts to choosing a PCA cutoff value.

Proof. Denote by $S^{\|} = (X_1^{\|}, \dots, X_n^{\|})$ and $S^{\perp} = (X_1^{\perp}, \dots, X_n^{\perp})$ the parallel and perpendicular components of the points $\{X_i\}$ with respect to T. Note that each X_i has the unique decomposition $X_i = X_i^{\|} + X_i^{\perp}$. We first decompose the Rademacher complexity into "parallel" and "perpendicular" terms:

$$\begin{aligned}\hat{R}_n(F;\{X_i\}) &= \frac{1}{n}\mathbb{E}_{\sigma}\left[\sup_{\|w\|\le 1}\sum_{i=1}^{n}\sigma_i(w\cdot X_i)\right]\\ &= \frac{1}{n}\mathbb{E}_{\sigma}\left[\sup_{\|w\|\le 1} w\cdot\sum_{i=1}^{n}\sigma_i(X_i^{\|} + X_i^{\perp})\right]\\ &\le \hat{R}_n(F;S^{\|}) + \hat{R}_n(F;S^{\perp}).\end{aligned} \tag{3}$$

We then proceed to bound the two terms in (3). To bound the first term, note that F restricted to T is a function class with linear-algebraic dimension $\dim(T)$, and furthermore our assumption that the data lies in the unit ball implies that the range of F is bounded by 1 in absolute value. Hence, the classic covering number estimate (see [49])

$$\mathcal{N}(F,t,\|\cdot\|_2) \le \left(\frac{3}{t}\right)^{\dim(T)}, \qquad \forall t \in (0,1) \tag{4}$$

applies. Substituting (4) into Dudley's integral (2) yields

$$\hat{R}_n(F; S^{\|}) \le 12 \int_0^\infty \sqrt{\frac{\log \mathcal{N}(t, F, \|\cdot\|_2)}{n}} dt$$
$$\le 12 \int_0^1 \sqrt{\frac{\dim(T)\log(3/t)}{n}} dt \le 17\sqrt{\frac{\dim(T)}{n}}. \tag{5}$$

The second term in (3) is bounded via a standard calculation:

$$\hat{R}_n(F; S^{\perp}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\|w\|_2\le 1} w \cdot \sum_{i=1}^n \sigma_i X_i^{\perp}\right] = \frac{1}{n}\mathbb{E}_\sigma\left\|\sum_{i=1}^n \sigma_i X_i^{\perp}\right\|_2$$
$$\le \frac{1}{n}\left(\mathbb{E}_\sigma\left\|\sum_{i=1}^n \sigma_i X_i^{\perp}\right\|_2^2\right)^{1/2},$$

where the second equality follows from the dual characterization of the ℓ_2 norm and the inequality is Jensen's. Now by independence of the Rademacher variables σ_i,

$$\mathbb{E}_\sigma\left\|\sum_{i=1}^n \sigma_i X_i^{\perp}\right\|_2^2 = \mathbb{E}_\sigma \sum_{1\le i,j\le n} \sigma_i\sigma_j (X_i^{\perp}\cdot X_j^{\perp}) = \sum_{i=1}^n \left\|X_i^{\perp}\right\|_2^2$$
$$= \sum_{i=1}^n \|P_T(X_i) - X_i\|_2^2 \le n\eta,$$

which implies $\hat{R}_n(F; S^{\perp}) \le \sqrt{\eta/n}$ and together with (3) and (5) proves the claim. □

Corollary 3.2. *Let (X_i, Y_i) be an i.i.d. sample of size n, where each $X_i \in \mathbb{R}^N$ satisfies $\|X_i\|_2 \le 1$. Then for all $\delta > 0$, with probability at least $1-\delta$, for every $w \in \mathbb{R}^N$ with $\|w\|_2 \le 1$, and every k-dimensional subspace T to which the sample is η-close, we have*

$$\mathbb{E}[L(w\cdot X, Y)] \le \frac{1}{n}\sum_{i=1}^n L(w\cdot X_i, Y_i) + 34\sqrt{\frac{k}{n}} + 2\sqrt{\frac{\eta}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $L(u, y) = (1-uy)_+$ is the hinge loss.

Proof. Follows from the Rademacher generalization bound [43, Theorem 3.1], the complexity estimate in Theorem 3.1, and an application of Talagrand's contraction lemma [50] to incorporate the hinge loss. □

Remark. The expected loss is bounded in Corollary 3.2 in terms of the empirical loss on the *original sample points* $\{X_i\}$; we can easily bound it also in terms of the empirical loss on the *projected points* $\{P_T(X_i)\}$, because the hinge-loss is 1-Lipschitz, and thus

$$\left|\frac{1}{n}\sum_{i=1}^n L(w\cdot X_i, Y_i) - \frac{1}{n}\sum_{i=1}^n L(w\cdot X_i^{\|}, Y_i)\right| \le \frac{1}{n}\sum_{i=1}^n \left\|X_i^{\perp}\right\|_2 \le \left(\frac{1}{n}\sum_{i=1}^n \left\|X_i^{\perp}\right\|_2^2\right)^{1/2} \le \sqrt{\eta}. \tag{6}$$

The linear classifier (choice of w) can therefore be computed by considering either the original sample or the projected points.

4. Adaptive dimensionality reduction: metric case

In this section we extend the statistical analysis of Section 3 from Euclidean spaces to the general metric case. Suppose $(\mathcal{X}, \rho)$ is a metric space and we receive the training sample (X_i, Y_i), $i = 1, \dots, n$, with $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$. Following von Luxburg and Bousquet [5] and Gottlieb et al. [11], the classifier we construct will be a Lipschitz function (whose predictions are computed via Lipschitz extension that in turn uses approximate nearest neighbor search)—but with the added twist of a dimensionality reduction preprocessing step. To motivate this approach, let us recall some results from [5, 11]. von Luxburg and Bousquet [5] made the simple but powerful observation that 1-Nearest-Neighbor (1-NN) classification is essentially equivalent to computing a real-valued Lipschitz extension h from the ± 1-labeled sample points, and classifying test points by thresholding h at 0. This made the 1-NN classifier amenable to analysis by Rademacher complexity and fat-shattering dimension techniques. In particular, the 1-NN classifier (and hence the Lipschitz extension it induces) can be computed efficiently within any fixed additive precision, formalized as follows.

Theorem 4.1. *(See [11].) Let $(\mathcal{X}, \rho)$ be a metric space, and fix $0 < \varepsilon < \frac{1}{32}$. For a sample S consisting of n labeled points $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$, with $d = \mathrm{ddim}(\{x_1, \dots, x_n\})$, let $f^* : \mathcal{X} \to \mathbb{R}$ be a* Lipschitz extension *of S, meaning that (a) $f^*(x_i) = y_i$ for all $i \in [n]$, (b) f^* is L-Lipschitz (c) there is no $f' : \mathcal{X} \to \mathbb{R}$ satisfying (a) with a Lipschitz constant smaller than L. Then there is a function $\tilde{f} : \mathcal{X} \to \mathbb{R}$ such that*

(i) *$\tilde{f}(x)$ can be evaluated at each $x \in \mathcal{X}$ in time $2^{O(d)} \log n + \varepsilon^{-O(d)}$, after an initial computation of $(2^{O(d)} \log n + \varepsilon^{-O(d)})n$ time;*
(ii) *$|\tilde{f}(x) - f^*(x)| \le 2\varepsilon$ for all $x \in \mathcal{X}$.*

As shown ibid., the ε-approximation causes a very mild degradation of the generalization bounds. The exponential dependence of the runtime—both the precomputation and the evaluation on new points—on $\mathrm{ddim}(S)$ stands to benefit a great deal from even a slight reduction in dimensionality.

The remainder of this section is organized as follows. In Section 4.1, we formalize the notion of "nearly" low-dimensional data in a metric space and discuss its implication for Rademacher complexity. We say that $S = \{x_i\} \subset \mathcal{X}$ is (η, d)-*elastic* if there is a $\tilde{S} = \{\tilde{x}_i\} \subset \mathcal{X}$ such that $\frac{1}{n} \sum_{i=1}^{n} \rho(x_i, \tilde{x}_i) \le \eta$ and $\mathrm{ddim}(\tilde{S}) \le d$. We can prove that if our data set is (η, d)-elastic, then the Rademacher complexity it induces on Lipschitz functions can be bounded in terms of η and d alone (Theorem 4.4), independently of the ambient dimension $\mathrm{ddim}(\mathcal{X})$. Similarly to the Euclidean case (Theorem 3.1), we then use in Section 4.2 these Rademacher complexity estimates to obtain data-dependent error bounds (Theorem 4.5).

In Section 4.3, we describe how to convert our distortion-based statistical bounds (for the Rademacher complexity and generalization error) into an effective classification procedure. To this end, we develop a novel bicriteria approximation algorithm presented in Section 5. Informally, given a set $S \subset \mathcal{X}$ and a target doubling dimension d, our method efficiently computes a set $\tilde{S}$ with $\mathrm{ddim}(\tilde{S}) \approx d$ and approximately minimal the distortion η. In the sample-preprocessing step that is applied before classifying new points, we iterate the bicriteria algorithm to find a near-optimal tradeoff between dimensionality and distortion. Having found an $\tilde{S}$ achieving near-optimal tradeoff, we apply on it the machinery of Theorem 4.1, and exploit its low dimensionality for fast approximate nearest-neighbor search.

4.1. Rademacher bounds

We begin by obtaining complexity estimates for Lipschitz functions in (nearly) doubling spaces. This was done in [11] in terms of the fat-shattering dimension, but here we obtain data-dependent bounds by direct control over the covering numbers. The following "covering numbers by covering numbers" lemma is a variant of the classic estimate [51].

Lemma 4.2. *Let F_L be the collection of L-Lipschitz functions mapping the metric space $(\mathcal{X}, \rho)$ to $[-1, 1]$, and endow F_L with the L_∞ metric:*

$$\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|, \qquad f, g \in F_L.$$

Then the covering numbers of F_L may be estimated in terms of the covering numbers of $\mathcal{X}$:

$$\mathcal{N}(\varepsilon, F_L, \|\cdot\|_\infty) \le \left(\frac{8}{\varepsilon}\right)^{\mathcal{N}(\varepsilon/2L, \mathcal{X}, \rho)}, \qquad \forall \varepsilon \in (0, 1).$$

Hence, for a space $(\mathcal{X}, \rho)$ with diameter 1,

$$\log \mathcal{N}(\varepsilon, F_L, \|\cdot\|_\infty) \le \left(\frac{4L}{\varepsilon}\right)^{\mathrm{ddim}(\mathcal{X})} \log\left(\frac{8}{\varepsilon}\right), \qquad \forall \varepsilon \in (0, 1).$$

Equipped with the covering numbers estimate, we proceed to bound the Rademacher complexity of Lipschitz functions on doubling spaces.[4]

Theorem 4.3. *Let F_L be the collection of L-Lipschitz $[-1, 1]$-valued functions defined on a metric space (S, ρ) with diameter 1 and doubling dimension d. Then $\hat{R}_n(F_L; S) = O\left(\frac{L}{n^{1/(d+1)}}\right)$.*

[4] Analogous bounds were obtained by [5] in less explicit form.

Proof. Recall that $\|f\|_2 \le \|f\|_\infty$ implies $\mathcal{N}(\varepsilon, F, \|\cdot\|_2) \le \mathcal{N}(\varepsilon, F, \|\cdot\|_\infty)$. Substituting the estimate in Lemma 4.2 into Dudley's integral (2), we have

$$
\begin{aligned}
\hat{R}_n(F_L; S) &\le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_\alpha^\infty \sqrt{\frac{\log \mathcal{N}(t, F_L, \|\cdot\|_\infty)}{n}}dt\right)\\
&\le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_\alpha^2 \sqrt{\frac{\left(\frac{4L}{t}\right)^d \log\left(\frac{8}{t}\right)}{n}}dt\right)\\
&\le \inf_{\alpha\ge 0}\left(4\alpha + 12\int_\alpha^2 \sqrt{\frac{\left(\frac{4L}{t}\right)^d \left(\frac{8}{t}\right)}{n}}dt\right)\\
&\le \inf_{\alpha\ge 0}\left(4\alpha + \frac{34(4L)^{d/2}}{\sqrt{n}}\int_\alpha^2 \left(\frac{1}{t}\right)^{(d+1)/2} dt\right)\\
&= \inf_{\alpha\ge 0}\left(4\alpha + \frac{34(4L)^{d/2}}{\sqrt{n}}\left(\frac{d-1}{2}\right)\left(\frac{1}{\alpha^{(d-1)/2}} - 2^{-(d+1)/2}\right)\right)\\
&= 4^{\frac{d-2}{d+1}}((d-1)K)^{\frac{2}{d+1}} + K\left(\left(8^{-\frac{2}{d+1}}((d-1)K)^{\frac{2}{d+1}}\right)^{\frac{d-1}{2}} - 2^{-(d+1)/2}\right)\\
&\le 8K^{\frac{2}{d+1}} + dK\left(K^{\frac{d-1}{d+1}} - 2^{-(d+1)/2}\right) = O(K^{\frac{2}{d+1}}),
\end{aligned}
$$

where

$$
K = \frac{34(4L)^{d/2}}{\sqrt{n}}\left(\frac{d-1}{2}\right).
$$

Thus, $\hat{R}_n(F_L; S) = O\left(\frac{L}{n^{1/(d+1)}}\right)$, as claimed. □

We can now quantify the savings earned by a low-distortion dimensionality reduction.

Theorem 4.4. *Let $(\mathcal{X}, \rho)$ be a metric space with diameter 1, and consider the two n-point sets $S, \tilde{S} \subset \mathcal{X}$, where S is (η, d)-elastic with witness $\tilde{S}$. Let F_L be the collection of all L-Lipschitz, $[-1, 1]$-valued functions on $\mathcal{X}$. Then $\hat{R}_n(F_L; S) = O\left(L(1/n^{1/(d+1)} + \eta)\right)$.*

Proof. For $X_i \in S$ and $\tilde{X}_i \in \tilde{S}$, define $\delta_i(f) = f(X_i) - f(\tilde{X}_i)$. Then

$$
\begin{aligned}
\hat{R}_n(F_L; S) &= \mathbb{E}\sup_{f\in F_L}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) = \mathbb{E}\sup_{f\in F_L}\frac{1}{n}\sum_{i=1}^n \sigma_i (f(\tilde{X}_i) + \delta_i(f))\\
&\le \hat{R}_n(F_L; \tilde{S}) + \mathbb{E}\sup_{f\in F_L}\frac{1}{n}\sum_{i=1}^n \sigma_i\delta_i(f).
\end{aligned} \tag{7}
$$

Now the Lipschitz property and our definition of distortion imply that

$$
\left|\sum_{i=1}^n \sigma_i\delta_i(f)\right| \le \sum_{i=1}^n |\delta_i(f)| \le L\sum_{i=1}^n \rho(X_i, \tilde{X}_i) \le Ln\eta,
$$

and hence $\mathbb{E}\sup_{f\in F_L}\frac{1}{n}\sum_{i=1}^n \sigma_i\delta_i(f) \le L\eta$. The other term in (7) is bounded by invoking Theorem 4.3. □

Remark. Comparing the Rademacher estimate $O\left(\frac{1}{n^{1/(d+1)}} + \eta\right)$ from Theorem 4.4 with the bound $O(\sqrt{d/n} + \sqrt{\eta/n})$ from the Euclidean case (Theorem 3.1), we see an exponential gap in the dependence on the dimension. The gap's origin is the bound (4) for linear classifiers compared with Lemma 4.2 for Lipschitz functions (the latter estimate is essentially tight [51]).

4.2. Generalization bounds

For $f : \mathcal{X} \to [-1, 1]$, define the *margin* of f on the labeled example (x, y) by $yf(x)$. The γ*-margin loss*, $0 < \gamma < 1$, that f incurs on (x, y) is $L_\gamma(f(x), y) = \min(\max(0, 1 - yf(x)/\gamma), 1)$, which charges a value of 1 for predicting the wrong sign, i.e., $yf(x) < 0$, charges nothing for predicting correctly with confidence $yf(x) \geq \gamma$, and interpolates linearly between these regimes. Since $L_\gamma(f(x), y) \leq \mathbb{1}_{\{yf(x)<\gamma\}}$, the sample's γ-margin loss is at most its γ-margin misclassification error.

Theorem 4.5. *Let F_L be the collection of L-Lipschitz functions mapping a metric space $\mathcal{X}$ of diameter 1 to $[-1, 1]$. If the i.i.d. sample $(X_i, Y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \ldots, n$, is (η, d)-elastic, then for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in F_L$ and all $\gamma \in (0, 1)$:*

$$\mathbb{P}(\mathrm{sgn}(f(X)) \neq Y) \leq \frac{1}{n}\sum_{i=1}^{n} L_\gamma(f(X_i), Y_i) + O\left(\frac{L(1/n^{1/(d+1)} + \eta)}{\gamma} + \sqrt{\frac{\log\log(L/\gamma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Proof. We invoke [43, Theorem 4.5] to bound the classification error in terms of sample margin loss and Rademacher complexity and the latter is bounded via Theorem 4.4. □

Remark. The expected loss is bounded in Theorem 4.5 in terms of the empirical loss on the *original sample points* $\{X_i\}$; we can easily bound it also in terms of the empirical loss on the *perturbed points* $\left\{\tilde{X}_i\right\}$, because the loss $L_\gamma(f(\cdot), Y_i)$ is (L/γ)-Lipschitz, and thus

$$\left|\frac{1}{n}\sum_{i=1}^{n} L_\gamma(f(X_i), Y_i) - \frac{1}{n}\sum_{i=1}^{n} L_\gamma(f(\tilde{X}_i), Y_i)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\frac{L}{\gamma}\rho(x_i, \tilde{x}_i) \leq \frac{L}{\gamma}\eta. \tag{8}$$

4.3. Classification procedure

Theorem 4.5 provides a statistical optimality criterion for the dimensionality-distortion tradeoff. Unlike the Euclidean case, where the well-known PCA optimized this tradeoff, the metric case requires the bicriteria approximation algorithm which we describe in Section 5. To recap, given a set $S \subset \mathcal{X}$ and a target doubling dimension d, this algorithm efficiently computes a set $\tilde{S}$ with $\mathrm{ddim}(\tilde{S}) \approx d$, which approximately minimizes the distortion η. We may iterate this algorithm over all $d \in \left\{1, \ldots, \log_2|S|\right\}$—since the doubling dimension of the metric space (S, ρ) is at most $\log_2|S|$—to optimize the complexity term[5] in Theorem 4.5.

Once a nearly optimal witness $\tilde{S}$ of (η, d)-elasticity has been constructed, we predict the value at a test point $x \in \mathcal{X}$ by a thresholded Lipschitz extension from $\tilde{S}$, as provided in Theorem 4.1.

5. Approximately optimizing the dimension-distortion tradeoff

In this section we cast the computation of an (η, d)-elasticity witness of a given finite set as an optimization problem and design for it a polynomial-time bicriteria approximation algorithm. Let $(\mathcal{X}, \rho)$ be a finite metric space. For a point v and a point set T, define $\rho(v, T) = \min_{w\in T}\rho(v, w)$. Given two point sets S, T, define the *cost of mapping* S to T to be $\sum_{v\in S}\rho(v, T)$. Define the *Low-Dimensional Mapping (LDM)* problem as follows: Given a point set $S \subseteq \mathcal{X}$ and a target dimension $d \geq 1$, find $T \subseteq S$ with $\mathrm{ddim}(T) \leq d$ such that the cost of mapping S to T is minimized.[6] An (α, β)-bicriteria approximate solution to the LDM problem is a subset $V \subset S$, such that the cost of mapping S to V is at most α times the cost of mapping S to an optimal T (of $\mathrm{ddim}(T) \leq d$), and also $\mathrm{ddim}(V) \leq \beta d$. We prove the following theorem.

Theorem 5.1. *The Low-Dimensional Mapping problem admits an $(O(1), O(1))$-bicriteria approximation in runtime $2^{O(\mathrm{ddim}(S))}n + O(n\log^4 n)$, where $n = |S|$.*

The presentation of the algorithm proceeds in four steps. In the first step, we modify LDM by adding to it yet another constraint: We simply extract from S a point hierarchy $\mathcal{S}$ (see Section 5.1), and require that the LDM solution not only be a subset of S, but that it also possesses a hierarchy which is a sub-hierarchy of $\mathcal{S}$. Lemma 5.2 demonstrates that this additional requirement can be fulfilled without significantly altering the cost and dimension of the optimal solution.

The second step of the presentation is an integer program (IP) which models the modified LDM problem, in Section 5.2. We show that a low-cost solution to the LDM problem implies a low-cost solution to the IP, and vice-versa (Lemma 5.3).

[5] Since L/γ multiplies the main term $1/n^{1/(d+1)} + \eta$ in the error bound, and the other term is of order $\log\log(L/\gamma)$, the optimization may effectively be carried out oblivious to L and γ.

[6] The LDM problem differs from k-median (or k-medoid) in that it imposes a bound on $\mathrm{ddim}(T)$ rather than on $|T|$.

Unfortunately, finding an optimal solution to the IP seems difficult, hence the third step, in Section 5.3, is to relax some of the IP constraints and derive a linear program (LP). We also give a rounding scheme that recovers from the LP solution an integral solution, and show that this integral solution indeed provides an $(O(1), O(1))$-bicriteria approximation (Lemma 5.4). The final step, in Section 5.4, shows that the LP can be solved in the runtime stated above (Lemma 5.5), thereby completing the proof of Theorem 5.1.

Remark. The presented algorithm has very large (though constant) approximation factors. The introduced techniques can yield much tighter bounds, by creating many different point hierarchies instead of only a single one. We have chosen the current presentation for simplicity.

5.1. Point hierarchies

Let S be a point set, and assume by scaling that it has diameter 1 and minimum interpoint distance $\delta > 0$. A *hierarchy* $\mathcal{S}$ of a set S is a sequence of nested sets $S_0 \subseteq \ldots \subseteq S_t$; here, $t = \lceil \log_2(1/\delta) \rceil$ and $S_t = S$, while S_0 consists of a single point. Set S_i must possess a *packing* property, which asserts that $\rho(v, w) \geq 2^{-i}$ for all $v, w \in S_i$, and a *c-covering* property for $c \geq 1$ (with respect to S_{i+1}), which asserts that for each $v \in S_{i+1}$ there exists $w \in S_i$ with $\rho(v, w) < c \cdot 2^{-i}$. Set S_i is called a 2^{-i}*-net* of the hierarchy. We remark that every point set S possesses, for every $c \geq 1$, at least one hierarchy.

We modify LDM by adding to it yet another constraint. Given set S with 1-covering hierarchy $\mathcal{S}$, the solution must also possess a hierarchy which is a sub-hierarchy of $\mathcal{S}$. (This implies a more structured solution, for which we can construct an IP in Section 5.2.) We show that this additional requirement can be fulfilled without significantly altering the cost and dimension of the optimal solution.

Lemma 5.2. *Let S be a point set, and let $\mathcal{S} = (S_0, \ldots, S_t)$ be a hierarchy for S with a 1-covering property. For every subset $T \subset S$ with doubling dimension $d := \mathrm{ddim}(T)$, there exists a set V satisfying $T \subseteq V \subseteq S$, and an associated hierarchy $\mathcal{V} = (V_0, \ldots, V_t)$ with the following properties.*

1. *Dimension:* $\mathrm{ddim}(V) \leq d' := 4d + 1$.
2. *Covering: Every point $v \in V_i$ is 4-covered by some point in V_{i-1}, and 5-covered by some point of V_k for all $k < i$.*
3. *Heredity: $\mathcal{V}$ is a* sub-hierarchy *of $\mathcal{S}$, meaning that $V_i \subseteq S_i$ for all $i \in [t]$.*

Proof. First extract from the set T an arbitrary 1-covering hierarchy $\mathcal{T} = (T_0, \ldots, T_t)$. Note that each point $v \in T_i$ is necessarily within distance $2 \cdot 2^{-i}$ of some point in S_i; this is because $v \in S_t$, and by the 1-covering property of $\mathcal{S}$, there must be some point $w \in S_i$ within distance $\sum_{j=i}^{t} 2^{-j} < 2 \cdot 2^{-i}$.

Initialize the hierarchy $\mathcal{V}$ by setting $V_0 = S_0$. Construct V_i for $i > 0$ by first including in V_i all points of V_{i-1}. Then, for each $v \in T_i$ (in an arbitrary order), if v is not within distance $2 \cdot 2^{-i}$ of a point already included in V_i, add to V_i the point $v' \in S_i$ closest to v. (Recall from above that $\rho(v, v') < 2 \cdot 2^{-i}$.) Let the final set V_t also include all points of T_t—that is all of T—and $V = V_t$.

Observe that $\mathcal{V}$ is a sub-hierarchy of $\mathcal{S}$, thus it inherits the packing property of hierarchy $\mathcal{S}$. Further, since $\mathcal{T}$ obeyed a 1-cover property, every point in T_i is within distance 2^{-i+1} of some point in T_{i-1}, and so the scheme above implies that any point in V_i must be within distance $2 \cdot 2^{-i} + 2^{-i+1} + 2 \cdot 2^{-(i-1)} = 4 \cdot 2^{-i+1}$ of some point in V_{i-1}. (Here, we have used the triangle inequality from V_i to T_i, T_i to T_{i-1} and finally T_{i-1} to V_{i-1}.) Likewise, since every point in T_i is, for any $k < i$, within distance $2 \cdot 2^{-k}$ of some point in T_k, we get that any point in V_i must be within distance $2 \cdot 2^{-k} + 2 \cdot 2^{-i} + 2 \cdot 2^{-k} \leq 5 \cdot 2^{-k}$ of some point in V_k.

Turning to the dimension, consider an arbitrary ball of radius 2^{-i} centered at any point $v \in V$. Let B be the set of points of V contained in the ball; we show that B can be covered by a small number of balls of radius 2^{-i-1}.

Define $B' \subseteq B$ to include point of V found only in levels V_k for $k \geq i + 3$. B' can be covered by balls of radius 2^{-i-1} centered at points of T_{i+3}: By construction, any point first appearing in V_{i+3} is within distance $2 \cdot 2^{-i-3} < 2^{-i-1}$ of a point of T_{i+3}. Any point first appearing in level V_k for $k > i + 3$ is within distance $2 \cdot 2^{-k}$ of a point in T_k, and that point is within distance $2 \cdot 2^{-i-3}$ of its ancestor in T_{i+3}, for a total distance of $2 \cdot 2^{-k} + 2 \cdot 2^{-i-3} < 2^{-i-1}$. It follows that all points of B' can be covered by a set of balls of radius 2^{-i-1} centered at points of T_{i+3} within distance $2^{-i} + 2^{-i-1} < 2^{-i+1}$ of v. By the doubling property of T, these account for at most 2^{4d} balls.

Now consider the remaining points $B \setminus B'$. These points are found in V_k for $k \leq i + 2$. When $k \leq i$, by construction V_k possesses minimum interpoint distance $2c \cdot 2^{-i}$, and so only a single point among all these levels may be found in B. When $k = i + 1$, we charge each point of V_k to its nearest point in T_k, and note that the points of T_k are within distance $2 \cdot 2^{-k} + 2^{-i} = 4 \cdot 2^{-k}$ of v. There can be at most 4^d such points. A similar argument gives fewer than 8^d points when $k = i + 2$. It follows that B can be covered by $2^{4d} + 1 + 4^d + 8^d < 2^{4d+1}$ smaller balls. $\square$

Please cite this article in press as: L.-A. Gottlieb et al., Adaptive metric dimensionality reduction, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.10.040

5.2. An integer program

We now show an integer program[7] that encapsulates a near-optimal solution to the modified LDM problem; we later relax it to a linear program in Section 5.3. Denote the input by $S = \{v_1, \dots, v_n\}$ and the target dimension by $d \geq 1$, and let $\mathcal{S}$ be a hierarchy for S with a 1-covering property. We shall assume, following Section 5.1, that all interpoint distances in S are in the range $[\delta, 1]$, and the hierarchy possesses $t = \lceil \log_2(1/\delta) \rceil$ levels. The optimal IP solution will imply a subset $W \subset S$ equipped with a hierarchy $\mathcal{W} = (W_0, \dots, W_t)$ that is a sub-hierarchy of $\mathcal{S}$; we will show in Lemma 5.3 that $W = W_t$ constructed in this way is indeed a bicriteria approximation to the modified LDM problem, and therefore to the original LDM problem as well.

We introduce a set Z of 0–1 variables for the hierarchy $\mathcal{S}$; each variable $z_j^i \in Z$ corresponds to a point $v_j \in S_i$, so clearly $|Z| \leq nt$. The IP imposes in Constraint (9) that $z_j^i \in \{0, 1\}$, and this variable is intended to be an indicator for whether $v_j \in W_i$. The IP requires in Constraint (10) that $z_j^i \leq z_j^{i+1}$, to enforce that the hierarchy $\mathcal{W}$ is nested, i.e., if a point is present in W_i then it should be present also in W_{i+1}, and in fact also in W_k for all $k > i$. When convenient, we may refer to distance between variables where we mean distance between their corresponding points (i.e., distance from z_j^i actually means distance from v_j).

We shall now define the i-level *neighborhood* of a point v_j to be the points in S_i that are relatively close to v_j, using a parameter $\alpha \geq 1$. Formally, when $v_j \in S_i$, let $N_j^i(\alpha) \subseteq Z$ include all variables z_k^i for which $\rho(v_j, v_k) \leq \alpha \cdot 2^{-i}$, and call v_j the *center* of $N_j^i(\alpha)$. If $v_j \notin S_i$, then let $w \in S_i$ be the nearest neighbor of v_j in S_i (recall that $\rho(v_j, w) < 2 \cdot 2^{-i}$), call w the *center* of $N_j^i(\alpha)$, and define $N_j^i(\alpha) \subseteq Z$ to include all variables z_k^i for which $\rho(w, v_k) \leq \alpha \cdot 2^{-i}$. So in both cases, $N_j^i(\alpha)$ is the set of variables z_k^i within distance $\alpha \cdot 2^{-i}$ from the center. Constraint (11) of the IP bounds the number of points that W includes from the set $N_j^i(\alpha)$ for each $\alpha \in \{7, 24, 75, 588, 612\}$; by the packing property for doubling spaces of dimension $d' := 4d + 1$ (Lemma 5.2) it can be formulated as $\sum_{z \in N_j^i(\alpha)} z \leq \lceil 2^{\log \alpha} \rceil^{d'} \leq (2\alpha)^{d'}$. The IP also imposes a covering property, by requiring in Constraint (12) that $\sum_{z \in N_j^i(7)} z \geq z_j^t$. Because S possesses a 1-covering property, this constraint ensures indirectly that any point in W_{i+1} is 8-covered by some point in W_i (recall $N_j^i(\alpha)$ might be defined via the nearest neighbor in S_i), and 9-covered by some point in W_k for all $k < i$.

We further introduce in Constraint (14) a set C of n cost variables $c_j \geq 0$ to represent the point mapping cost $\rho(v_j, W)$, and these are enforced by Constraints (15)–(16).

The complete integer program is as follows.

$$
\begin{aligned}
\min \quad & \sum_j c_j \\
\text{s.t.} \quad & z_j^i \in \{0,1\} && \forall z_j^i \in Z && (9)\\
& z_j^i \leq z_j^{i+1} && \forall z_j^i \in Z,\ i < t && (10)\\
& \sum_{z \in N_j^i(\alpha)} z \leq (2\alpha)^{d'} && \forall \alpha \in \{7, 24, 75, 588, 612\},\ i \in [0..t],\ v_j \in S && (11)\\
& \sum_{z \in N_j^i(7)} z \geq z_j^t && \forall i \in [0..t],\ \forall z_j^t \in Z && (12)\\
& \sum_{z \in N_j^i(24)} z \geq \frac{1}{(2 \cdot 24)^{d'}} \sum_{z \in N_j^k(24)} z && \forall i, k \in [0..t], i < k,\ \forall v_j \in S && (13)\\
& c_j \geq 0 && \forall v_j \in S && (14)\\
& z_j^t + \tfrac{c_j}{\delta} \geq 1 && \forall v_j \in S && (15)\\
& z_j^t + \tfrac{c_j}{2^{-i}} + \sum_{z \in N_j^i(12)} z \geq 1 && \forall i \in [0..t],\ \forall v_j \in S && (16)
\end{aligned}
$$

Recall that T denotes an optimal solution for the original low-dimensional mapping problem on input (S, d), and let C^* be the cost of mapping S to T. Let V be the set guaranteed by Lemma 5.2 (applied with covering $c = 1$), and since $T \subseteq V$ the cost of mapping S to V is no greater than C^*. The following lemma proves a bi-directional relationship between the IP and LDM, to eventually bound the IP solution W in terms of the LDM solution T. We remark that Constraint (13) is not necessary for the following lemma, but will later play a central role in the proof of Lemma 5.4.

[7] Formally, this is a mixed integer linear program, since only some of the variables are constrained to be integers.

Lemma 5.3. *Let (S, d) be an input for the LDM problem, and let $T \subseteq V$ be solutions as above. Then*

(a) *V implies a feasible solution to the IP of objective value at most C^*.*
(b) *A feasible solution to the IP with objective value C' yields (efficiently) an LDM solution W with $\mathrm{ddim}(W) \le \log 150 \cdot d' < 29d + 8$ and cost of mapping S to W at most $32C'$.*

Proof. To prove part (a), we need to show that assigning the indicator variables in Z and cost variables C according to V yields a feasible solution with the stated objective value. Note that $\mathcal{V}$ is nested, so it satisfies Constraint (10). Further, the doubling dimension of V (as given in Lemma 5.2) implies that all points obey packing Constraint (11). The covering properties of V are in fact tighter than those required by Constraint (12).

Turning to the IP objective value, let us show that setting $c_j = \rho(v_j, V)$ satisfies the cost Constraints (15)–(16). Verifying Constraint (15) is easy; we assume that $c_j < \delta$ (as otherwise the constraint is clearly satisfied), and then necessarily $v_j \in V$ and thus $z_j^t = 1$. To verify Constraint (16) for a given i, we may assume that $c_j < 2^{-i}$ (otherwise the constraint is clearly satisfied). Let v^* be the closest neighbor to v_j in V. The point v^* is 9-covered by some point $v_p \in V_i$ (using Constraint (12), just as we explained about $\mathcal{W}$), and so $\rho(v_j, v_p) \le \rho(v_j, v^*) + \rho(v^*, v_p) \le 2^{-i} + 9 \cdot 2^{-i} = 10 \cdot 2^{-i}$. Now, the distance from v_j to the closest point in S_i is less than $2 \cdot 2^{-i}$, so v_p is within distance $10 \cdot 2^{-i} + 2 \cdot 2^{-i} = 12 \cdot 2^{-i}$ of the center of $N_j^i(12)$. We thus find a variable z_p^i included in $N_j^i(12)$ that has value 1 (recall $v_p \in V_i$).

We next claim that Constraint (13) is actually extraneous for this IP, and follows from Constraint (12): Constraint (13) simply means that if $N_j^k(24)$ contains at least one non-zero variable, then so does $N_j^i(24)$. (Recall that by Constraint (11), $N_j^k(24)$ contains at most $(2 \cdot 24)^{d'}$ non-zero variables.) But if $N_j^k(24)$ contains a non-zero variable, then this variable is necessarily 9-covered by some non-zero variable in level i of hierarchy $\mathcal{V}$ (by Constraint (12)). The covering variable must be within distance $9 \cdot 2^{-i} + 24 \cdot 2^{-k}$ of the center of $N_j^k(24)$, within distance $9 \cdot 2^{-i} + 24 \cdot 2^{-k} + 2 \cdot 2^{-k} \le 22 \cdot 2^{-i}$ of v_j, and within distance $22 \cdot 2^{-i} + 2 \cdot 2^{-i} = 24 \cdot 2^{-i}$ of the center of $N_j^i(24)$. So $N_j^i(24)$ contains the covering non-zero variable.

To prove part (b), given a solution to the IP, let the indicator variables Z determine the points of LDM solution $W \subset S$, as well as its hierarchy. We need to bound the dimension and the mapping cost of W. Consider a ball of radius $2^{-i} < r \le 2^{-i+1}$ centered at $v_j \in W$, and let us show that it can be covered by a bounded number of balls of radius at most $2^{-i-1} < \frac{r}{2}$. Every point covered by v_j's ball is within distance $9 \cdot 2^{-i-5} < 2^{-i-1}$ of some covering point in W_{i+5}: To bound the number of these covering points in W_{i+5}, observe they are all within distance $9 \cdot 2^{-i-5} + r \le 9 \cdot 2^{-i-5} + 2^{-i+1} = 73 \cdot 2^{-i-5}$ of v_j, and within distance $75 \cdot 2^{-i-5}$ of a point $v \in S_{i+5}$ covering v_j, and that by Constraint (11), there are at most $(2 \cdot 75)^{d'}$ net-points of W_{i+5} within distance $75 \cdot 2^{-i-5}$ of v. It follows that v_j's ball can be covered by $(2 \cdot 75)^{d'}$ balls of radius $9 \cdot 2^{-i-5} < 2^{-i-1}$ (and whose centers are in W), and therefore $\mathrm{ddim}(W) \le \log 150 \cdot d'$.

Turning to the mapping cost, we will demonstrate that $\rho(v_j, W) \le 32c_j$. If $\rho(v_j, W) = 0$ the claim holds trivially, so we may assume $2^{-p} \le \rho(v_j, W) < 2^{-(p-1)}$ for some integer p. If $p \ge t - 4$ then using Constraint (15) we have that $c_j \ge \delta \ge 2^{-t} \ge 2^{-(p-1)-5} > \frac{1}{32}\rho(v_j, W)$. Otherwise, we shall use Constraint (16) for $i = p + 5 \le t$. The distance from v_j to any point of $N_j^i(24)$ is at most $2 \cdot 2^{-i} + 24 \cdot 2^{-i} = 26 \cdot 2^{-i} < 2^{-p}$, but since $\rho(v_j, W) \ge 2^{-p}$, no point of $N_j^i(24)$ is contained in W and thus $z_j^t = \sum_{z \in N_j^i(24)} z = 0$. Constraint (16) now implies $c_j \ge 2^{-i} = 2^{-p-5} \ge \frac{1}{32}\rho(v_j, W)$. $\square$

5.3. A linear program

While the IP gives a good approximation to the LDM problem, we do not know how to solve this IP in polynomial time. Instead, we create an LP by relaxing the integrality constraints (9) into linear constraints $z_j^i \in [0, 1]$. This LP can be solved quickly, as shown in Section 5.4. After solving the LP, we recover a solution to the LDM problem by rounding the Z variables to integers, as follows:

1. If $z_j^t \ge \frac{1}{2}$, then z_j^t is rounded up to 1.
2. For each level $i = 0, \ldots, t$: Let $\hat{\mathcal{N}}^i$ be the set of all neighborhoods $N_j^i(24)$. Extract from $\hat{\mathcal{N}}^i$ a maximal subset $\hat{\mathcal{N}}^i$ whose elements obey the following: (i) For each $N_j^i(24) \in \hat{\mathcal{N}}^i$ there is some $k \ge i$ such that $\sum_{z \in N_j^k(24)} z \ge \frac{1}{4}$. (ii) Elements of $\hat{\mathcal{N}}^i$ do not intersect. For each element $N_j^i(24) \in \hat{\mathcal{N}}^i$, we round up to 1 its center z_l^i (where v_l is the nearest neighbor of v_j in S_i), as well as every variable z_l^k with $k > i$.
3. All other variables of Z are rounded down to 0.

These rounded variables Z correspond (in an obvious manner) to an integral solution W' with hierarchy $\mathcal{W}'$. The following lemma completes the first half of Theorem 5.1.

Lemma 5.4. *W' is a $(624, 82 + o(1))$-bicriteria approximate solution to the LDM problem on S.*

Proof. Before analyzing W', we enumerate three properties of its hierarchy $\mathcal{W}'$.

(i) Nested. When a variable of level i is rounded up in rounding step 2, all corresponding variables in levels $k > i$ are also rounded up. This implies that $\mathcal{W}'$ is nested.

(ii) Packing. For later use, we will need to show that after the rounding, the number of 1-valued variables found in each $N_j^i(g)$ is small, when $g = 588$. By Constraint (11), the sum of the pre-rounded variables $z_k^i \in N_j^i(g)$ is at most $(2g)^{d'}$. If $i = t$, then step 1 rounds up only variables z_k^t of value $\frac{1}{2}$ and higher, so after this rounding step G_j^t contains at most $2 \cdot (2g)^{d'}$ points of W_t'. For general $i \in [t]$, variables of $N_j^i(g)$ may be rounded up due to rounding step 2 acting on level i. This step stipulates that a variable $z_l^i \in N_j^i(g)$ may be rounded up if z_l^i is the center of a distinct subset $N_l^i(24) \in \hat{\mathcal{N}}^i$. Inclusion in $\hat{\mathcal{N}}^i$ requires $\sum_{z \in N_l^k(24)} z \geq \frac{1}{4}$ for some $k \geq i$, and so Constraint (13) implies that $\sum_{z \in N_l^i(24)} z \geq \frac{1}{4(2 \cdot 24)^{d'}}$. Now, since z_l^i is in both $N_j^i(g)$ and $N_l^i(24)$, all points in $N_l^i(24)$ are within distance $g + 24$ of the center of G_j^i, and so by Constraint (11) rounding step 2 may place at most $4(2 \cdot 24)^{d'} \cdot (2 \cdot (g+24))^{d'} < (2 \cdot (g+24))^{2d'}$ points of W_i' into the ball.

Further, rounding step 2 acting on levels $k < i$ may add points to ball $N_j^i(g)$. Since points in each nested level k possess packing 2^{-k}, and the radius of our ball is at most $g \cdot 2^i$, levels $k \leq i - \log g$ can together add just a single point. Levels $i - \log g < k < i$ may each add at most $(2 \cdot (g+24))^{d'}$ additional points to $N_j^i(g)$, accounting for $(2 \cdot (g+24))^{d'}$ total points. It follows that the total number of points in the ball is bounded by $2(2 \cdot (g+24))^{2d'}$.

(iii) Covering. We first consider a variable z_j^t rounded up in rounding step 1, and show it will be 74-covered in each level W_i' of the hierarchy. Since $z_j^t \geq \frac{1}{2}$, Constraint (12) implies that for the pre-rounded variables, $\sum_{z \in N_j^i(24)} z \geq \sum_{z \in N_j^i(7)} z \geq \frac{1}{2}$. By construction of rounding step 2, a variable of $N_j^i(24)$ or one in a nearby set in $\hat{\mathcal{N}}^i$ is rounded to 1, and the distance of this variable from v_j is less than $(3 \cdot 24 + 2) \cdot 2^{-i} = 74 \cdot 2^{-i}$.

We turn to a variable z_j^i rounded to 1 in step 2, and demonstrate that it too is 74-covered in each hierarchy level $k < i$. Since z_j^i was chosen to be rounded, there must exists $k \geq i$ with $\sum_{z \in N_j^k(24)} z \geq \frac{1}{4}$, and so a variable in every set $N_j^h(24)$ (or in a nearby set in $\hat{\mathcal{N}}^h$) for all $h < k$ must be rounded as well. It follows that z_j^i is $3 \cdot 24 < 74$-covered by a variable in each set $N_j^h(24)$ (or in a nearby set in $\hat{\mathcal{N}}^h$) for all $h < i$.

Having enumerated the properties of the hierarchy, we can now prove the doubling dimension of W'. First fix a radius $2^{-i} < r \leq 2^{-i+1}$ and a center $v_j \in W_t'$, and we will show that this ball can be covered by a fixed number of balls with radius at most $2^{-i-1} < \frac{r}{2}$: Each point covered by v_j's ball is within distance $74 \cdot 2^{-i-8} < 2^{-i-1}$ of some covering point in W_{i+8}, and so all covering points in W_{i+8} are within distance $74 \cdot 2^{-i-8} + 2^{-i+1} = 74 \cdot 2^{-i-8} + 512 \cdot 2^{-i-8} = 586 \cdot 2^{-i-8}$ of v_j and within distance $588 \cdot 2^{-i-8}$ of a point $v \in S_{i+8}$ covering v_j. Recalling that $g = 588$, by the packing argument above there are at most $2(2 \cdot (g+24))^{2d'}$ net-points of W_{i+8} within distance $g \cdot 2^{-i-8}$ of v, and this implies a doubling dimension of $2d' \log(2 \cdot (g+24)) + 1 = 2d' \log 1224 + 1 < 82d + 1$.

It remains to bound the mapping cost. By Lemma 5.3(a), the cost of an optimal LP solution is at most $32C^*$. Consider the mapping cost of a point v_j. If the corresponding variable z_j^t was rounded up to 1 then the mapping cost $\rho(v_j, W') = 0 \leq c_j$, i.e., at most the contribution of this point to the LP objective. Hence, we may restrict attention to a variable $z_j^t < \frac{1}{2}$ that was subsequently rounded down. We want to show that $\rho(v_j, W')$ is not much more than the LP cost c_j. First, $c_j \geq \frac{\delta}{2}$ by Constraint (15). Now take the highest level i for which $c_j < \frac{2^{-i}}{4}$; by Constraint (16), it must be that $\sum_{z \in N_j^i(24)} \geq \frac{1}{4}$. Then by rounding step 2, a variable within distance $72 + 2 \cdot 2^{-i} = 74 \cdot 2^{-i}$ of v_j must be rounded up. Hence, the LP cost $c_j \geq \frac{2^{-i-1}}{4} = \frac{2^{-i}}{8}$ is at least $1/592$-fraction of the mapping cost $\rho(v_j, W')$. Altogether, we achieve an approximation of $32 + 592 = 624$ to the optimal cost. □

5.4. LP solver

To solve the linear program, we utilize the framework presented by Young [52] for LPs of following form: Given non-negative matrices P, C, vectors p, c and precision $\beta > 0$, find a non-negative vector x such that $Px \leq p$ (LP packing constraint) and $Cx \geq c$ (LP covering constraint). Young shows that if there exists a feasible solution to the input instance, then a solution to a relaxation of the input program, specifically $Px \leq (1+\beta)p$ and $Cx \geq c$, can be found in time $O(mr(\log m)/\beta^2)$, where m is the number of constraints in the program and r is the maximum number of constraints in which a single variable may appear. We will show how to model our LP in a way consistent with Young's framework, and obtain an algorithm that achieves the approximation bounds of Lemma 5.4 with the runtime claimed by Theorem 5.1. Lemma 5.5 below completes the proof of Theorem 5.1.

Lemma 5.5. *An algorithm realizing the bounds of Lemma 5.4 can be computed in time* $2^{O(\mathrm{ddim}(S))}n + O(n \log^4 n)$.

Proof. To define the LP, we must first create a hierarchy for S, which can be done in time $\min\{O(tn^2), 2^{O(\mathrm{ddim}(S))}tn\}$, as in [10,53]. After solving the LP, the rounding can be done in this time bound as well.

To solve the LP, we first must modify the constraints to be of the form $Px \leq p$ and $Cx \geq c$. This can be done easily by introducing complementary constraints $\bar{z}^i_j \in [0,1]$, and setting $z^i_j + \bar{z}^i_j = 1$. For example, constraint $z^i_j \geq z^t_j$ now becomes $z^i_j + \bar{z}^t_j \geq 1$. A similar approach works for the other constraints as well.

We now count the number of basic constraints. Note that $j \in [1,n]$ and $i \in [1,t]$, so a simple count gives $m = O(t^2 n)$ constraints (where the quadratic term comes from constraint (13)). To bound r, the maximum number of constraints in which a single variable may appear, we note that this can always be bounded by $O(1)$ if we just make copies of variable z^i_j. (That is, two copies of the form $z^{i\,\prime}_j = z^i_j$, $z^{i\,\prime\prime}_j = z^i_j$, then two copies of each copy, etc.) So $r = O(1)$ and the bound on m increases to $O(t^2 n + n\log n)$.

Finally, we must choose a value for β. The variable copying procedure above creates a dependency chain of $O(\log n)$ variables, which will yield additive errors unless $\beta = O(1/\log n)$. Similarly, constraint (10) creates a chain of $O(t)$ variables, so $\beta = O(1/t)$. It suffices to take $\beta = O(1/(t\log n))$, and the stated runtime follows. □

6. Conclusion

We developed learning algorithms that adapt to the intrinsic dimensionality of the data. Our algorithms exploit training sets that are close to being low-dimensional to achieve improved runtime and more optimistic generalization bounds. For linear classifiers in Euclidean spaces, we showed data-dependent generalization bounds that can be optimized by PCA, which is probably the most widely used dimension-reduction technique. For Lipschitz classifiers in general metric spaces, we demonstrated similar data-dependent generalization bounds, which suggest a metric analogue of PCA. We then designed an algorithm that computes the approximate dimensionality of metric data.

An outstanding question left open is to understand the scenarios in which reducing the dimension (by PCA or other techniques) before running a learning algorithm would achieve *better accuracy* (as opposed to better runtime and generalization bounds). It would also be interesting to investigate further our notion of metric dimension reduction (analogous to PCA), e.g., by devising for it better or more practical algorithms, or by finding other contexts where it could be useful.

Acknowledgements

We thank the anonymous referees for constructive suggestions and Ramon van Handel for helpful correspondence.

References

[1] P. Enflo, On the nonexistence of uniform homeomorphisms between L_p-spaces, Ark. Mat. 8 (1969) 103–105.
[2] A. Naor, G. Schechtman, Planar earthmover is not in l_1, SIAM J. Comput. 37 (2007) 804–826, http://dx.doi.org/10.1137/05064206X.
[3] A. Andoni, R. Krauthgamer, The computational hardness of estimating edit distance, SIAM J. Comput. 39 (6) (2010) 2398–2429, http://dx.doi.org/10.1137/080716530.
[4] C.A. Micchelli, M. Pontil, A function representation for learning in Banach spaces, in: COLT, 2004, pp. 255–269.
[5] U. von Luxburg, O. Bousquet, Distance-based classification with Lipschitz functions, J. Mach. Learn. Res. 5 (2004) 669–695.
[6] M. Hein, O. Bousquet, B. Schölkopf, Maximal margin classification for metric spaces, J. Comput. System Sci. 71 (3) (2005) 333–359.
[7] R. Der, D. Lee, Large-margin classification in Banach spaces, in: AISTATS, 2007, pp. 91–98.
[8] H. Zhang, Y. Xu, J. Zhang, Reproducing kernel Banach spaces for machine learning, J. Mach. Learn. Res. 10 (2009) 2741–2775, http://dl.acm.org/citation.cfm?id=1577069.1755878.
[9] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.
[10] R. Krauthgamer, J.R. Lee, Navigating nets: simple algorithms for proximity search, in: SODA, 2004, pp. 791–801.
[11] R.K. Lee-Ad Gottlieb, Aryeh Kontorovich, Efficient classification for metric data, IEEE Trans. Inform. Theory 60 (9) (2014) 5750–5759.
[12] L.-A. Gottlieb, A. Kontorovich, Nearly optimal classification for semimetrics, arXiv:1502.06208, http://arxiv.org/abs/1502.06208, 2015.
[13] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[14] V. Koltchinskii, D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, Ann. Statist. 30 (1) (2002) 1–50, http://dx.doi.org/10.1214/aos/1015362182.
[15] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Inf. Sci. Stat., Springer, 2007.
[16] C.J.C. Burges, Dimension reduction: a guided tour, Found. Trends Mach. Learn. 2 (4) (2010) 275–365.
[17] M.-F. Balcan, A. Blum, S. Vempala, Kernels as features: on kernels, margins, and low-dimensional mappings, Mach. Learn. 65 (1) (2006) 79–94.
[18] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: NIPS, 2007, pp. 1177–1184.
[19] Q. Shi, C. Shen, R. Hill, A. van den Hengel, Is margin preserved after random projection?, in: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 2012.
[20] S. Paul, C. Boutsidis, M. Magdon-Ismail, P. Drineas, Random projections for support vector machines, in: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, 2013, pp. 498–506.
[21] J. Bi, K.P. Bennett, M.J. Embrechts, C.M. Breneman, M. Song, Dimensionality reduction via sparse support vector machines, J. Mach. Learn. Res. 3 (2003) 1229–1243.
[22] K. Fukumizu, F.R. Bach, M.I. Jordan, Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, J. Mach. Learn. Res. 5 (2004) 73–99, http://dl.acm.org/citation.cfm?id=1005332.1005335.
[23] K. Huang, S. Aviyente, Large margin dimension reduction for sparse image classification, in: SSP, 2007, pp. 773–777.
[24] K.R. Varshney, A.S. Willsky, Linear dimensionality reduction for margin-based classification: high-dimensional data and sensor networks, IEEE Trans. Signal Process. 59 (6) (2011) 2496–2512.
[25] I. Borg, P.J. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, 2005.

[26] L.-A. Gottlieb, R. Krauthgamer, Proximity algorithms for nearly-doubling spaces, in: APPROX-RANDOM, 2010, pp. 192–204.
[27] B. Schölkopf, J. Shawe-Taylor, A. Smola, R. Williamson, Kernel-dependent support vector error bounds, in: ICANN, 1999, pp. 103–108.
[28] R.C. Williamson, A.J. Smola, B. Schölkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, IEEE Trans. Inform. Theory 47 (6) (2001) 2516–2532, http://dx.doi.org/10.1109/18.945262.
[29] S. Mendelson, On the performance of kernel classes, J. Mach. Learn. Res. 4 (2003) 759–771, http://www.jmlr.org/papers/v4/mendelson03a.html.
[30] G. Blanchard, O. Bousquet, P. Massart, Statistical performance of support vector machines, Ann. Statist. 36 (2) (2008) 489–531, http://dx.doi.org/10.1214/009053607000000839.
[31] D. Hsu, S.M. Kakade, T. Zhang, Random design analysis of ridge regression, Found. Comput. Math. 14 (3) (2014) 569–600, http://dx.doi.org/10.1007/s10208-014-9192-1.
[32] S. Sabato, N. Srebro, N. Tishby, Tight sample complexity of large-margin learning, in: NIPS, 2010, pp. 2038–2046.
[33] G. Blanchard, L. Zwald, Finite-dimensional projection for classification and statistical learning, IEEE Trans. Inform. Theory 54 (9) (2008) 4169–4182, http://dx.doi.org/10.1109/TIT.2008.926312.
[34] S. Mosci, L. Rosasco, A. Verri, Dimensionality reduction and generalization, in: Machine Learning, Proceedings of the Twenty-Fourth International Conference, ICML 2007, 2007, pp. 657–664.
[35] P.J. Bickel, B. Li, Local polynomial regression on unknown manifolds, in: R. Liu, W. Strawderman, C.-H. Zhang (Eds.), Complex Datasets and Inverse Problems, in: Lecture Notes–Monograph Series, vol. 54, Institute of Mathematical Statistics, 2007, pp. 177–186.
[36] S. Kpotufe, Fast, smooth and adaptive regression in metric spaces, in: Advances in Neural Information Processing Systems, vol. 22, 2009, pp. 1024–1032.
[37] S. Kpotufe, k-NN regression adapts to local intrinsic dimension, in: Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 729–737.
[38] S. Kpotufe, S. Dasgupta, A tree-based regressor that adapts to intrinsic dimension, J. Comput. System Sci. 78 (5) (2012) 1496–1515, http://dx.doi.org/10.1016/j.jcss.2012.01.002.
[39] S. Kpotufe, V.K. Garg, Adaptivity to local smoothness and dimension in kernel regression, in: Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 3075–3083.
[40] P.N. Lee-Ad Gottlieb, Aryeh Kontorovich, Near-optimal sample compression for nearest neighbors, in: Neural Information Processing Systems, NIPS, 2014, pp. 370–378.
[41] A. Kontorovich, R. Weiss, A Bayes consistent 1-NN classifier, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, 2015, pp. 480–488.
[42] A. Gupta, R. Krauthgamer, J.R. Lee, Bounded geometries, fractals, and low-distortion embeddings, in: FOCS, 2003, pp. 534–543.
[43] A.T. Mehryar Mohri, Afshin Rostamizadeh, Foundations of Machine Learning, The MIT Press, 2012.
[44] L. Wasserman, All of Nonparametric Statistics, Springer Texts in Statistics, Springer, New York, 2006.
[45] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (5) (1998) 1926–1940.
[46] R.M. Dudley, Uniform Central Limit Theorems, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 1999.
[47] S. Kakade, A. Tewari, Dudley's theorem, fat shattering dimension, packing numbers. Lecture 15, Toyota Technological Institute at Chicago, 2008, available at http://ttic.uchicago.edu/~tewari/lectures/lecture15.pdf.
[48] A. Blum, Random projection, margins, kernels, and feature-selection, in: Subspace, Latent Structure and Feature Selection, 2005, pp. 52–68.
[49] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Invent. Math. 152 (1) (2003) 37–55, http://dx.doi.org/10.1007/s00222-002-0266-3.
[50] M. Ledoux, M. Talagrand, Probability in Banach Spaces, Springer-Verlag, 1991.
[51] A.N. Kolmogorov, V.M. Tihomirov, ε-Entropy and ε-capacity of sets in functional space, Amer. Math. Soc. Transl. Ser. 2 17 (1961) 277–364.
[52] N.E. Young, Sequential and parallel algorithms for mixed packing and covering, in: FOCS, 2001, pp. 538–546.
[53] R. Cole, L.-A. Gottlieb, Searching dynamic point sets in spaces with bounded doubling dimension, in: 38th Annual ACM Symposium on Theory of Computing, 2006, pp. 574–583.