Adaptive Metric Dimensionality Reduction[*]

Lee-Ad Gottlieb[1], Aryeh Kontorovich[2,**], and Robert Krauthgamer[3,***]

[1] Ariel University, Ariel, Israel
[2] Ben-Gurion University of the Negev, Beer Sheva, Israel
[3] Weizmann Institute of Science, Rehovot, Israel

Abstract. We study data-adaptive 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, which yields a new theoretical explanation for empirically reported improvements gained by preprocessing Euclidean data by PCA (Principal Components Analysis) prior to constructing a linear classifier. 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 [9, 25, 1]. 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. [23, 22, 16, 8, 33].

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, and algorithmically, i.e. computational runtime. 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 [22, 13].

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 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.

Previous work has mainly addressed statistical efficiency in Hilbertian spaces. Scholkopf, Shawe-Taylor, Smola, and Williamson [29] noted the folklore fact that the intrinsic dimensionality of data affects the generalization performance of SVM on that data, and they provided a rigorous explanation for this phenomenon by deriving generalization bounds expressed in terms of the singular values of the training set. These results are a first step towards establishing a connection between Principal Components Analysis (PCA) and linear classification (in fact SVM). However, their generalization bounds are somewhat involved, and hold only for the case of zero training-error. Moreover, these results do not lead to any computational speedup, as the algorithm employed is SVM, without (say) a PCA-based dimensionality reduction. Other attempts to analyze the statistical benefits of low intrinsic dimensionality include [5, 19], however this phenomenon still lacks a simple explanation with user-friendly bounds.

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 [3, 18], but we are not aware of rigorous analysis that extends such bounds to the case where the sample is "close" to a low-dimensional subspace.

Two geometric notions put forth by Sabato, Srebro and Tishby [28] for the purpose of providing tight bounds on the sample complexity, effactually represent "low intrinsic dimensionality". However, these results are statistical in nature, and do not address at all the issue of computational efficiency. Our notion of low dimension may seem similar to theirs, but it is in fact quite different — our definition depends only on the (observed) training sample, while theirs depend on the data's entire (unknown) distribution.

Our contribution. We present classification algorithms that adapt to the *intrinsic* dimensionality of the data, and can exploit a training set that is *close* to being low-dimensional for improved accuracy and runtime complexity. We start with the scenario of a Hilbertian space, which is technically simpler. Let the observed sample be $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^N \times \{-1, 1\}$, and suppose that $\{x_1, \ldots, x_n\}$ 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 tradeoff between the reduced dimension and the distortion, which can be optimized efficiently by performing PCA. To the best of our knowledge, our analysis provides the first rigorous theory for selecting a cutoff value for the singular values, in any supervised learning setting. Algorithmically, our approach amounts to running PCA with a cutoff value implied by Corollary 1, constructing a linear classifier on the projected data $(P_T(x_1), y_1), \ldots, (P_T(x_n), y_n)$, and "lifting" this linear classifier to $\mathbb{R}^N$, with the low dimensionality of T being exploited to speed up the classifier's construction.

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 $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$, where $(\mathcal{X}, \rho)$ is some metric space,. The statistical framework proposed by [22], where classifiers are realized by Lipschitz functions, was extended by [13] 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. First, we establish in Section 4 new generalization bounds for the scenario where there is a multiset $\tilde{S} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$ of low doubling dimension, whose distortion $\eta = \sum_i \rho(x_i, \tilde{x}_i)$ is small. In this case, the Lipschitz extension classifier will generalize well, regardless of the ambient dimension $\mathrm{ddim}(\mathcal{X})$; see Theorem 3. Next, 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 4. The overall classification algorithm operates by computing $\tilde{S}$ and constructing a Lipschitz classifier on the modified training set $(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)$, exploiting its low doubling dimension to compute a classifier faster, using for example [13].

An important feature of our method is that the generalization bounds depend only on the *intrinsic* dimension of the *training set*, and not on the dimension of (or potential points in) the ambient space. Similarly, the intrinsic low dimensionality of the observed data is exploited to design faster algorithms.

Related work. There is a plethora of literature on dimensionality reduction, see e.g. [21, 7], and thus we restrict the ensuing discussion to results addressing supervised learning. Previously, only Euclidean dimension reduction was considered, and chiefly for the purpose of improving runtime efficiency This was realized by projecting the data onto a random low-dimensional subspace — a data-oblivious technique, see e.g. [2, 27, 26]. On the other hand, data-dependent

dimensionality reduction techniques have been observed empirically to improve or 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. [4, 10, 17, 31]. Remarkably, these techniques in some sense defy standard margin theory because orthogonal projection is liable to decrease the separation margin. Our analysis in Section 3 sheds new light on the matter.

There is little previous work on dimension reduction in general metric spaces. MDS (Multi-Dimensional Scaling) is a generalization of PCA, whose input is metric (the pairwise distances); however, its output is Euclidean and thus MDS is effective only for metrics that are "nearly" Euclidean. [14] considered another metric dimension reduction problem: *removing* from an input set S as few points as possible, so as to obtain a large subset of low doubling dimension. While close in spirit, their objective is technically different from ours, and the problem seem to require rather different techniques.

2   Definitions and Notation

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 $(\mathcal{X}, \rho)$. The Lipschitz constant of a function $f : \mathcal{X} \to \mathbb{R}$, denoted by $\|f\|_{\mathrm{Lip}}$, is defined to be the infimum $L \geq 0$ that satisfies $|f(x) - f(y)| \leq L \cdot \rho(x, y)$ for all $x, y \in \mathcal{X}$.

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. $\lambda_{\mathcal{X}}$ 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. [15]. We will use $|\cdot|$ to denote the cardinality of finite metric spaces.

Rademacher complexity. For any n points $Z_1, \ldots, Z_n$ in some set $\mathcal{Z}$ and any collection of functions G mapping $\mathcal{Z}$ to a bounded range, we may define the Rademacher complexity of G evaluated at the n points: $\hat{R}_n(G; \{Z_i\}) = \frac{1}{n} \cdot$

$\mathbb{E} \sup_{g \in G} \sum_{i=1}^{n} \sigma_i g(Z_i)$, where the expectation is over the iid random variables σ_i that take on ± 1 with probability $1/2$. The seminal work of [3] and [18] established the central role of Rademacher complexities in generalization bounds.

3   Adaptive Dimensionality Reduction: Euclidean Case

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), $i = 1, \ldots, n$, with $(X_i, Y_i) \in \mathbb{R}^N \times \{-1, 1\}$, and without loss of generality we take $\|X_i\|_2 \leq 1$ and the hypothesis class $H = \{x \mapsto \mathrm{sgn}(w \cdot x) : \|w\|_2 \leq 1\}$. 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 $\mathbb{R}^N$, Massart's lemma [24, Theorem 3.3] implies that $\hat{R}_n(H) \leq \sqrt{2k \log(en/k)/n}$, 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 [6] is that data separable by margin γ effectively lies in an $\tilde{O}(1/\gamma^2)$-dimensional space.

In this section, we consider the case where 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 $\frac{1}{n} \sum_{i=1}^n \|P_T(X_i) - X_i\|_2^2 \leq \eta$ (where $P_T(\cdot)$ 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 1). As a consequence, we obtain a bound on the expected hinge-loss (Corollary 1). These results both motivate and guide the use of PCA for classification.

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

We prove this theorem in the full version [12]. Notice that the Rademacher complexity is independent of the ambient dimension N. Also note the tension between $\dim(T)$ and η in the bound above — as we seek a lower-dimensional approximation, we are liable to incur a larger distortion.

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

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

where $L(u, y) = |u| \, \mathbb{1}_{\{yu<0\}}$ is the hinge loss.

Proof. Follows from the Rademacher generalization bound [24, Theorem 3.1], the complexity estimate in Theorem 1, and an application of Talagrand's contraction lemma [20] to incorporate the hinge loss.   □

Implicit in Corollary 1 is a tradeoff between dimensionality reduction and distortion. Algorithmically, this tradeoff may be optimized using PCA. It suffices to compute the singular value decomposition once, with runtime complexity $O(n^3 + Nn^2)$ [11]. Then for each $1 \leq k \leq N$, we obtain the lowest-distortion k-dimensional subspace $T^{(k)}$, corresponding to the top k singular values. We then choose the value $1 \leq k \leq N$ which minimizes the generalization bound of Corollary 1 and construct a low-dimensional linear classifier on the projected data $(P_T(x_1), y_1), \ldots, (P_T(x_n), y_n)$, which is "lifted" to $\mathbb{R}^N$. Our generalization

bound holds in the original space $\mathbb{R}^N$ even without the projection and lifting, although heuristically we expect improved perfmance for norm-regularized classifiers since now w is allowed to be large without "wasting" magnitude on irrelevant dimensions.

As PCA is already employed heuristically as a denoising filtering step in the supervised classification setting [4, 17, 31], Corollary 1 provides apparently the first rigorous theory for choosing the best cutoff for the PCA singular values.

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), $i = 1, \ldots, n$, with $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$. Following [22] and [13], 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.

In Section 4.1, we formalize the notion of "nearly" low-dimensional data in a metric space and discuss its implication for Rademacher complexity. Given $S = \{x_i\} \subset \mathcal{X}$, we say that $\tilde{S} = \{\tilde{x}_i\} \subset \mathcal{X}$ is an (η, D)-*perturbation* of S if $\sum_{i=1}^n \rho(x_i, \tilde{x}_i) \le \eta$ and $\mathrm{ddim}(\tilde{S}) \le D$. If our data admits an (η, D)-perturbation, we can prove that the Rademacher complexity it induces on Lipschitz functions can be bounded in terms of η and D alone (Theorem 2), independently of the ambient dimension $\mathrm{ddim}(\mathcal{X})$. As in the Euclidean case (Theorem 1), Rademacher estimates imply data-dependent error bounds, stated in Theorem 3.

In Section 4.3, we describe how to convert our perturbation-based Rademacher bounds 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. As a preprocessing step, we iterate the bicriteria algorithm to find a near-optimal tradeoff between dimensionality and distortion. Having found a near-optimal (η, D)-perturbation $\tilde{S}$, we employ the machinery developed in [13] to 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 [13] in terms of the fat-shattering dimension, but here we obtain data-dependent bounds by direct control over the covering numbers. The proof appears in the full version [12].

Theorem 2. *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 $\tilde{S}$ is an $(\eta n^{D/(D+1)}, D)$-perturbation of S. Let F_L be the collection of all L-Lipschitz, $[-1, 1]$-valued functions on $\mathcal{X}$. Then $\hat{R}_n(F_L; S) = O\left(\frac{L(1+\eta)}{n^{1/(D+1)}}\right)$.*

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, charges nothing for predicting correctly with confidence $yf(x) \geq \gamma$, and for $0 < yf(x) < \gamma$ linearly interpolates between 0 and 1. Since $L_\gamma(f(x), y) \leq \mathbb{1}_{\{yf(x) < \gamma\}}$, the sample margin loss lower-bounds the margin misclassification error.

Theorem 3. *Let F_L be the collection of L-Lipschitz functions mapping the metric space $\mathcal{X}$ of diameter 1 to $[-1, 1]$. If the iid sample $(X_i, Y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \ldots, n$, admits an $(\eta n^{D/(D+1)}, D)$-perturbation then for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in F_L$ and all $\gamma \in (0, 1)$:*

$$\mathbb{P}(\mathrm{sgn}(f(X)) \neq Y) \leq \tfrac{1}{n} \sum_i L_\gamma(f(X_i), Y_i) + O\left(\tfrac{L(1+\eta)}{\gamma n^{1/(D+1)}} + \sqrt{\tfrac{\lg\lg(1/\gamma)}{n}} + \sqrt{\tfrac{\lg(1/\delta)}{n}}\right).$$

Proof. We invoke [24, Theorem 4.5] to bound the classification error in terms of sample margin loss and Rademacher complexity and the latter is bounded via Theorem 2.   □

4.3   Classification Procedure

Theorem 3 provides a statistical optimality criterion for the dimensionality-distortion tradeoff.[1] Unlike the Euclidean case, where a simple PCA optimized this tradeoff, the metric case requires a novel bicriteria approximation algorithm, described 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 $\mathrm{ddim}(\tilde{S}) \approx D$, which approximately minimizes the distortion η. We may iterate this algorithm over all $D \in \{1, \ldots, \log_2 |S|\}$ — since the doubling dimension of the metric space (S, ρ) is at most $\log_2 |S|$ — to optimize the complexity[2] term in Theorem 3.

Once a nearly optimal $(\eta n^{D/(D+1)}, D)$-perturbation $\tilde{S}$ has been computed, we predict the value at a test point $x \in \mathcal{X}$ by a thresholded Lipschitz extension from $\tilde{S}$, which algorithmically amounts to an approximate nearest-neighbor classifier. The efficient implementation of this method (as well as technicalities stemming from its approximate nature) are discussed in [13]. Their algorithm computes an ε-approximate Lipschitz extension in preprocessing time $2^{O(D)} n \log n$ and test-point evaluation time $2^{O(D)} \log n + \varepsilon^{-O(D)}$. The latter also allows one to efficiently decide on which sample points (if any) the classifier should be allowed to err, with corresponding savings in the Lipschitz constant[3] (and hence lower complexity).

---

[1] Although the estimate in Theorem 2 was given as $O(L(1+\eta)/\gamma n^{1/(D+1)})$ for readability, its proof yields explicit, easily computable bounds.

[2] Since L/γ multiplies $(1+\eta)/n^{1/(D+1)}$ in the error bound, the optimization may be carried out oblivious to L and γ.

[3] The complexity term in Theorem 3 scales as L/γ, hence the final classifier can be normalized to have Lipschitz constant 1 — so no further stratification over L is necessary. We do, however, need to stratify over the doubling dimension D (see [30]).

5   Approximating Intrinsic Dimension and Perturbation

In this section we consider the computation of an (η, D)-perturbation (of the observed data) as an optimization problem, and design for it a polynomial-time bicriteria approximation algorithm. As before, 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.[4] 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 4. *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 $n = |S|$.*

In presenting the algorithm, we first give in Section 5.2 an integer program (IP) that models this problem. We show that an optimal solution to the LDM problem implies a solution to the IP, and also that an optimal solution to the integer program gives a bicriteria approximation to the LDM problem (Lemma 2). However, finding an optimal solution to the IP seems difficult; we thus relax in Section 5.3 some of the IP constraints, and derive a linear program (LP) that can be solved in the runtime stated above (Lemma 4). Further, we give a rounding scheme that recovers from the LP solution an integral solution, and then show in Lemma 3 that the integral solution indeed provides an $(O(1), O(1))$-bicriteria approximation, thereby completing the proof of Theorem 4.[5]

5.1   Preliminaries

Point hierarchies. Let S be a point set, and assume by scaling 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. Every point set S possesses one or more hierarchies for each value of $c \geq 1$. We will later need the following lemma, which extracts from an optimal solution a more structured sub-solution.

---

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

[5] We remark that 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.

Lemma 1. *Let S be a point set, and let $\mathcal{S}$ be a hierarchy for S with a c-covering property. For every subset $T \subset S$ with doubling dimension $D := \mathrm{ddim}(T)$, there exists a set T' satisfying $T \subseteq T' \subseteq S$, and an associated hierarchy $\mathcal{T}'$ with the following properties:*

1. $\mathrm{ddim}(T') \leq D' := \log_2(2^{3D} + 1) = 3D + o(1)$.
2. *Every point $v \in T_i'$ is $4c$-covered by some point in T_{i-1}', and $5c$-covered by some point of T_k' for all $k < i$.*
3. $\mathcal{T}'$ *is a* sub-hierarchy *of $\mathcal{S}$, meaning that $T_i' \subseteq S_i$ for all $i \in [t]$.*

Proof. First take set T and extract from it an arbitrary c-covering hierarchy $\mathcal{T}$ composed of nets T_i. Note that each point $v \in T_i$ is necessarily within distance $2c \cdot 2^{-i}$ of some point in S_i: This is because v exists in S_t, and by the c-covering property of $\mathcal{S}$, $v \in S_t$ must within distance $\sum_{j=t}^{i} c \cdot 2^{-j} = 2c \cdot 2^{-i} - c \cdot 2^{-t} < 2c \cdot 2^{-i}$ of some point $w \in S_i$.

We initialize the hierarchy $\mathcal{T}'$ by setting $T_0' = S_0$. Construct T_i' for $i > 0$ by first including in T_i' all points of T_{i-1}'. Then, for each $v \in T_i$, if v is not within distance $2c \cdot 2^{-i}$ of a point already included in T_i', then add to T_i' the point $v' \in S_i$ closest to v. (Recall from above that $\rho(v, v') < 2c \cdot 2^{-i}$.) Clearly, $\mathcal{T}'$ inherits the packing property of hierarchy $\mathcal{S}$. Further, since $\mathcal{T}$ obeyed a c-cover property, the scheme above ensures that any point $w \in S_i'$ must be within distance $2c \cdot 2^{-i} + c \cdot 2^{-i+1} + 2c \cdot 2^{-i+1} \leq 4c \cdot 2^{-i+1}$ of some point in T_{i-1}', and within distance $2c \cdot 2^{-i} + 2c \cdot 2^{-k} + 2c \cdot 2^{-k} \leq 5c \cdot 2^{-k}$ of some point in any T_k', $k < i$.

Turning to the dimension, $\mathcal{T}$ possessed dimension D, and $\mathcal{T}'$ may be viewed as 'moving' each net point a distance strictly less than $2c \cdot 2^{-i}$, which can increase the dimension by a multiplicative factor of 3. Further, the retention of points of each T_{i-1}' in T_i' can add 1 to the doubling constant, as an added point may be the center of a new ball of radius 2^i. $\qquad\square$

5.2   An Integer Program

The integer program below encapsulates a near-optimal solution to LDM, and will be relaxed to a linear program in Section 5.3. Denote the input by $S = \{v_1, \ldots, v_n\}$ and $D \geq 1$, 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 S are in the range $[\delta, 1]$, and the hierarchy possesses $t = \lceil \log_2(1/\delta) \rceil$ levels. We construct from an optimal IP solution a subset $S' \subset S$ equipped with a hierarchy $\mathcal{S}'$ that is a sub-hierarchy of $\mathcal{S}$; we will show in Lemma 2 that S' constructed in this way is indeed a bicriteria approximation to the LDM problem.

We introduce a set Z of 0-1 variables for the hierarchy $\mathcal{S}$; variable $z_j^i \in Z$ corresponds to a point $v_j \in S_i$. Clearly $|Z| \leq nt$. The IP imposes in Constraint (1) that $z_j^i \in \{0, 1\}$, intended to be an indicator variable for whether v_j appears in S_i' (level i of the hierarchy $\mathcal{S}'$ of S'). The IP requires in Constraint (2) that $z_j^i \leq z_j^{i+1}$, which enforces the nested property in the hierarchy $\mathcal{S}'$. When convenient, we may refer to distance between variables where we mean distance between their corresponding points.

Let us define the i-level *neighborhood* of a point v_j to be the net-points of S_i that are relatively close to v_j. Formally, when $v_j \in S_i$, let $E_j^i \subseteq Z$ include all variables z_k^i for which $\rho(v_j, v_k) \leq e \cdot 2^{-i}$, for $e := 7$. If $v_j \notin S_i$, then let $w \in S_i$ be the nearest neighbor of v_j in S_i (notice that $\rho(v_j, w) < 2 \cdot 2^{-i}$), and define $E_j^i \subseteq Z$ to include all variables z_k^i for which $\rho(w, v_k) \leq e \cdot 2^{-i}$. We similarly define three more neighbor sets: $F_j^i \subseteq Z$ for $f := 12$, $G_j^i \subseteq Z$ for $g := 114$, and $H_j^i \subseteq Z$ for $h := f + g = 126$. The IP imposes on F_j^i, G_j^i, H_j^i (or the corresponding points in S') the packing property for doubling spaces of dimension $D' := \log_2(2^{3D} + 1)$ of the form $\sum_{z \in F_j^i} z \leq (2f)^{D'}$, see Constraints (4)-(6). The IP imposes also covering property, as follows. Constraint (3) requires that $\sum_{z \in E_j^i} z \geq z_j^t$, which implies that every $v_j \in S_i'$ is $(e + 2)$-covered by some point in S_k' for all $k < i$.

We further introduce a set C of n cost variables c_j, intended to represent the point mapping cost $\rho(v_j, T')$, and this is enforced by Constraints (7)-(8). The complete integer program is as follows.

$$
\begin{aligned}
\min \quad & \sum_j c_j \\
\text{s.t.} \quad & z_j^i \in \{0, 1\} && \forall z_j^i \in Z && (1) \\
& z_j^i \leq z_j^{i+1} && \forall z_j^i \in Z && (2) \\
& \sum_{z \in E_j^i} z \geq z_j^t && \forall i,\ z_j^t \in Z && (3) \\
& \sum_{z \in F_j^i} z \leq (2f)^{D'} && \forall i,\ v_j \in S && (4) \\
& \sum_{z \in G_j^i} z \leq (2g)^{D'} && \forall i,\ v_j \in S && (5) \\
& \sum_{z \in H_j^i} z \leq (2h)^{D'} && \forall i,\ v_j \in S && (6) \\
& z_j^t + \tfrac{c_j}{\delta} \geq 1 && \forall v_j \in S && (7) \\
& z_j^t + \tfrac{c_j}{2^{-i}} + \sum_{z \in F_j^i} z \geq 1 && \forall i,\ v_j \in S && (8) \\
& \sum_{z \in F_j^i} z \geq \tfrac{1}{(2f)^{D'}} \sum_{z \in F_j^k} z && \forall i < k,\ v_j \in S && (9)
\end{aligned}
$$

Recall that T is the optimal solution for the low-dimensional mapping problem on input (S, D), and let C^* be the cost of mapping S to T. Let T' be the set given by Lemma 1, and the cost of mapping S to T' cannot be greater than C^*. The following lemma proves a bi-directional relationship between the IP and LDM, relating IP solution S' to LDM solutions T.[6]

---

[6] Constraints (6) and (9) are not necessary for the purposes of the following lemma, but will later play a central role in the proof of Lemma 3.

Lemma 2. *Let (S, D) be an input for the LDM problem.*

(a). *Then T' yields (in the obvious manner) a feasible solution to the IP of cost at most C^*.*

(b). *A feasible solution to the IP with objective value C yields S' that is a bicriteria approximate solution to LDM, with $\mathrm{ddim}(S') \leq (3 \log 228)D + o(1)$ and cost of mapping S to S' at most $32C$.*

Proof. For part (a), we need to show that assigning the variables in Z and C according to T' yields a feasible solution with the stated mapping cost. Note that $\mathcal{T}'$ is nested, so it satisfies Constraint (2). Further, the doubling dimension of T implies that all points obey packing constraints (4)-(6). The covering properties of $\mathcal{T}'$ are tighter than those required by Constraint (3). Constraints (7)-(8) are valid, because if $z_j^t = 0$, then necessarily $\rho(v_j, T')$ must be large enough to satisfy these constraints.

We then claim that Constraint (9) is actually extraneous for this IP, since it is trivially satisfied by any hierarchy possessing $(e + 2)$-covering (Constraint (3)): Since F_j^k contains at most $(2f)^{\log(2^{3d}+1)}$ non-zero variables (Constraint (4)), Constraint (9) simply means that if F_j^k contains at least one non-zero variable, then so does F_j^i. But if F_j^k contains a non-zero variable, then this variable is necessarily $(e + 2)$-covered by some non-zero variable in hierarchical level i. Further, the non-zero covering variable must be in F_j^i, since F_j^i contains all variables within distance $f \cdot 2^{-i} - 2 \cdot 2^{-i} > (e + 2) \cdot 2^{-i}$ of v_j.

Turning to the IP cost, a point v_j included in T' clearly implies that $c_j = 0$. For a point v_j not included in T' ($z_j^t = 0$) Constraint (7) requires that $c_j \geq \delta$, but this is not greater than $\rho(v_j, T')$. If in addition $\sum_{z \in F_j^i} z = 0$ for a minimal i, then by Constraint (8), we must assign cost $c_j = 2^{-i}$. We will show that $c_j \leq \rho(v_j, T')$. In what follows, let v_k be the closest neighbor to v_j in T', and let $2^{-p} \leq \rho(v_j, v_k) < 2^{-(p-1)}$. We demonstrate that when $i < p$, v_j does not incur the cost of constraint (8): v_k is $(e + 2)$-covered by some point $w \in T_i'$, and so $\rho(v_j, w) \leq \rho(v_j, v_k) + \rho(v_k, w) \leq 2^{-p+1} + (e + 2) \cdot 2^{-i} \leq (e + 3) \cdot 2^{-i}$. Now, the distance from v_j to the closest point in S_i is less than $2 \cdot 2^{-i}$, so w is within distance $(e + 3) \cdot 2^{-i} + 2 \cdot 2^{-i} = (e + 5) \cdot 2^{-i} = f \cdot 2^{-i}$ of the center point of F_j^i, and so w's variable is included in F_j^i. It follows that $\sum_{z \in F_j^i} z \geq 1$, and so Constraint (8) does not impose a cost when $i < p$. We conclude that $c_j \leq \rho(v_j, S')$.

We proceed to prove part (b), and show that S' has the stated dimension and cost. Concerning the dimension, recall first that every point $v_j \in S_t'$ is within distance $(e + 2) \cdot 2^{-i}$ of some point in S_i'. Consider a ball of radius $2(e + 2) \cdot 2^{-i}$ centered at any point $v \in S_i'$, and we will show that this large ball can be covered by a fixed number of balls of half-radius $(e + 2) \cdot 2^{-i}$ centered at points of S_i': Each point covered by the large ball is also covered by a half-radius ball centered at some point $w \in S_i'$, and clearly $\rho(v, w) \leq 3(e + 2) \cdot 2^{-i} < g \cdot 2^{-i}$. By Constraint (5), there are at most $(2g)^{3D+o(1))}$ net-points of S_i' within distance

$3(e+2) \cdot 2^{-i}$ of v, and this implies a doubling dimension of $\log(2^{3D}+1)\log 2g = (3\log 228)D + o(1)$.[7]

Turning to mapping cost, we will demonstrate that for set S', $\rho(v_j, S') \le 32c_j$. As above, let v_k be the closest neighbor to v_j in S', and let $2^{-p} \le \rho(v_j, v_k) < 2^{-(p-1)}$. We will show that whenever $i \ge p+5$, v_j incurs the cost of constraint (8) (and recall that by Constraint (7), $c_j \ge \delta$): The distance from v_j to any point of F_k^i is at most $2 \cdot 2^{-i} + f \cdot 2^{-i} \ge= (f+2) \cdot 2^i$. Since the distance from v_j to v_k is as least $2^{-p} = 32 \cdot 2^{-i} > (f+2) \cdot 2^{-i}$, no point of F_k^i is contained in S_i'. It follows that $\sum_{z \in F_j^i} z = 0$, and so c_j must be set equal to at least 2^{-i}. We conclude that $c_j \ge \frac{1}{32}\rho(v_j, S')$. □

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 (1) 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 $\mathcal{F}^i$ be the set of all neighborhoods F_j^i. Extract from $\mathcal{F}^i$ a maximal subset $\hat{\mathcal{F}}^i$ whose elements obey the following: (i) For each $F_j^i \in \hat{\mathcal{F}}^i$ there is some $k \ge i$ such that $\sum_{z \in F_j^k} z \ge \frac{1}{4}$. (ii) Elements of $\hat{\mathcal{F}}^i$ do not intersect. For each element $F_j^i \in \hat{\mathcal{F}}^i$, 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 S'' with hierarchy $\mathcal{S}''$. The following lemma completes the first half of Theorem 4.

Lemma 3. S'' is a $(336, 4\log_2 252 + o(1))$-bicriteria approximate solution to the LDM problem on S.

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

(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 $\mathcal{S}''$ is nested.

(ii) Packing. We will show that after the rounding, the number of 1-valued variables found in each G_j^i is small. By Constraint (5), the sum of the pre-rounded variables $z_k^i \in G_j^i$ 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

---
[7] Choosing $g = 3(e+2)$ would give a tighter bound, but the current value of g will be useful later.

most $2 \cdot (2g)^{D'}$ points of S''_t. For general $i \in [t]$, variables of G^i_j may be rounded up due to rounding step 2 acting on level i. This step stipulates that a variable $z^i_l \in G^i_j$ may be rounded up if z^i_l is the center of a distinct subset $F^i_l \in \hat{\mathcal{F}}^i$. Inclusion in $\hat{\mathcal{F}}^i$ requires $\sum_{z \in F^k_l} z \geq \frac{1}{4}$ for some $k \geq i$, and so Constraint (9) implies that $\sum_{z \in F^i_l} z \geq \frac{1}{4(2f)^{D'}}$. Now, since z^i_l is in both G^i_j and F^i_l, all points in F^i_l are within distance $g + f = h$ of the center of G^i_j, and so by Constraint (6) rounding step 2 may place at most $4(2f)^{\log(2^{3D}+1)} \cdot (2h)^{\log(2^{3D}+1)} = (2h)^{4D+o(1)}$ points of S''_i into the ball.

Further, rounding step 2 acting on levels $k < i$ may add points to ball G^i_j. 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 $(2h)^{4D+o(1)}$ additional points to G^i_j, accounting for $(2h)^{4D+O(1)}$ total points. It follows that the total number of points in the ball is bounded by $(2h)^{4D+O(1)}$.

(iii) Covering. We first consider a variable z^t_j rounded up in rounding step 1, and show it will be $(3f + 2)$-covered in each level S''_i of the hierarchy. Since $z^t_j \geq \frac{1}{2}$, Constraint (3) implies that for the pre-rounded variables, $\sum_{z \in F^i_j} z \geq \sum_{z \in E^i_j} z \geq \frac{1}{2}$. By construction of rounding step 2, a variable of F^i_j or one in a nearby set in $\hat{\mathcal{F}}^i$ is rounded to 1, and the distance of this variable from v_j is less than $(3f + 2) \cdot 2^{-i}$.

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

Having enumerated the properties of the hierarchy, we can now prove the doubling dimension of S''. Take any ball B of radius $2(3f + 2) \cdot 2^{-i}$ centered at a point of S''_i. Since every point of S''_k is $(3f + 2)$-covered by some point in S''_i, the points of S''_k covered by B are all covered by a set of balls of radius $(3f + 2)$ within distance $3(3f + 2) \cdot 2^{-1} = g \cdot 2^{-1}$ of the center point. By the packing property proved above, there exist fewer than $(2h)^{4D+o(1)}$ such points, implying a doubling dimension of $4D \log 252 + o(1)$.

It remains to bound the mapping cost. By Lemma 2(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^t_j was rounded up to 1 then the mapping cost $\rho(v_j, S'') = 0 \leq c_j$, i.e., at most the contribution of this point to the LP objective. Hence, we may restrict attention to a variable $z^t_j < \frac{1}{2}$ that was subsequently rounded down. We want to show that $\rho(v_j, S'')$ is not much more than the LP cost c_j. First, $c_j \geq \frac{\delta}{2}$ by Constraint (7). Now take the highest level i for which $c_j < \frac{2^{-i}}{4}$; by Constraint (8), it must be that $\sum_{z \in F^i_j} \geq \frac{1}{4}$. Then by rounding step 2, a variable within distance $(3f + 2) \cdot 2^{-i} = 38 \cdot 2^{-i}$ 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 $1/304$-fraction of the mapping

cost $\rho(v_j, S'')$. Altogether, we achieve an approximation of $32 + 304 = 336$ to the optimal cost. □

5.4 LP Solver

To solve the linear program, we utilize the framework presented by [32] 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. In [12], we show how to model our LP in a way consistent with Young's framework, with $m = O(t^2 n + n \log n)$, $r = O(1)$, and $\beta = O(1/(t \log n))$. This yields an algorithm that achieves the approximation bounds of Lemma 3 with the runtime claimed by Theorem 4. Lemma 4 below completes the proof of Theorem 4.

Lemma 4. *An algorithm realizing the bounds of Lemma 3 can be computed in time* $2^{O(\mathrm{ddim}(S))}n + O(n \log^4 n)$.

References

[1] Andoni, A., Krauthgamer, R.: The computational hardness of estimating edit distance. SIAM J. Comput. 39(6), 2398–2429 (2010)

[2] Balcan, M.F., Blum, A., Vempala, S.: Kernels as features: On kernels, margins, and low-dimensional mappings. Mach. Learn. 65(1), 79–94 (2006)

[3] Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. JMLR 3, 463–482 (2002)

[4] Bi, J., Bennett, K.P., Embrechts, M.J., Breneman, C.M., Song, M.: Dimensionality reduction via sparse support vector machines. JMLR 3, 1229–1243 (2003)

[5] Blanchard, G., Zwald, L.: Finite-dimensional projection for classification and statistical learning. IEEE Trans. Inform. Theory 54(9), 4169–4182 (2008), http://dx.doi.org/10.1109/TIT.2008.926312

[6] Blum, A.: Random projection, margins, kernels, and feature-selection. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) SLSFS 2005. LNCS, vol. 3940, pp. 52–68. Springer, Heidelberg (2006)

[7] Burges, C.J.C.: Dimension reduction: A guided tour. Foundations and Trends in Machine Learning 2(4) (2010)

[8] Der, R., Lee, D.: Large-Margin Classification in Banach Spaces. In: AISTATS 2007, pp. 91–98 (2007)

[9] Enflo, P.: On the nonexistence of uniform homeomorphisms between L_p-spaces. Ark. Mat. 8, 103–105 (1969)

[10] Fukumizu, K., Bach, F.R., Jordan, M.I.: Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. JMLR 5, 73–99 (2004)

[11] Golub, G.H., Van Loan, C.F.: Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)

[12] Gottlieb, L.A., Kontorovich, A., Krauthgamer, R.: Adaptive metric dimensionality reduction (2013), `http://arxiv.org/abs/1302.2752`

[13] Gottlieb, L.A., Kontorovich, L., Krauthgamer, R.: Efficient classification for metric data. In: COLT, pp. 433–440 (2010)

[14] Gottlieb, L.A., Krauthgamer, R.: Proximity algorithms for nearly-doubling spaces. In: Serna, M., Shaltiel, R., Jansen, K., Rolim, J. (eds.) APPROX and RANDOM 2010. LNCS, vol. 6302, pp. 192–204. Springer, Heidelberg (2010)

[15] Gupta, A., Krauthgamer, R., Lee, J.R.: Bounded geometries, fractals, and low-distortion embeddings. In: FOCS, pp. 534–543 (2003)

[16] Hein, M., Bousquet, O., Schölkopf, B.: Maximal margin classification for metric spaces. J. Comput. Syst. Sci. 71(3), 333–359 (2005)

[17] Huang, K., Aviyente, S.: Large margin dimension reduction for sparse image classification. In: SSP, pp. 773–777 (2007)

[18] Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Ann. Statist. 30(1), 1–50 (2002)

[19] Kpotufe, S., Dasgupta, S.: A tree-based regressor that adapts to intrinsic dimension. J. Comput. Syst. Sci. 78(5), 1496–1515 (2012), `http://dx.doi.org/10.1016/j.jcss.2012.01.002`

[20] Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer (1991)

[21] Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Information Science and Statistics. Springer (2007)

[22] von Luxburg, U., Bousquet, O.: Distance-based classification with lipschitz functions. Journal of Machine Learning Research 5, 669–695 (2004)

[23] Micchelli, C.A., Pontil, M.: A function representation for learning in banach spaces. In: Shawe-Taylor, J., Singer, Y. (eds.) COLT 2004. LNCS (LNAI), vol. 3120, pp. 255–269. Springer, Heidelberg (2004)

[24] Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press (2012)

[25] Naor, A., Schechtman, G.: Planar earthmover is not in l_1. SIAM J. Comput. 37, 804–826 (2007)

[26] Paul, S., Boutsidis, C., Magdon-Ismail, M., Drineas, P.: Random projections for support vector machines. CoRR abs/1211.6085 (2012)

[27] Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: NIPS (2007)

[28] Sabato, S., Srebro, N., Tishby, N.: Tight sample complexity of large-margin learning. In: NIPS, pp. 2038–2046 (2010)

[29] Schölkopf, B., Shawe-Taylor, J., Smola, A., Williamson, R.: Kernel-dependent support vector error bounds. In: ICANN (1999)

[30] Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory 44(5), 1926–1940 (1998)

[31] Varshney, K.R., Willsky, A.S.: Linear dimensionality reduction for margin-based classification: High-dimensional data and sensor networks. IEEE Transactions on Signal Processing 59(6), 2496–2512 (2011)

[32] Young, N.E.: Sequential and parallel algorithms for mixed packing and covering. In: FOCS, pp. 538–546 (2001)

[33] Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel banach spaces for machine learning. J. Mach. Learn. Res. 10, 2741–2775 (2009)