Nearly optimal classification for semimetrics

Lee-Ad Gottlieb
Ariel University

Aryeh Kontorovich
Ben Gurion University

Pinhas Nisnevitch
Tel-Aviv University

Abstract

We initiate the rigorous study of classification in semimetric spaces, which are point sets with a distance function that is non-negative and symmetric, but need not satisfy the triangle inequality. We define the *density dimension* dens and discover that it plays a central role in the statistical and algorithmic feasibility of learning in semimetric spaces. We compute this quantity for several widely used semimetrics and present nearly optimal sample compression algorithms, which are then used to obtain generalization guarantees, including fast rates.

Our claim of near-optimality holds in both computational and statistical senses. When the sample has radius R and margin γ, we show that it can be compressed down to roughly $d = (R/\gamma)^{\mathrm{dens}}$ points, and further that finding a significantly better compression is algorithmically intractable unless P=NP. This compression implies generalization via standard Occam-type arguments, to which we provide a nearly matching lower bound.

1 Introduction

The problem of learning in non-metric spaces has been of significant recent interest, being the subject of a 2010 COLT workshop and a central topic of all three SIMBAD conferences. In this paper, we initiate the study of efficient statistical learning in *semimetric* spaces, which are point sets endowed with a distance function that is non-negative and symmetric but may not satisfy the triangle inequality [40][1]. Without the latter, quite a bit of structure is lost — for example, semimetric spaces admit convergent sequences without a Cauchy subsequence [11]. We are not aware of any rigorous learning results in semimetric spaces prior to this work.

Background and motivation. Much of the existing machinery for classification algorithms, as well as generalization bounds, depends strongly on the data residing in a Hilbert space. For some important applications, this structural constraint severely limits the applicability of existing methods. Indeed, it is often the case that the data is naturally endowed with some distance function strongly dissimilar to the familiar Euclidean norm.

Consider images, for example. Although these can be naively represented as coordinate-vectors in $\mathbb{R}^d$, the Euclidean (or even ℓ_p) distance between the representative vectors does not correspond well to the one perceived by human vision. Instead, the earthmover distance is commonly used in vision applications [33]. Yet representing earthmover distances using any fixed ℓ_p norm unavoidably introduces very large interpoint distortion [32], potentially corrupting the data geometry before the learning process has even begun. Nor is this issue mitigated by kernelization, as kernels necessarily embed the data in a Hilbert space, again incurring the aforementioned distortion. A similar issue arises for strings: These can be naively treated as vectors endowed with different ℓ_p metrics, but a much more natural metric over strings is the edit distance, which is similarly known to be strongly non-Euclidean [1]. Additional limitations of kernel methods are articulated in [4].

These concerns have led researchers to seek out algorithmic and statistical approaches that apply in greater generality. A particularly fruitful recent direction has focused on metric spaces, which are point sets endowed with a distance function that is non-negative

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

[1] Some authors use the term "semimetric" to mean *pseudometrics*. These preserve much of the structure of metrics, the only difference being that they allow distinct points to have distance 0. Our usage appears to be more standard.

and symmetric, and also satisfies the triangle equality. While metric spaces are significantly more general than Hilbertian ones, they still do not capture many common distance functions used by practitioners. These non-metric distances include the Jensen-Shannon divergence, which appears in statistical applications [14, 16], the k-median Hausdorff distances, and the ℓ_p distances with $0 < p < 1$, which appear in vision applications [29, 25] — all of which are semimetrics. An additional line of work [29, 25, 24, 38] underscored the effectiveness of non-metric distances in various applications (mainly vision), and among these, semimetrics again play a prominent role [8, 12, 15, 22, 26, 23].

Main results. We initiate the rigorous study of classification for semimetric spaces. We define the *density dimension* (dens = dens($\mathcal{X}$)) of a semimetric space $\mathcal{X}$ as the logarithm of the *density constant* $\mu = \mu(\mathcal{X})$, which intuitively is the smallest number such that any r-radius open ball in $\mathcal{X}$ contains at most μ points at mutual interpoint distance at least $r/2$; a formal definition is given in Equation (2). We then demonstrate that dens plays a central role in the statistical and algorithmic feasibility of learning in this setting by showing that it controls the packing numbers of $\mathcal{X}$. Crucially for learning, this insight implies that there is one standard technique that survives violations of the triangle inequality — namely, sample compression. Denoting by R and γ the sample radius and margin, respectively, we can achieve the latter by extracting a γ-net from the sample (Theorem 2). This compresses the sample from size n to $(R/\gamma)^{O(\text{dens})}$, which is nearly optimal unless P=NP.

On the statistical front, we give a compression-based generalization bound that smoothly interpolates between the consistent $(R/\gamma)^{O(\text{dens})}/n$ and agnostic $\sqrt{(R/\gamma)^{O(\text{dens})}/n}$ decay regimes (Theorem 7). This "fast rate" holds for general compression schemes. Applied to margin-based semimetric sample-compression schemes, it becomes amenable to efficient Structural Risk Minimization. The lower bound in Theorem 10 shows that even under margin assumptions, there exist adversarial distributions forcing the sample complexity to be exponential in dens.

To demonstrate the applicability of our framework, we compute the density dimension of the three popular semimetrics enumerated above: Jensen-Shannon divergence, ℓ_p distances with $0 < p < 1$, and k-median Hausdorff distances (Theorem 12). Along the way, we discover that the latter (for $k = 1$) is in fact universal for *all* semimetrics; this surprising fact may be of independent interest (Lemma 11).

Related work. In a series of papers, [2, 5, 3, 4] developed a theory of learning with similarity functions, which resemble kernels but relax the requirement of being positive definite. Learning is accomplished by embedding the data into an appropriate Euclidean space and performing large-margin separation. Hence, this approach effectively extracts the implicit Euclidean structure encoded in the similarity function, but does not seem well-suited for inherently non-Euclidean data. In [37] this framework was extended to dissimilarity functions, obtaining analogous results.

For metric spaces, it is known that a sample of size exponential in the doubling dimension (ddim) suffices to achieve low generalization error [36, 17, 34, 27], and that exponential dependence on ddim is in general unavoidable [34]. As for algorithmic runtimes, the naive nearest-neighbor classifier evaluates queries in $O(n)$ time (where n is the sample size); however, an approximate nearest neighbor can be found in time $2^{O(\text{ddim})} \log n$. If one desires runtimes depending not on n but on the geometry (say, margin γ) of the data, one may achieve a sample compression scheme of size $\gamma^{-O(\text{ddim})}$, and it is NP-hard to achieve a significantly better compression [18].

As we show in the Appendix, the above results characterizing learning in metric space do not carry over to semimetrics. More precisely, the doubling dimension of a semimetric does not control its packing numbers, as it does in metric spaces. Although we succeed in showing that the density constant does indeed control the packing numbers even in semimetrics, this does not necessarily imply portability of learning algorithms for metric spaces into semimetrics. For example, although the nearest-neighbor classifier is still well-defined in semimetric spaces, and may naively be evaluated on queries in $O(n)$ time, relaxing to approximate nearest neighbors no longer provides the exponential speedup in query time that it does in metric spaces. Simply put, without the triangle inequality, the hierarchy-based search methods, such as [9, 17] and related approaches, all break down.

Paper outline. After presenting our basic definitions in Section 2, we give packing bounds and net-construction algorithms in Section 3. In Section 4 we give upper and lower bounds on sample complexity for learning in semimetrics. The density dimension of some common semimetircs is computed in Sectino 5.

2 Preliminaries

Semimetric spaces. Throughout this paper, our instance space $\mathcal{X}$ will be endowed with a semimetric $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, which is a non-negative symmetric function verifying $\rho(x, x') = 0 \iff x = x'$ for all $x, x' \in \mathcal{X}$. If the semimetric space $(\mathcal{X}, \rho)$ additionally satisfies the triangle inequality, $\rho(x,x') \le \rho(x,x'') + \rho(x'',x')$ for all $x,x',x'' \in \mathcal{X}$, then ρ is a *metric*. The distance between two sets A,B in a semimetric space is defined by $\rho(A,B) = \inf_{x \in A, x' \in B} \rho(x,x')$. For $x \in \mathcal{X}$ and $r > 0$, denote by $B_r(x) = \{y \in \mathcal{X} : \rho(x,y) < r\}$ the open r-ball about x. The *radius* of a set is the radius of the smallest ball containing it, $\mathrm{rad}(A) = \inf\{r > 0 : \exists x \in A, A \subseteq B_r(x)\}$ and $\mathrm{diam}(A) := \sup_{x,x' \in A} \rho(x,x')$.

Doubling and density dimensions. Let $\lambda = \lambda(\mathcal{X})$ be the smallest number such that every open ball in $\mathcal{X}$ can be covered by λ open balls of half the radius, where all balls are centered at points of $\mathcal{X}$. Formally,

$$\begin{aligned}\lambda(\mathcal{X}) = \min\{&\lambda \in \mathbb{N} : \forall x \in \mathcal{X}, r > 0 \\ &\exists x_1, \ldots, x_\lambda \in \mathcal{X} : B_r(x) \subseteq \cup_{i=1}^{\lambda} B_{r/2}(x_i)\}.\end{aligned}$$

Then λ is the *doubling constant* of $\mathcal{X}$, and the *doubling dimension* of $\mathcal{X}$ is $\mathrm{ddim}(\mathcal{X}) = \log_2 \lambda$.

An *r-net* of a set $A \subseteq \mathcal{X}$ is any *maximal* subset A having mutual interpoint distance at least r. The r-packing number $\mathcal{M}(r,A)$ of A is the maximum size of any r-net of A:

$$\begin{aligned}\mathcal{M}(r,A) = \max\{&|E| : E \subseteq A, \\ &(x,y \in E) \wedge (x \neq y) \implies \rho(x,y) \ge r\}.\end{aligned} \tag{1}$$

The *density constant* $\mu(\mathcal{X})$ was defined in [20] as the smallest number such that any open r-radius ball in $\mathcal{X}$ contains at most μ points at mutual interpoint distance at least $r/2$:

$$\begin{aligned}\mu(\mathcal{X}) = \min\{&\mu \in \mathbb{N} : (x \in \mathcal{X}) \wedge (r > 0) \\ &\implies \mathcal{M}\left(\frac{r}{2}, B_r(x)\right) \le \mu\},\end{aligned} \tag{2}$$

and we define the *density dimension* of $\mathcal{X}$ by $\mathrm{dens}(\mathcal{X}) = \log_2 \mu(\mathcal{X})$. An important property of the density dimension is that it is *hereditary*: for $S \subset \mathcal{X}$, we have $\mu(S) \le \mu(\mathcal{X})$; the doubling dimension is only approximately hereditary [20].

It will be convenient to define $\mathrm{Log}\,(x) := \log_2 \lceil x \rceil$, and we will make frequent use of the identity

$$\mu(S)^{\mathrm{Log}(\alpha)} = \lceil \alpha \rceil^{\mathrm{dens}(S)}. \tag{3}$$

Learning model. We work in the standard *agnostic* learning model [30, 34], whereby the learner receives a sample S consisting of n labeled examples (X_i, Y_i), drawn iid from an unknown distribution over $\mathcal{X} \times \{-1,1\}$. All subsequent probabilities and expectations will be with respect to this distribution. Based on the training sample S, the learner produces a *hypothesis* $h : \mathcal{X} \to \{-1,1\}$, whose *empirical error* is defined by $\widehat{\mathrm{err}}(h) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{h(X_i) \neq Y_i\}}$ and whose *generalization error* is defined by $\mathrm{err}(h) = \mathbb{P}(h(X) \neq Y)$.

Sub-sample, margin, and induced 1-NN. In a slight abuse of notation, we will blur the distinction between $S \subset \mathcal{X}$ as a collection of points in a semimetric space and $S \in (\mathcal{X} \times \{-1,1\})^n$ as a sequence of labeled examples. Thus, the notion of a *sub-sample* $\tilde{S} \subset S$ partitioned into its positively and negatively labeled subsets as $\tilde{S} = \tilde{S}_+ \cup \tilde{S}_-$ is well-defined. The *margin* of $\tilde{S}$, defined by $\mathrm{marg}(\tilde{S}) = \rho(\tilde{S}_+, \tilde{S}_-)$, is the minimum distance between a pair of opposite-labeled points. In degenerate cases where one of $\tilde{S}_+, \tilde{S}_-$ is empty, $\mathrm{marg}(\tilde{S}) = \infty$. (For ease of presentation, we assume that the margin is strictly less than $\mathrm{rad}(S)$, and so $\mathrm{rad}(S)/\mathrm{marg}(S) > 1$. In the case of equality, substituting any value less than the margin will cause the relevant claim to hold.) A sub-sample $\tilde{S}$ naturally induces the 1-NN classifier $h_{\tilde{S}}$, via $h_{\tilde{S}}(x) = \mathrm{sign}(\rho(x, \tilde{S}_-) - \rho(x, \tilde{S}_+))$.

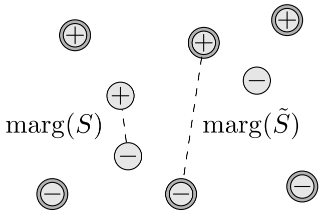

Figure 1: In this example, the sub-sample $\tilde{S} \subset S$ is indicated by double circles. It is always the case that $\mathrm{marg}(\tilde{S}) \ge \mathrm{marg}(S)$.

The problem of *nearest-neighbor condensing* is to produce the minimal subsample $\tilde{S} \subset S$ so that the 1-NN classifier $h_{\tilde{S}}$ is *consistent* with S, i.e. has zero training error. In the *inconsistent* version of this problem, one is given a parameter $\varepsilon > 0$ and tasked with producing a minimal subsample $\tilde{S} \subset S$ whose induced $h_{\tilde{S}}$ has training error at most ε.

3 Packing bounds and algorithms

The central contribution of this section is the following lemma, which demonstrates that for a semimetric space, a bound on its density dimension implies one on its packing numbers.

Lemma 1. *For any point set S in a semimetric space $\mathcal{X}$ and $r > 0$, the size of any r-net of S is $k = (\mathrm{rad}(S)/r)^{O(\mathrm{dens}(S))}$, and furthermore, such an r-net can be computed in time $O(k|S|)$.*

Proof. To bound the size of a maximal r-net $C \subset S$, suppose its radius is R. Partition C into clusters by extracting from C an arbitrary net D with minimum interpoint distance $R/2$, and assigning each point $p \in C$ to a cluster centered at the nearest neighbor of p in D. Then apply the procedure recursively to each cluster, halving the previous radius, until reaching point sets with minimum interpoint distance at least r. By repeatedly applying the definition of the density constant, the size of C is bounded by $\mu(S)^{\operatorname{Log}(\operatorname{rad}(S)/r)} = (\operatorname{rad}(S)/r)^{O(\operatorname{dens}(S))}$.

The actual r-net is constructed in a greedy fashion. Initialize set $C = \emptyset$, and for every point in S, add it to C if its closest neighbor in C is at distance r or greater. Since $|C| \leq k$, the total runtime is $O(k|S|)$. See Algorithm 1. □

Algorithm 1 Brute-force net construction

```
Require: sample $S$, margin $r$
Ensure: $C$ is an $r$-net for $S$
  for $x \in S$ do
    if $\rho(x, C) \geq r$ then
      $C = C \cup \{x\}$
    end if
  end for
```

Having demonstrated the existence of a small r-net, we can now consider the problems of producing both consistent (lossless) and inconsistent (lossy) 1-NN classifiers for the sample (see Section 2).

Consistent case. For a labeled sample S, recall that the *margin* of S is the minimum distance between oppositely labeled points in S, as defined formally in Section 2. The margin of a given sample can be computed in time $\Theta(|S|^2)$ by considering all pairs of points. We begin with a consistent classifier, whose generalization performance with explicit constants is analyzed in Theorem 6(i). Informally, the latter states that a 1-nearest neighbor classifier induced by a sub-sample of size k has generalization error $O\left(\frac{k \log n + \log \frac{1}{\delta}}{n}\right)$.

Theorem 2. *Let S be a sample set equipped with a semimetric distance function, and let the margin $\gamma = \operatorname{marg}(S)$ be given. In time $O(k|S|)$ we can construct a nearest-neighbor classifier that achieves zero training error on S, where $k = (\operatorname{rad}(S)/\gamma)^{O(\operatorname{dens}(S))}$. The evaluation time for a test point is $\Theta(k)$, and with probability $1-\delta$, the resulting classifier has generalization error $O\left(\frac{k \log n + \log \frac{1}{\delta}}{n}\right)$.*

Proof. We build a γ-net C for S in time $O(k|S|)$, as in Lemma 1. Since γ is the margin, by construction every point in S has the same label as its nearest neighbor in C, and so the nearest neighbor classifier with respect to C has zero sample error.

Given a test point x, we assign it the same label as its nearest neighbor in C. Then the generalization bounds follow from Theorem 6(i). For the runtime, $O(k)$ operations are clearly sufficient to find the nearest neighbor. □

Remark 3. *If the margin is not known in advance, then it must be computed, and the runtime in Theorem 2 grows to $O(n^2)$. In this case we can give an alternate construction that achieves the runtime of $O(k|S|)$ of the Theorem. Extract sets $S_+, S_- \in S$ of oppositely labeled points in S, and for each set build a subset in a manner similar to the proof of Lemma 1: Let R be the maximum among the radii of S_+, S_-. Partition S_+ into clusters by extracting from S_+ an arbitrary net D with minimum interpoint distance $R/2$, and assigning each point $p \in S_+$ to a cluster centered at the nearest neighbor of p in D. Then apply the procedure recursively to each cluster in D, halving the previous radius. This construction is done to S_- in parallel, and terminates when the union of the subsets for S_+ and S_- is consistent with S. This must occur within $O(\log(\operatorname{rad}(S)/\operatorname{marg}(S)))$ iterations, producing a consistent set of size k in time $O(kn \log(\operatorname{rad}(S)/\operatorname{marg}(S))) = O(kn)$ (where equality follows from the* $\log(\operatorname{rad}(S)/\operatorname{marg}(S))$ *term being subsumed in the definition of k).*

The procedure in Theorem 2 compresses S, producing a consistent sub-sample C. Immediate from the theorem is that the smaller the compressed set C, the better the generalization bounds of the classifier. However, as [18] recently demonstrated, even in metric spaces, it is NP-hard to approximate the size of the minimum consistent subset to within a factor $2^{O(\operatorname{ddim}(S)\log(\operatorname{rad}(S)/\operatorname{marg}(S))^{1-o(1)}} = 2^{O(\operatorname{dens}(S)\log(\operatorname{rad}(S)/\operatorname{marg}(S))^{1-o(1)}}$. This means that choosing the net of Lemma 1 is close to the optimal construction for a consistent subset of S.

Inconsistent case. It is natural to ask whether allowing the classifier nonzero sample error results in improved generalization bounds. This is indeed generally the case, as the bound in Theorem 7 indicates. Informally, the latter shows that a 1-nearest neighbor classifier induced by a sub-sample of size k with sample error ε has generalization error $Q(k, \varepsilon) = O\left(\varepsilon + \frac{1}{n}\log\frac{n^k}{\delta} + \sqrt{\frac{\varepsilon}{n}\log\frac{n^k}{\delta}}\right)$. Optimizing this bound is an instance of Structural Risk Minimization (SRM). Unfortunately, we can show SRM to be infeasible for this problem, and that the generalization guarantees of Theorem 2 are nearly the best that can be obtained by way of Theorem 7:

Theorem 4. *Given a set S equipped with a metric or semimetric distance function, let $S^* \subset S$ be a sub-sample for which the generalization bound $Q(d, \varepsilon)$ in Theorem 7 (for a fixed constant δ) is minimized. Then it is NP-hard to compute any subset of S achieving a generalization bound within factor $2^{O((\operatorname{dens}(S)\log(\operatorname{rad}(S)/\operatorname{marg}(S)))^{1-o(1)}}$ of the generalization bound induced by S^*.*

Proof. The proof is via reduction from the minimum consistent subset problem mentioned above. Fix the confidence level δ in the bound, let T be an instance of the minimum consistent subset problem, and put $m = |T|$. For some large value p, replace each point $t_i \in T$ with a (similarly labeled) set of p points $s_{i,1}, \ldots, s_{i,p}$ obeying the line metric, with $\rho(s_{i,a}, s_{i,b}) = \phi|a-b|$ for an infinitesimally small ϕ. Put $\rho(s_{i,a}, s_{j,b}) = \rho(t_i, t_j)$. The new set is S, with $n = |S| = pm$.

Consider a subset $S' \subset S$. If the 1-NN rule on S' misclassifies a point of S, say $s_{i,a}$, then in fact it must misclassify all p points $s_{i,b}$, $b \in [1,p]$. So an inconsistent subset of S achieves a value of $Q(|S'|, p/n) = \Omega(p/n)$ in the generalization bound.

Now consider the consistent subset of S consisting of $m = n/p$ points $s_{i,1}$ for $i \in [1,m]$. By Theorem 6(i), this classifier achieves a generalization bound of $O\left(\frac{m \log n}{n}\right) = O\left(\frac{\log n}{p}\right)$. So when $p = \Omega(\sqrt{n \log n})$, this consistent classifier is better than any inconsistent classifier. Now a consistent subset of size $d \leq m$ has generalization bound $O\left(\frac{d \log n}{n}\right)$. As it is NP-hard to find a subset whose size is within a factor $2^{O(\mathrm{dens}(S) \log(\mathrm{rad}(S)/\mathrm{marg}(S))^{1-o(1)}}$ of the smallest consistent subset, it is NP-hard to find a consistent subset with generalization bound within a factor $2^{O(\mathrm{dens}(S) \log(\mathrm{rad}(S)/\mathrm{marg}(S))^{1-o(1)}}$ of the optimal consistent subset, and the theorem follows. □

In light of the hardness result established in Theorem 4, we specialize the goal from one of seeking a small nearly consistent sub-sample to one where the sub-sample must be a γ-net. In this case, the relevant generalization bound is provided by Theorem 9. As before, we wish to perform SRM for this bound. Fortunately, we are able to compute the latter exactly in polynomial time, and even more efficiently if we are willing to settle for a solution within a constant factor of the optimal. The proof of the following theorem follows the lines of [17].

Theorem 5. *Given a sample set S equipped with a semimetric:*

(a) A nearest-neighbor classifier minimizing the generalization bound of Theorem 9 can be computed in randomized time $O(|S|^{4.373})$.

(b) A nearest-neighbor classifier whose generalization bound is within factor 2 of optimal can be computed in deterministic time $O(|S|^2 \log |S|)$.

Each of these classifiers can be evaluated on test points in time $\left(\frac{\mathrm{rad}(S)}{\gamma}\right)^{O(\mathrm{dens}(S))}$, where γ is the margin imposed by the SRM procedure.

Proof. For each of these solutions, we enumerate and sort in increasing order the distances between all oppositely labeled point pairs in S, in total time $O(|S|^2 \log |S|)$. Each distance constitutes a separate guess for the optimal margin to "impose" on S. That is, for each distance γ, we will remove from S some points to ensure that all opposite labeled pairs are more than γ far apart.

To accomplish this, we iteratively build a new graph G. We initialize G with vertices representing the points of S. At each round we add to G an edge between the next closest pair of opposite labeled points, as given by the sorted enumeration above. This distance is the margin of the current round: Points connected by an edge in G represent pairs that are too close together for the current margin, and we need to compute how many points would need to be removed from G in order for no edge to remain in the graph. (However, no points or edges will actually removed from G.) As observed by [17], this task is precisely the problem of bipartite vertex cover. By König's theorem, the minimum vertex cover problem in bipartite graphs is equivalent to the maximum matching problem, and a maximum matching in bipartite graphs can be computed in randomized time $O(n^{2.373})$ [31, 39]. So for each candidate margin, we can compute in $O(n^{2.373})$ time the number of points that must be removed from the current graph G in order to remove all edges. For $O(n^2)$ possible margins, this amounts to $O(n^{4.373})$ time. Having computed for each interpoint distance the number of points required to be deleted to achieve this distance, we choose the distance-number pair which minimizes the bound of Theorem 9. We then remove the corresponding points from S, and use the algorithm of Lemma 1 to construct a net satisfying the margin bound.

The runtime improvement in (b) comes from a faster vertex-cover computation. It is well known that a 2-approximation to vertex cover can be computed (in arbitrary graphs) by a greedy algorithm in time linear in the graph size $O(|V^+ \cup V^-| + |E|) = O(n^2)$, see e.g. [6]. This algorithm simply chooses any edge and removes both endpoints, until no edges remain. We apply this algorithm to our setting: A new edge is added to G only if both endpoints survive in the already computed cover, and then both endpoints are marked as removed in the solution to the new graph. Having computed for each interpoint distance the number of points required to be deleted to achieve this distance, we choose the distance-number pair which minimizes the bound of Theorem 9. We then remove the corresponding points from S, and use the algorithm of Lemma 1 to construct a net satisfying the margin bound. The runtime is dominated by the time required to sort the distances.

For both algorithms, a new point is classified by finding its nearest neighbor in the extracted net. □

4 Generalization guarantees

In this section, we provide general sample compression bounds, which then will be specialized to the nearest-neighbor classifier proposed above. Theorem 7 presents a smooth interpolation between two classic bounds: the consistent case with rate $\tilde{O}(1/n)$, and the agnostic case with rate $\tilde{O}(1/\sqrt{n})$. Applied to margin-based semimetric sample-compression schemes, this result yields the efficiently computable and optimizable bound in Theorem 9, which is nearly optimal (as shown in Theorem 4). Finally, the lower bound in Theorem 10 shows that even under margin assumptions, there exist adversarial distributions forcing the sample complexity to be exponential in dens.

4.1 Sample compression schemes

We use the notion of a *sample compression scheme* in the sense of [21], where it is treated in full rigor. Informally, a learning algorithm maps a sample S of size n to a hypothesis h_S. It is a d-sample compression scheme if a sub-sample of size d suffices to produce a hypothesis that agrees with the labels of all the n points. It is an *ε-lossy* d-sample compression scheme if a sub-sample of size d suffices to produce a hypothesis that disagrees with the labels of at most εn of the n sample points.

The algorithm need not know d and ε in advance. We say that the sample S is *(d, ε)-compressible* if the algorithm succeeds in finding an ε-lossy d-sample compression scheme for this particular sample. In this case:

Theorem 6 ([21])**.** *For any distribution over $\mathcal{X} \times \{-1, 1\}$, any $n \in \mathbb{N}$ and any $0 < \delta < 1$, with probability at least $1 - \delta$ over the random sample S of size n, the following holds:*

(i) If S is $(d, 0)$-compressible, then $\operatorname{err}(h_S) \leq \frac{1}{n-d}\left((d+1)\log n + \log\frac{1}{\delta}\right)$.

(ii) If S is (d, ε)-compressible, then $\operatorname{err}(h_S) \leq \frac{\varepsilon n}{n-d} + \sqrt{\frac{(d+2)\log n + \log\frac{1}{\delta}}{2(n-d)}}$.

The generalizing power of sample compression was independently discovered by [28, 13], and later elaborated upon by [21]. The bounds above are already quite usable, but they feature an abrupt transition from the $(\log n)/n$ decay in the lossless ($\varepsilon = 0$) regime to the $\sqrt{(\log n)/n}$ decay in the lossy regime. We now provide a smooth interpolation between the two (such results are known in the literature as "fast rates" [10]; see also a related result in [34]):

Theorem 7. *Fix a distribution over $\mathcal{X} \times \{-1, 1\}$, an $n \in \mathbb{N}$ and $0 < \delta < 1$. With probability at least $1 - \delta$ over the random sample S of size n, the following holds for all $0 \leq \varepsilon \leq \frac{1}{2}$: If S is (d, ε)-compressible, then, putting $\tilde{\varepsilon} = \varepsilon n/(n-d)$, we have*

$$\begin{aligned}\operatorname{err}(h_S) &\leq \tilde{\varepsilon} + \frac{2}{3(n-d)}\log\frac{n^{d+2}}{\delta} \\ &\quad + \sqrt{\frac{9\tilde{\varepsilon}(1-\tilde{\varepsilon})}{2(n-d)}\log\frac{n^{d+2}}{\delta}} \\ &=: Q(d, \varepsilon). \end{aligned} \tag{4}$$

Proof. Deferred to the journal version. □

4.2 Margin-based nearest neighbor compression

We now specialize the general sample compression result of Theorem 7 to our setting, where $h_{S'}$ induced by a sub-sample $S' \subset S$ is given by the 1-NN classifier defined in Section 2. Any sample S of size n is trivially $(n, 0)$-compressible and $(0, \frac{1}{2})$-compressible — the former is achieved by not compressing at all, and the latter by a constant predictor. Now d and ε cannot simultaneously be made arbitrarily small, and for non-degenerate samples S, the bound Q in Theorem 7 will have a nontrivial minimal value Q^*. Theorem 4 shows that computing Q^* is intractable and the algorithm in Theorem 5 solves a tractable modification of this problem. For $k \in \mathbb{N}$ and $\gamma > 0$, let us say that the sample S is *(k, γ)-separable* if it admits a sub-sample $S' \subset S$ such that $|S \setminus S'| \leq k$ and $\operatorname{marg}(S') > \gamma$, and observe that separability implies compressibility:

Lemma 8. *If S is (k, γ)-separable then it is $\left(\lceil \operatorname{rad}(S)/\gamma \rceil^{\operatorname{dens}(S)}, k/|S|\right)$-compressible.*

Proof. Suppose $S' \subset S$ is a witness of (k, γ)-separability. Being pessimistic, we will allow our lossy sample compression scheme to mislabel all of $S \setminus S'$, but not any of S', giving it a sample error $\varepsilon \leq k/|S|$. Now by construction, S' is $(0, \gamma)$-separable, and thus a γ-net $\tilde{S} \subset S'$ suffices to recover the correct labels of S' via 1-nearest neighbor. Lemma 1 provides the estimate

$$|\tilde{S}| \leq \mu(S)^{\operatorname{Log}(\operatorname{rad}(S)/\gamma)} = \lceil \operatorname{rad}(S)/\gamma \rceil^{\operatorname{dens}(S)},$$

whence the compression bound. □

These observations culminate in an efficiently optimizable margin-based generalization bound:

Theorem 9. *Fix a distribution over $\mathcal{X}$, an $n \in \mathbb{N}$ and $0 < \delta < 1$. With probability at least $1 - \delta$ over the random sample S of size n, the following holds for all $0 \le k \le n/2$: If S is (k, γ)-separable with witness S', then*

$$\operatorname{err}(h_{S'}) \le Q(d, k/n) =: R(k, \gamma),$$

where Q is defined in (4) and

$$d = \mu(S')^{\operatorname{Log}\left(\operatorname{rad}(S')/\gamma\right)} = \lceil \operatorname{rad}(S')/\gamma \rceil^{\operatorname{dens}(S)}.$$

Furthermore, the minimizer (k^*, γ^*) of $R(\cdot, \cdot)$ is efficiently computable.

4.3 Sample complexity lower bound

The following result shows that even under margin assumptions, a sample of size exponential in dens will be required for some distributions.

Theorem 10. *There are universal constants $c, \delta > 0$ such that for every semimetric space $(\mathcal{X}, \rho)$ with* $\operatorname{dens}(\mathcal{X}) > 6$ *and any learning algorithm mapping samples S of size n to hypotheses $h_n : \mathcal{X} \to \{-1, 1\}$, there is a distribution $\mathbb{P}$ over $\mathcal{X}$ and a target concept $f : \mathcal{X} \to \{-1, 1\}$, such that* $\operatorname{err}(f) = 0$ *yet*

$$\mathbb{P}\left(\operatorname{err}(h_n) \ge \frac{c \lceil \operatorname{rad}(S)/\operatorname{marg}(S) \rceil^{\operatorname{dens}(\mathcal{X})}}{n}\right) \ge 1 - \delta.$$

Proof. Deferred to the journal version. □

5 Density dimension of some common semimetrics

In this section we demonstrate the utility of the density dimension by calculating its value under some common semimetric distance functions on d-dimensional vectors. The first of these functions is the Jensen-Shannon divergence, equivalent [14] to the ℓ_2-squared distance function $\ell_2^2(x, y) = \sum_{i=1}^d |x_i - y_i|^2$. We also consider the non-metric ℓ_p-spaces for $0 < p < 1$, $\ell_p(x, y) = (\sum_{i=1}^d |x_i - y_i|^p)^{1/p}$. Finally, we consider the k-median Hausdorff distance.

Recall that the usual Hausdorff distance is a metric defined on any two point sets A and B, and we shall make the simplifying assumption that $|A| = |B| = m$. Let $l(a, b)$ for all $a \in A$ and $b \in B$ be a vector distance function — for simplicity we shall assume the Euclidean ℓ_2 distance — and $l(a, B)$ be the distance from $a \in A$ to its nearest neighbor in B. The Hausdorff distance is the maximum distance between a point in A (or B) and its nearest neighbor in B (respectively, in A): $\max\{\max_{a \in A} l(a, B), \max_{b \in B} l(b, A)\}$. [22] define the k-median Hausdorff distance (in the terminology of [25], but perhaps more aptly termed the k-rank Hausdorff distance) by setting $h_k(A, B)$ to be the k-th smallest value in the vector $v = (l(a_1, B), \ldots, l(a_d, B))$, and then the k-rank Hausdorff distance is $H_k(A, B) = \max\{h_k(A, B), h_k(B, A)\}$. Note that $H_m(A, B)$ recovers the classic metric Hausdorff distance (and we require $k \le m$). On the other hand, we can show that $H_1(A, B)$ is sufficiently robust to be *universal* for all semimetrics — that is, any semimetric can be realized by the distance function $H_1(A, B)$:

Lemma 11. *If ρ is a semimetric on a point set X of size n, then ρ can be realized as the H_1 distance (induced by $l = \ell_2$ as above) over subsets of $\mathbb{R}$ of size n.*

Proof. Put $D = \operatorname{diam}(X)$ and replace each point $x_i \in X$ with a set $A_i \subset \mathbb{R}$ of size n as follows. For $a_{i,j} \in A_i$, if $j \ge i$, then set $a_{i,j} = 2D((i+1)n + j)$, and otherwise set $a_{i,j} = a_{j,i} + \rho(x_i, x_j)$.

Consider any pair $x_i, x_j \in X$ for $i < j$. Clearly $\ell_2(a_{i,j} - a_{j,i}) = \ell_2(a_{i,j} - a_{i,j} - \rho(x_i, x_j)) = \rho(x_i, x_j)$, so $H_1(A_i, A_j)$ is at most this value. On the other hand, for any k, p we can show that $\ell_2(a_{i,k} - a_{j,p}) \ge D$ whenever $k \ne j$ or $p \ne i$): If $i \le k$, we have $a_{i,k} = 2D((i+1)n + k)$, and otherwise $2D((k+1)n + i) \le a_{i,k} \le 2D((k+1)n + i) + D$. Similarly, if $j \le p$ we have $a_{j,p} = 2D((j+1)n + p)$, and otherwise $2D((p+1)n + j) \le a_{j,p} \le 2D((p+1)n + j) + D$. Since by assumption $i \ne j$, the two terms differ by at least D unless both $j = k$ and $i = p$. □

We bound the density dimension under these three distance functions.

Theorem 12. *A set of m d-dimensional vectors has density dimension: $O(d)$ under ℓ_2-squared, $O(d/p^2)$ under ℓ_p ($0 < p < 1$), and $O(k(d + \log m))$ under H_k.*

Proof. We begin with a standard proof that a set of d-dimension Euclidean vectors has density dimension $O(d)$. Take any radius 4 ball, and we bound the size of a 2-net of points within this ball. By the triangle inequality, 1-radius balls centered at the 2-net points do not intersect, and so the density constant of the space is bounded by the number of 1-radius balls whose centers can be packed into the 4-radius ball. Since a piece of a 1-radius ball may escape the 4-radius ball, by the triangle inequality this term is bounded by the number of 1-radius balls that can be packed into a 5-radius ball. The ratio of the volume of a 5-radius ball to that of a 1-radius ball is 5^d, which bounds the density constant of d-dimensional Euclidean space.

For ℓ_2^2, we embed this space into ℓ_2 by simply retaining the vectors and changing only the distance function. In other words, we take the square root of all the distances, which is known as a *snowflake* operator. To bound the number of 2-net points within a ball of radius 4 in ℓ_2^2, consider instead a larger 1-net in the 4-radius ball. After the embedding, it is a 2-radius Euclidean ball containing a 1-net, and so the density constant of ℓ_2^2 is $2^{O(d)}$ as well.

For ℓ_p $(0 < p < 1)$, let us consider a snowflake of this function, that is $\ell_p^{p/2}(x,y) = (\sum_{i=1}^d |x_i - y_i|^p)^{1/2}$. We can show that the vectors under this distance function can be embedded into $O(d/p^2)$-dimensional Euclidean space with only constant distortion: Considering each coordinate separately, the distance operator $|x_i - y_i|^{p/2}$ on a single coordinate has the effect of embedding all points on a line into a helix. It is known that this embedding can be realized in $O(1/p^2)$-dimensional Euclidean space with arbitrarily small distortion (see [35] for $\frac{1}{2} < p < 1$, and [19, 7] for $0 < p \leq \frac{1}{2}$). We create such an embedding for each coordinate and then concatenate the coordinate embeddings into a single vector. This yields an embedding from d-dimensional $\ell_p^{p/2}$ into $O(d/p^2)$-dimensional Euclidean space with arbitrarily small distortion. Then a 1-net inside a $2^{2/p}$-radius ball in the original ℓ_p-space is a 1-net inside a $(2+\varepsilon)$-radius ball in the target Euclidean space (for arbitrarily small ε), and so its density dimension is $O(d/p^2)$.

For the k-rank Hausdorff distance, note that for all vector sets A, B and subsets $A' \subset A$ and $B' \subset B$, $H_k(A', B') \geq H_k(A, B)$ (since h_k is non-decreasing under deletions). Further, there always exist A', B' of size k satisfying $H_k(A', B') = H_k(A, B)$ (by combining pairs from $h_k(A, B)$ and $h_k(B, A)$ in a prudent fashion). Now consider a set $\mathcal{A}$ of vector sets all within distance 2 of center set $A_c \in \mathcal{A}$ and at mutual inter-set distance at least 1. We will show that $|\mathcal{A}| = 2^{O(kd)} m^k$, from which the item follows. To prove this, take in turn each subset $A'_c \subset A_c$ of size k (there are $\binom{m}{k} < m^k$ such subsets), and let $\mathcal{A}'$ contain all sets $A'_i \subset A_i$ of size k for which $H_k(A'_c, A'_i) = H_k(A_c, A_i)$. $\mathcal{A}'$ has radius 2 and inter-set distance at least 1. To complete the proof, we will show that $n = |\mathcal{A}'| \leq 2^{O(d+k)}$, from which it follows that $|\mathcal{A}| < |\mathcal{A}'| m^k = 2^{O(kd)} m^k$:

Since $H_k(A'_c, A'_i) \leq 2$ for all $A'_i \in \mathcal{A}'$, we have that $h_k(A'_i, A'_c) \leq 2$, and so every vector is within Euclidean distance 2 of one of the k vectors of A'_c. Let each vector of A'_c be the center of a 2-radius Euclidean ball. Clearly, the vectors of each A'_i fall into k different Euclidean balls, and A'_i falls into one of 2^k different ball configurations. Let $C \subset \mathcal{A}'$ $(p = |C|)$ include all sets falling into some specific configuration. Within each Euclidean ball, an optimal configuration clusters vectors from the p sets into $2^{O(d)}$ groups, so that the intergroup distance is at least 1, and the intragroup size is $\frac{p}{2^{O(d)}}$. Repeated over k balls, we require $\frac{p}{2^{O(kd)}} \leq 1$, and so $p \leq 2^{O(kd)}$ and $n \leq 2^k p = 2^{O(kd)}$.

□

We leave it as an open problem to improve on the dependence of k in our bound of the density dimension of the Hausdorff distance.

We conclude this section with an illustration of how the theory developed in this paper explains the success of the greedy net-based compression algorithm, even in the case of semimetrics. We present results for the Hausdorff semimetric applied to the Covertype dataset, found in the UCI Machine Learning Repository.[2] This dataset contains 7 different label types, which we treated as 21 separate binary classification problems; we report results for labels 2 vs. 5, 1 vs. 4, and 4 vs. 7.

data set	**original size**	**down to %**
Covertype 2 vs. 5	2000	97
Covertype 1 vs. 4	2000	25
Covertype 4 vs. 7	2000	2

Figure 2: Summary of the performance of semimetric sample compression algorithm.

Acknowledgements. Thanks to Shay Moran for pointing out the analogous sample compression result of [34]. A special thanks to Yevgeni Korsunsky for carefully reading the manuscript and providing very helpful comments.

References

[1] Alexandr Andoni and Robert Krauthgamer. The computational hardness of estimating edit distance. *SIAM J. Comput.*, 39(6):2398–2429, April 2010.

[2] Maria-Florina Balcan and Avrim Blum. On a theory of learning with similarity functions. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pages 73–80, 2006.

[3] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 287–298, 2008.

[4] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

[2] `http://tinyurl.com/cover-data`

[5] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 671–680, 2008.

[6] Reuven Bar-Yehuda and Shimon Even. A linear-time approximation algorithm for the weighted vertex cover problem. *J. Algorithms*, 2(2):198–203, 1981.

[7] Yair Bartal, Ben Recht, and Leonard J. Schulman. Dimensionality reduction: Beyond the johnson-lindenstrauss bound. In *SODA*, 2011.

[8] R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. In *Physics-Based Modeling in Computer Vision*, 1995.

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

[10] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[11] Dennis K. Burke. Cauchy sequences in semimetric spaces. *Proc. AMS*, 33(1):161–164, 1972.

[12] I. Cox, M. Miller, S.M. Omohundro, and P.N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *Intl. Conf. Pat. Rec.*, 1996.

[13] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[14] Bend Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory (ISIT)*, June 2004.

[15] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Trans. PAMI*, 21(12):1312–1328, 1999.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[17] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

[18] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 370–378, 2014.

[19] Lee-Ad Gottlieb and Robert Krauthgamer. A nonlinear approach to dimension reduction. In *SODA*, 2011.

[20] Lee-Ad Gottlieb and Robert Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Discrete Math.*, 27(4):1759–1769, 2013.

[21] Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[22] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. PAMI*, 15(9):850–863, 1993.

[23] J. Puzicha, J. Buhmann, Y. Rubner, C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, 1999.

[24] David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Condensing image databases when retrieval is based on non-metric distances. In *ICCV*, pages 596–601, 1998.

[25] David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(6):583–600, 2000.

[26] A. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. PAMI*, 19(12):1386–1391, 1997.

[27] Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 892–900, 2014.

[28] Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability, unpublished. 1986.

[29] M. Dubuisson, A. Jain. A modified hausdorff distance for object matching. In *Intl. Conf. Pat. Rec.*, 1994.

[30] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.

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

[32] Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1. *SIAM J. Comput.*, 37:804–826, June 2007.

[33] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[34] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[35] Michel Talagrand. Approximating a helix in finitely many dimensions. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 28(3):355–363, 1992.

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

[37] L. Wang, C. Yang, and J. Feng. On learning with dissimilarity functions. In *ICML*, 2007.

[38] Daphna Weinshall, David W. Jacobs, and Yoram Gdalyahu. Classification in non-metric spaces. In *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*, pages 838–846, 1998.

[39] Virginia Vassilevska Williams. Breaking the Coppersmith-Winograd barrier. In *STOC '12: Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, New York, NY, USA, 2012. ACM Press.

[40] Wallace Alvin Wilson. On semi-metric spaces. *American Journal of Mathematics*, 53(2):361–373, 1931.