

Nearly optimal classification for semimetrics*

Lee-Ad Gottlieb

*Department of Computer Science
Ariel University
Ariel, Israel*

LEEAD@ARIEL.AC.IL

Aryeh Kontorovich

*Department of Computer Science
Ben-Gurion University
Beer Sheva, Israel*

KARYEH@CS.BGU.AC.IL

Pinhas Nisnevitch

*Department of Computer Science
Tel-Aviv University
Tel-Aviv, Israel*

PINHASN@MAIL.TAU.AC.IL

Editor: Kevin Murphy

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 $d = (R/\gamma)^{\text{dens}}$ 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.

Keywords: semimetric, classification, compression, generalization

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

*. An extended abstract of this paper appeared in AISTATS 2016. L. Gottlieb was supported in part by the Israel Science Foundation (grant No. 755/15). A. Kontorovich was partially supported by the Israel Science Foundation (grants No. 1141/12 and 755/15) and a Yahoo Faculty award.

but may not satisfy the triangle inequality (Wilson, 1931)¹. Without the latter, quite a bit of structure is lost — for example, semimetric spaces admit convergent sequences without a Cauchy subsequence (Burke, 1972). 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 (Rubner et al., 2000). Yet representing earthmover distances using any fixed ℓ_p norm unavoidably introduces very large interpoint distortion (Naor and Schechtman, 2007), 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 (Andoni and Krauthgamer, 2010). Additional limitations of kernel methods are articulated in Balcan et al. (2008b).

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 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 (Fuglede and Topsøe, 2004; Goodfellow et al., 2014), the k -median Hausdorff distances, and the ℓ_p distances with $0 < p < 1$, which appear in vision applications (M. Dubuisson, A. Jain, 1994; Jacobs et al., 2000) — all of which are semimetrics. An additional line of work (M. Dubuisson, A. Jain, 1994; Jacobs et al., 2000, 1998; Weinshall et al., 1998) underscored the effectiveness of non-metric distances in various applications (mainly vision), and among these, semimetrics again play a prominent role (Basri et al., 1995; Cox et al., 1996; Gdalyahu and Weinshall, 1999; Huttenlocher et al., 1993; Jain and Zongker, 1997; J. Puzicha, J. Buhmann, Y. Rubner, C. Tomasi, 1999).

Main results. We initiate the rigorous study of classification for semimetric spaces. We define the *density dimension* ($\mathbf{dens} = \mathbf{dens}(\mathcal{X})$) 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 \mathbf{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

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.

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 $P=NP$ (Theorem 4).

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 8). 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 11 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 13). 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 12).

Related work. In a series of papers, Balcan et al. (2008c,a,b) 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 Wang et al. (2007) 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 (von Luxburg and Bousquet, 2004; Gottlieb et al., 2014a; Shalev-Shwartz and Ben-David, 2014; Kontorovich and Weiss, 2014), and that exponential dependence on ddim is in general unavoidable (Shalev-Shwartz and Ben-David, 2014). 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 (Gottlieb et al., 2014b).

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 (Lemma 17). Simply put, without the triangle inequality, the hierarchy-based search methods, such as Beygelzimer et al. (2006); Gottlieb et al. (2014a) 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 semimetrics is computed in Section 5. The Appendix contains a detailed discussion demonstrating the inapplicability of existing metric-space techniques to the semimetric case.

2. Preliminaries

Semimetric spaces. Throughout this paper, our instance space \mathcal{X} will be endowed with a semimetric $\rho : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$, 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}, ρ) additionally satisfies the triangle inequality, $\rho(x, x') \leq \rho(x, x'') + \rho(x'', x')$ 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, $\text{rad}(A) = \inf \{r > 0 : \exists x \in A, A \subseteq B_r(x)\}$ and $\text{diam}(A) := \sup_{x, x' \in A} \rho(x, x')$.

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,

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

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

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 :

$$\mathcal{M}(r, A) = \max\{|E| : E \subseteq A, (x, y \in E) \wedge (x \neq y) \implies \rho(x, y) \geq r\}. \quad (1)$$

The *density constant* $\mu(\mathcal{X})$ was defined in Gottlieb and Krauthgamer (2013) 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 $r/2$:

$$\mu(\mathcal{X}) = \min\{\mu \in \mathbb{N} : (x \in \mathcal{X}) \wedge (r > 0) \implies \mathcal{M}\left(\frac{r}{2}, B_r(x)\right) \leq \mu\}, \quad (2)$$

and we define the *density dimension* of \mathcal{X} by $\text{dens}(\mathcal{X}) = \log_2 \mu(\mathcal{X})$. A closely related notion was introduced by Vol'berg and Konyagin (1987) under the term *uniform metric dimension*.

An important property of the density dimension is that it is *hereditary*: for $S \subset \mathcal{X}$, we have $\mu(S) \leq \mu(\mathcal{X})$; the doubling dimension is only approximately hereditary (Gottlieb and Krauthgamer, 2013).

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

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

Learning model. We work in the standard *agnostic* learning model (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014), 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* $h : \mathcal{X} \rightarrow \{-1, 1\}$, whose *empirical error* is defined by $\widehat{\text{err}}(h) = n^{-1} \sum_{i=1}^n \mathbf{1}_{\{h(X_i) \neq Y_i\}}$ and whose *generalization error* is defined by $\text{err}(h) = \mathbb{P}(h(X) \neq Y)$.

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 $\text{marg}(\tilde{S}) = \rho(\tilde{S}_+, \tilde{S}_-)$, is the minimum distance between a pair of opposite-labeled points (see Fig. 1). In degenerate cases where one of \tilde{S}_+, \tilde{S}_- is empty, $\text{marg}(\tilde{S}) = \infty$. (For ease of presentation, we assume that the margin is strictly less than $\text{rad}(S)$, and so $\text{rad}(S)/\text{marg}(S) > 1$. 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 $h_{\tilde{S}}(x) = \text{sign}(\rho(x, \tilde{S}_-) - \rho(x, \tilde{S}_+))$.

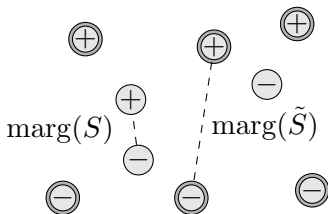


Figure 1: In this example, the sub-sample $\tilde{S} \subset S$ is indicated by double circles. It is always the case that $\text{marg}(\tilde{S}) \geq \text{marg}(S)$.

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 $k = \lceil \text{rad}(S)/r \rceil^{\text{dens}(S)}$, 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 $\mu(S)^{\text{Log}(\text{rad}(S)/r)} = \lceil \text{rad}(S)/r \rceil^{\text{dens}(S)}$.

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 7(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 $\gamma = \text{marg}(S)$ be given. In time $O(k|S|)$ we can construct a nearest-neighbor classifier that achieves zero training error on S , where $k = (\text{rad}(S)/\gamma)^{O(\text{dens}(S))}$. The evaluation on a test point requires $\Theta(k)$ distance computations, 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 7(i). For the runtime, $O(k)$ operations are clearly sufficient to find the nearest neighbor (see also Lemmas 16 and 17). ■

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 $O(\log(\text{rad}(S)/\text{marg}(S)))$ iterations, producing a consistent set of size k in time $O(kn \log(\text{rad}(S)/\text{marg}(S))) = O(kn)$ (where equality follows from the $\log(\text{rad}(S)/\text{marg}(S))$ 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 Gottlieb et al. (2014b) recently demonstrated, even in metric spaces, it is NP-hard to approximate the size of the minimum consistent subset to within a factor $2^{O(\text{ddim}(S) \log(\text{rad}(S)/\text{marg}(S))^{1-o(1)})} = 2^{O(\text{dens}(S) \log(\text{rad}(S)/\text{marg}(S))^{1-o(1)})}$ (where the equality follows from Lemma 15). 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 8 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 8:

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 $Q(k, \varepsilon)$ in Theorem 8 (for a fixed constant δ) is minimized. Then it is NP-hard to compute any subset of S achieving a generalization bound within factor $2^{O((\text{dens}(S) \log(\text{rad}(S)/\text{marg}(S)))^{1-o(1)})}$ 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 $s_{i,1}, \dots, s_{i,p}$ 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 7(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 $2^{O(\text{dens}(S) \log(\text{rad}(S)/\text{marg}(S))^{1-o(1)})}$ of the smallest consistent subset, it is NP-hard to find a consistent subset with generalization bound within a factor $2^{O(\text{dens}(S) \log(\text{rad}(S)/\text{marg}(S))^{1-o(1)})}$ of the optimal consistent subset, and the theorem follows. \blacksquare

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 10. 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 Gottlieb et al. (2014a).

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

- (a) *A nearest-neighbor classifier minimizing the generalization bound of Theorem 10 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 $\left(\frac{\text{rad}(S)}{\gamma}\right)^{O(\text{dens}(S))}$, 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 be removed from G .) As observed by Gottlieb et al. (2014a), 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 $O(n^{2.373})$ (Mucha and Sankowski, 2004; Williams, 2012). 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 10. 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. Bar-Yehuda and Even (1981). 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 10. 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. ■

Remark 6 *It is worth asking whether a successful learner in (semi)metric space must necessarily perform some type of compression. Moran and Yehudayoff (2016) recently demonstrated that for classifiers of finite VC-dimension, learning and compression are in fact equivalent. However, no such result is known — or even conjectured — for infinite VC-dimension classifiers such as 1-NN and its variants.*

If a (semi)metric learner does choose to learn by compression, he is not necessarily restricted to learning via 1-NN, but can make use of any proximity-based function. But even then, Theorems 10 and 11 below demonstrate that from an information-theoretic standpoint, roughly $(\text{rad}(S)/\text{marg}(S))^{\text{dens}(\mathcal{X})}$ examples are both necessary in the worst case and always sufficient. One could still hope to exhibit a learning problem in a (semi)metric space where any learner achieving small generalization must solve some hard problem. Results of this type are known for finite automata under cryptographic assumptions (Kearns and Vazirani, 1997, Section 1.4); could something similar be shown for proximity-based learning?

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 8 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 10, which is nearly optimal (as shown in Theorem 4). Finally, the lower bound in Theorem 11 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 Graepel et al. (2005), 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. At this stage, we are not assuming anything about the computational efficiency of a compression scheme.

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 7 (Graepel et al. (2005)) *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 $\text{err}(h_S) \leq \frac{1}{n-d} \left((d+1) \log n + \log \frac{1}{\delta} \right)$.*
- (ii) *If S is (d, ε) -compressible, then $\text{err}(h_S) \leq \frac{\varepsilon n}{n-d} + \sqrt{\frac{(d+1) \log n + \log \frac{1}{\delta}}{2(n-d)}}$.*

The generalizing power of sample compression was independently discovered by Littlestone and Warmuth (1986); Devroye et al. (1996), and later elaborated upon by Graepel et al. (2005). 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” (Boucheron et al., 2005); see also a related result in Shalev-Shwartz and Ben-David (2014)):

Theorem 8 *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*

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

Proof We closely follow the argument in Graepel et al. (2005, Theorem 2), with the twist that instead of Hoeffding’s inequality, we use Bernstein’s. The particular form of the latter is due to Dasgupta and Hsu (2008, Lemma 1): if $\hat{p} \sim \text{Bin}(n, p)/n$ and $\delta > 0$, then the following holds with probability at least $1 - \delta$:

$$p \leq \hat{p} + \frac{2}{3n} \log \frac{1}{\delta} + \sqrt{\frac{9\hat{p}(1-\hat{p})}{2n} \log \frac{1}{\delta}}. \tag{5}$$

Now suppose that S is $(d, k/n)$ -compressible, as witnessed by some sub-sample $\tilde{S} \subset S$ of size d . In particular, the hypothesis $h_{\tilde{S}}$ induced by the sub-sample \tilde{S} makes k or fewer mistakes on the $n - d$ points in $S \setminus \tilde{S}$. Substituting $p = \text{err}(h_{\tilde{S}})$ and

$$\hat{p} = \widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) := \frac{1}{|S \setminus \tilde{S}|} \sum_{x \in S \setminus \tilde{S}} \mathbf{1}_{\{h_{\tilde{S}} \text{ errs on } x\}} \leq \frac{k}{n - d} = \tilde{\varepsilon}$$

into (5) yields that for fixed \tilde{S} and random $S \setminus \tilde{S}$, with probability at least $1 - \delta$,

$$\text{err}(h_{\tilde{S}}) \leq \widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) + \frac{2}{3(n - d)} \log \frac{1}{\delta} + \sqrt{\frac{9\tilde{\varepsilon}(1 - \tilde{\varepsilon})}{2(n - d)}} \log \frac{1}{\delta}, \quad (6)$$

where we used the monotonicity of $t \mapsto t(1 - t)$ on $[0, \frac{1}{2}]$. To see that (6) follows from (5), note that when \tilde{S} of size d is fixed and $S \setminus \tilde{S}$ is drawn iid $\sim \mathbb{P}$, we have $(n - d)\widehat{\text{err}}_{S \setminus \tilde{S}}(h_{\tilde{S}}) \sim \text{Bin}(n - d, \text{err}(h_{\tilde{S}}))$. To make (6) hold simultaneously for all $\tilde{S} \subseteq S$, divide δ by n^d — the number of ways to choose a (multi)set \tilde{S} of size d . To make the claim hold for all $d \in [n]$ and all $0 \leq \varepsilon < 1$, stratify (as in Graepel et al. (2005, Lemma 1)) over the n possible choices of d , which amounts to dividing δ by an additional factor of n . Unlike in Graepel et al. (2005), there is no need to stratify over the possible values of ε , since (5) holds for random (and not just fixed) \hat{p} . \blacksquare

4.2 Margin-based nearest neighbor compression

We now specialize the general sample compression result of Theorem 8 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 8 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 $\text{marg}(S') > \gamma$, and observe that separability implies compressibility:

Lemma 9 *If S is (k, γ) -separable then it is $(\lceil \text{rad}(S)/\gamma \rceil^{\text{dens}(S)}, k/|S|)$ -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

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

whence the compression bound. \blacksquare

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

Theorem 10 *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 $0 \leq k \leq n/2$: If S is (k, γ) -separable with witness S' , then*

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

where Q is defined in (4) and

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

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. (See also the recent agnostic lower bound for passive 1-NN, Kontorovich et al. (2016, Theorem 5), based on (Berend and Kontorovich, 2015, Equation (25)) an exact PAC lower bound (Kontorovich and Pinelis, 2016, Theorem 2.1).)

Theorem 11 *There are universal constants $c, \delta > 0$ such that for every semimetric space (\mathcal{X}, ρ) with $\text{dens}(\mathcal{X}) > 6$ and any learning algorithm mapping samples S of size n to hypotheses $h_n : \mathcal{X} \rightarrow \{-1, 1\}$, there is a distribution \mathbb{P} over \mathcal{X} and a target concept $f : \mathcal{X} \rightarrow \{-1, 1\}$, such that $\text{err}(f) = 0$ yet*

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

Proof The definition of the density constant implies the existence of an $r > 0$ and an $M \subseteq \mathcal{X}$ satisfying

- (a) $m := |M| = \mu(\mathcal{X}) = 2^{\text{dens}(\mathcal{X})}$
- (b) M is contained in a ball of radius $2r$
- (c) $\rho(x, x') \geq r$ for $x \neq x' \in M$

The assumption $\text{dens}(\mathcal{X}) > 6$ implies $m \geq 65$. Taking $0 < \varepsilon < \frac{1}{8}$, recall the standard VC lower bound argument (Blumer et al., 1989; A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, 1989), which constructs a distribution \mathcal{D} over M by putting a mass of $1 - 8\varepsilon$ on $x_1 \in M$ and distributing the remaining mass uniformly over the other $m - 1$ points. Suppose the target $f : M \rightarrow \{-1, 1\}$ is drawn uniformly at random from $F = \{-1, 1\}^M$, and some learning algorithm is given a sample $S = (X_1, \dots, X_n) \sim \mathcal{D}^n$, where $n = (m - 1)/(32\varepsilon)$.

First, we argue that with probability $\Omega(1)$, S will contain exactly 2 distinct points. Let E_1 be the event that S consists of 1 point; then

$$\mathbb{P}(E_1) = (1 - 8\varepsilon)^n + (m - 1) \left(\frac{8\varepsilon}{m - 1} \right)^n. \quad (7)$$

Since $n = (m - 1)/(32\varepsilon)$ and $m \geq 65$, the first term in the right-hand side of (7) is at most $e^{-(m-1)/32} \leq e^{-2} < 0.1353354$. The second term therein is (i) monotonically decreasing in m (for $m \geq 65$) and (ii) monotonically increasing in ε ; at $m = 65$ and $\varepsilon = 1/8$, we have the upper bound $65/64^{16} < 10^{-27}$. Thus $\mathbb{P}(E_1) < 0.1353355$.

Let E_2 be the event that S has more than 2 points. We decompose E_2 into two cases: E'_2 , where $x_1 \in S$ and E''_2 , where $x_1 \notin S$. Clearly, $\mathbb{P}(E''_2) \leq (8\varepsilon/(m - 1))^3 < 0.000004$. To analyze E'_2 , define $\xi_i = \mathbf{1}_{\{X_i \neq x_1\}}$ and note that under E'_2 , the random quantity $1 + Z := \sum_{i=1}^m \xi_i$ upper-bounds the number of distinct points in S . Hence, $\mathbb{P}(E'_2) \leq \mathbb{P}(Z \geq 2)$. Now $\mathbb{E}[Z] = n(8\varepsilon/(m - 1)) = 1/4$, and by the multiplicative Chernoff's bound,

$$\mathbb{P}(Z \geq 2) = \mathbb{P}(Z \geq 4\mathbb{E}Z) \leq e^{-\mathbb{E}Z} = e^{-1/4}.$$

We conclude that $\mathbb{P}(E_2) < 0.0003 + e^{-1/4} < 0.7789$ and hence $\mathbb{P}(E_1 \cup E_2) < 0.91415$.

Let E_3 be the event that S will consist exactly of 2 distinct points x, x' , and furthermore, that the randomly drawn f has $f(x) \neq f(x')$; then $\mathbb{P}(E_3) \geq \frac{1}{2}(1 - \mathbb{P}(E_1 \cup E_2)) > 0.04$. Conditioned on E_3 , the labels assigned by f to the remaining $m - 2$ points in M are iid Bernoulli(1/2) and their total mass under \mathcal{D} is at least $(m - 2)(8\varepsilon/(m - 1)) > 7.875\varepsilon$, since $m \geq 65$. Since the labels are random, no learning algorithm can do better than flipping a coin, and the algorithm will be correct on $\frac{1}{2}(m - 2)$ of the unseen points, in expectation. The probability of being correct on more than $\frac{3}{4}(m - 2)$ of the points is bounded by $e^{-(m-2)/8} < 0.000381$. Thus, with probability at least $\delta := 0.039$, any learning algorithm that observes n labeled examples sampled from \mathcal{D} will achieve an error of at least $(7.875/4)\varepsilon > 1.95\varepsilon$.

The above argument proves a sample complexity lower bound of $\Omega(m/\varepsilon)$. Inverting the bound for $\varepsilon = \text{err}(h_n)$ implies that with probability at least δ , the following events occur:

- (i) $\text{err}(h_n) > 1.95 \cdot \frac{m-1}{32n} > \frac{0.05m}{n}$
- (ii) the sample S contains exactly two distinct points, x, x'
- (iii) the randomly chosen f has $f(x) \neq f(x')$.

To finish the proof, it remains to establish the requisite relationship between the deterministic quantity $m = |M|$ and the random quantity

$$\mu(\mathcal{X})^{\text{Log}(\text{rad}(S)/\text{marg}(S))} = \lceil \text{rad}(S)/\text{marg}(S) \rceil^{\text{dens}(\mathcal{X})},$$

where $S \sim \mathcal{D}^n$. By construction, $\text{rad}(M) \leq 2r$, and by (ii,iii), $r \leq \text{marg}(S) \leq 2r$, whence $\text{Log}(\text{rad}(S)/\text{marg}(S)) = 1$ and

$$m = \mu(\mathcal{X})^{\text{Log}(\text{rad}(S)/\text{marg}(S))}.$$

and the claimed lower bound follows. ■

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 (Fuglede and Topsøe, 2004) 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): $\max\{\max_{a \in A} l(a, B), \max_{b \in B} l(b, A)\}$. Huttenlocher et al. (1993) define the k -median Hausdorff distance (in the terminology of Jacobs et al. (2000), 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 $v = (l(a_1, B), \dots, l(a_d, B))$, 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 $k \leq m$). 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 12 *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 $D = \text{diam}(X)$ 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 $j \geq i$, 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 $\ell_2(a_{i,j} - a_{j,i}) = \ell_2(a_{i,j} - a_{j,i} - \rho(x_i, x_j)) = \rho(x_i, x_j)$, so $H_1(A_i, A_j)$ is at most this value. On the other hand, for any k, p we can show that $\ell_2(a_{i,k} - a_{j,p}) \geq D$ whenever $k \neq j$ or $p \neq i$: If $i \leq k$, we have $a_{i,k} = 2D((i+1)n + k)$, and otherwise $2D((k+1)n + i) \leq a_{i,k} \leq 2D((k+1)n + i) + D$. Similarly, if $j \leq p$ we have $a_{j,p} = 2D((j+1)n + p)$, and otherwise $2D((p+1)n + j) \leq a_{j,p} \leq 2D((p+1)n + j) + D$. Since by assumption $i \neq j$, 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 13 *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 Talagrand (1992) for $\frac{1}{2} < p < 1$, and Gottlieb and Krauthgamer (2011); Bartal et al. (2011) 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, first note that since h_k is non-decreasing under deletions, we have for all vector sets A, B and subsets $A' \subset A$ and $B' \subset B$ that $H_k(A', B') \geq H_k(A, B)$. Also, we can show there always exist A', B' each of size exactly k that satisfy $H_k(A', B') = H_k(A, B)$: To see this, assume without loss of generality that $h_k(A, B) \geq h_k(B, A)$. Let $h_k(A, B)$ be determined by a set $A_1 \subset A$ of size exactly k , along with a set $B_1 \subset B$ of size $k_1 \leq k$, where the points of B_1 are the nearest neighbors of A_1 . Similarly, let $h_k(B, A)$ be determined by set $B_2 \subset B$ of size exactly k , along with nearest-neighbor set $A_2 \subset A$ of size $k_2 \leq k$. Now add to B' the k_1 vectors of B_1 , and for each $b \in B_1$, add to A' a distinct vector $a \in A_1$ for which b is its nearest neighbor. Further add to A' $\min\{k_2, k - k_1\}$ vectors of A_2 , and add to B' $\min\{k_2, k - k_1\}$ distinct vectors of B_2 that are the nearest neighbors of the relevant vectors in A_2 . If the resulting sets A', B' are of size k then we are done, and have $H_k(A', B') = H_k(A, B)$. Otherwise, we add to A' and B' additional points of A_1 and B_2 , respectively, until both sets are of size k – as A', B' already contain the respective nearest neighbors of B_2, A_1 , the additional points do not affect the distance.

Now consider a set \mathcal{A} of vector sets all within distance 2 of some 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 $n = |\mathcal{A}'| \leq 2^{O(kd)}$, 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 of A'_i 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 at most

k different Euclidean balls, and so there are at most 2^k different sets of k balls (i.e., ball configurations) into which A'_i may fall. Let $C \subset \mathcal{A}'$ ($p = |C|$) include all vertex sets falling into some specific ball configuration. For each pair of vertex sets in C , there must be a pair of vectors at distance at least 2 falling into the same 2-radius Euclidean ball. Then $p = 2^{O(kd)}$ 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 Covertypes 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 representative results below.

| data set | original size | % compressed down to |
|--------------------|---------------|----------------------|
| Covertypes 2 vs. 5 | 2000 | 97 |
| Covertypes 1 vs. 4 | 2000 | 25 |
| Covertypes 4 vs. 7 | 2000 | 2 |

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

Acknowledgments

Thanks to Daniel Hsu for bringing Equation (5) to our attention and Shay Moran for pointing out the analogous result of Shalev-Shwartz and Ben-David (2014). A special thanks to Yevgeni Korsunsky for carefully reading the manuscript and providing very helpful comments.

References

- A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant. A general lower bound on the number of examples needed for learning. *Inf. Comp.*, 82(3):247 – 261, 1989. ISSN 0890-5401. doi: [http://dx.doi.org/10.1016/0890-5401\(89\)90002-3](http://dx.doi.org/10.1016/0890-5401(89)90002-3). URL <http://www.sciencedirect.com/science/article/pii/0890540189900023>.
- Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997. URL citeseer.ist.psu.edu/alon97scalesensitive.html.
- Alexandr Andoni and Robert Krauthgamer. The computational hardness of estimating edit distance. *SIAM J. Comput.*, 39(6):2398–2429, April 2010. doi: 10.1137/080716530.

2. <http://tinyurl.com/cover-data>

- 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, 2008a. URL <http://colt2008.cs.helsinki.fi/papers/86-Balcan.pdf>.
- Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008b. doi: 10.1007/s10994-008-5059-5. URL <http://dx.doi.org/10.1007/s10994-008-5059-5>.
- 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, 2008c. doi: 10.1145/1374376.1374474. URL <http://doi.acm.org/10.1145/1374376.1374474>.
- Reuven Bar-Yehuda and Shimon Even. A linear-time approximation algorithm for the weighted vertex cover problem. *J. Algorithms*, 2(2):198–203, 1981. doi: 10.1016/0196-6774(81)90020-1. URL [http://dx.doi.org/10.1016/0196-6774\(81\)90020-1](http://dx.doi.org/10.1016/0196-6774(81)90020-1).
- Yair Bartal, Ben Recht, and Leonard J. Schulman. Dimensionality reduction: Beyond the johnson-lindenstrauss bound. In *SODA*, 2011.
- R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. In *Physics-Based Modeling in Computer Vision*, 1995. doi: 10.1109/PBMCV.1995.514678.
- Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015. URL <http://dl.acm.org/citation.cfm?id=2886797>.
- 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. ISBN 1-59593-383-2. doi: <http://doi.acm.org/10.1145/1143844.1143857>.
- Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.
- Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. ISSN 1262-3318. doi: 10.1051/ps:2005018. URL <http://dx.doi.org/10.1051/ps:2005018>.
- Dennis K. Burke. Cauchy sequences in semimetric spaces. *Proc. AMS*, 33(1):161–164, 1972. ISSN 00029939. URL <http://www.jstor.org/stable/2038192>.
- I. Cox, M. Miller, S.M. Omohundro, and P.N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *Intl. Conf. Pat. Rec.*, 1996. doi: 10.1109/ICPR.1996.546971.

- Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 208–215, 2008. doi: 10.1145/1390156.1390183. URL <http://doi.acm.org/10.1145/1390156.1390183>.
- 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. ISBN 0-387-94618-7.
- Bend Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory (ISIT)*, June 2004. doi: 10.1109/ISIT.2004.1365067.
- 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. ISSN 0162-8828. doi: 10.1109/34.817410.
- 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. URL <http://papers.nips.cc/paper/5423-generative-adversarial-nets>.
- Lee-Ad Gottlieb and Robert Krauthgamer. A nonlinear approach to dimension reduction. In *SODA*, 2011.
- Lee-Ad Gottlieb and Robert Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Discrete Math.*, 27(4):1759–1769, 2013.
- Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014a. doi: 10.1109/TIT.2014.2339840. URL <http://dx.doi.org/10.1109/TIT.2014.2339840>.
- 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, 2014b. URL <http://papers.nips.cc/paper/5528-near-optimal-sample-compression-for-nearest-neighbors>.
- Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.
- 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.
- D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. PAMI*, 15(9):850–863, 1993. ISSN 0162-8828. doi: 10.1109/34.232073.

- J. Puzicha, J. Buhmann, Y. Rubner, C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, 1999. doi: 10.1109/ICCV.1999.790412.
- 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.
- David W. Jacobs, Daphna Weinshall, and Yoram Gdalyahu. Classification with non-metric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(6):583–600, 2000. doi: 10.1109/34.862197. URL <http://doi.ieeecomputersociety.org/10.1109/34.862197>.
- A. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. PAMI*, 19(12):1386–1391, 1997. ISSN 0162-8828. doi: 10.1109/34.643899. URL <http://dx.doi.org/10.1109/34.643899>.
- Micheal Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1997.
- Aryeh Kontorovich and Iosif Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model. *CoRR*, abs/1606.08920, 2016. URL <http://arxiv.org/abs/1606.08920>.
- 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. URL <http://jmlr.org/proceedings/papers/v32/kontorovichb14.html>.
- Aryeh Kontorovich, Sivan Sabato, and Ruth Uerner. Active nearest-neighbor learning in metric spaces, arxiv:1605.06792. *CoRR*, abs/1605.06792, 2016. URL <http://arxiv.org/abs/1605.06792>.
- Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 791–801, January 2004.
- Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability, unpublished. 1986.
- M. Dubuisson, A. Jain. A modified hausdorff distance for object matching. In *Intl. Conf. Pat. Rec.*, 1994. doi: 10.1109/ICPR.1994.576361.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.
- Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL <http://doi.acm.org/10.1145/2890490>.

- 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. ISBN 0-7695-2228-9. doi: <http://dx.doi.org/10.1109/FOCS.2004.40>.
- Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1 . *SIAM J. Comput.*, 37: 804–826, June 2007. doi: 10.1137/05064206X.
- 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.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- Michel Talagrand. Approximating a helix in finitely many dimensions. *Annales de l’institut Henri Poincaré (B) Probabilités et Statistiques*, 28(3):355–363, 1992.
- A. L. Vol’berg and S. V. Konyagin. On measures with the doubling condition. *Izv. Akad. Nauk SSSR Ser. Mat.*, 51(3):666–675, 1987. ISSN 0373-2436.
- Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.
- L. Wang, C. Yang, and J. Feng. On learning with dissimilarity functions. In *ICML*, 2007. doi: 10.1145/1273496.1273621. URL <http://doi.acm.org/10.1145/1273496.1273621>.
- 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. URL <http://papers.nips.cc/paper/1581-classification-in-non-metric-spaces>.
- 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.
- Wallace Alvin Wilson. On semi-metric spaces. *American Journal of Mathematics*, 53(2): 361–373, 1931.

Appendix A. Inapplicability of metric techniques to semimetrics

Our proof that the density constant controls the size of a net in semimetric spaces (Lemma 1) is non-constructive, and as a result the construction times for nets in semimetrics are necessarily inferior to those of metric spaces. Likewise, estimating a margin of a point set is more challenging for semimetrics and requires a more sophisticated approach than for metric spaces (see Remark 3).

Unlike in metric spaces, where the covering numbers $\mathcal{N}(\cdot)$ and the packing numbers $\mathcal{M}(\cdot)$ are related via $\mathcal{M}(2\varepsilon) \leq \mathcal{N}(\varepsilon) \leq \mathcal{M}(\varepsilon)$ (see e.g., Alon et al. (1997)), violating the triangle inequality breaks this connection between covering and packing. Particularly, for

semimetrics, a doubling constant (while well-defined) does not imply a packing property such as that of Lemma 1. As a consequence, the bounds in the host of results constituting the theory of learning in doubling metric spaces are not applicable to semimetrics. This fact is captured in the following lemma:

Lemma 14 *In semimetric spaces, a bound on the doubling constant does not imply a packing number. In particular, There exist semimetric spaces \mathcal{X} of arbitrary cardinality with a universally bounded doubling constant $\lambda(\mathcal{X}) = O(1)$, such that \mathcal{X} contains a $\text{rad}(\mathcal{X})$ -net C of size $\Theta(|\mathcal{X}|)$.*

Proof Let \mathcal{X} be composed of two sets, A and A' . Put $A = \{a_1, \dots, a_n\}$, endowed with the line metric $\rho(a_i, a_j) = |i - j|$, so the maximum distance in A is $n - 1$. Note that $\lambda(A) = O(1)$. Define A' to consist of n points, such that

$$\rho(a'_i, a'_j) = \rho(a_i, a_j) + \phi \mathbf{1}_{\{i=j\}}, \quad (\phi > 0 \text{ infinitesimal}),$$

while $\rho(a'_i, a'_j) = n - 1$. This defines a semimetric on \mathcal{X} .

Clearly, A' forms a $\text{rad}(\mathcal{X})$ -net of size $|\mathcal{X}|/2$, and yet we can show that $\lambda(\mathcal{X}) = O(1)$. Indeed, consider any ball $B_r(x)$ in \mathcal{X} . Then all points in $B_r(x)$ can be covered by the same $\lambda(A) = O(1)$ balls of radius $\frac{r}{2}$ that cover $A \cap B_r(x)$. The claim follows. \blacksquare

In fact, a deeper principle underlies the results above: In metric spaces, the doubling and density constants are almost the same, while in semimetric spaces there may be a large gap between them. This is captured in the following lemma, which delineates the relationship between the doubling constant and density constant. (The first half of the lemma is due to Gottlieb and Krauthgamer (2013).)

Lemma 15 *Let \mathcal{X} be point set endowed with a metric distance function. Then*

- (a) $\lambda(\mathcal{X}) \leq \mu(\mathcal{X})$,
- (b) $\sqrt{\mu(\mathcal{X})} \leq \lambda(\mathcal{X})$.

Let \mathcal{Y} be a point set endowed with a semimetric distance function. Then

- (c) $\lambda(\mathcal{Y}) \leq \mu(\mathcal{Y})$,
- (d) $\mu(\mathcal{Y})$ may be as large as $\Theta(|\mathcal{Y}|)$, even when $\lambda(\mathcal{Y}) = O(1)$.

Proof To prove (a) and (c), that $\lambda \leq \mu$: Consider any open ball $B_r(x) \in \mathcal{X}$. Let C be a maximal collection of points at mutual interpoint distance at least $\frac{r}{2}$, and note that by definition $|C| \leq \mu(\mathcal{X})$. By the maximality of C , $|C|$ balls of radius $\frac{r}{2}$ centered at points of C cover all of $B_r(x)$, so $\lambda(\mathcal{X}) \leq |C| \leq \mu(\mathcal{X})$. For (b): again, consider any open ball $B_r(x) \in \mathcal{X}$, and let C be a maximal collection of points at mutual interpoint distance at least $\frac{r}{2}$. Now, by definition \mathcal{X} may be covered by $\lambda(\mathcal{X})$ balls of radius $\frac{r}{2}$, and each of these smaller balls may be covered by $\lambda(\mathcal{X})$ balls of radius $\frac{r}{4}$, so there exists a set of $\lambda^2(\mathcal{X})$ balls of radius $\frac{r}{4}$ covering all of X , and in particular C . By the triangle inequality, each ball of radius $\frac{r}{4}$ can cover at most one point of C , and so $|C| \leq \lambda^2(\mathcal{X})$. Finally, (d) follows

immediately from Lemma 14. ■

In metric spaces, the following bounds on exact and approximate nearest neighbor search are well-known.

Lemma 16 *Given a point set S equipped with a metric distance function, and a query point x :*

- (a) *Locating the exact nearest neighbor of x in S requires $\Theta(|S|)$ comparisons in the worst case.*
- (b) *A $(1 + \varepsilon)$ -approximate nearest neighbor of x in S can be found in time $2^{\text{ddim}(S)} \log |S| + \varepsilon^{-O(\text{ddim}(S))}$.*

Proof To prove (a), let S be a set of points obeying the line metric, i.e. the distance between $s_i, s_j \in S$ is $|i - j|$. Suppose x is at distance $n = |S|$ from s_i , and at distance $n + 1$ from all other points of S . Then s_i can be any point of S , and cannot be located without inspecting each point. The claim in (b) is the result of Krauthgamer and Lee (2004). ■

For semimetric spaces, we demonstrate that the situation is much worse:

Lemma 17 *Given a point set S equipped with a semimetric distance function, discovering an exact or approximate nearest neighbor requires $\Theta(|S|)$ comparisons in the worst case.*

Proof For the upper bound, trivially $O(|S|)$ time is sufficient to consider every point in S .

For the lower bound, suppose the query point q is at an infinitesimally small distance from a single point $s_0 \in S$, and at distance $\text{rad}(S)$ from all other points of S . Then s_0 can be any point in S , and cannot be located without inspecting each point: Without the triangle inequality, the distance between one pair of points has no bearing on any other distance. ■