Two Regimes of Generalization for Non-Linear Metric Learning

Anonymous authors
Paper under double-blind review

Abstract

A common approach to metric learning is to seek an embedding of the input data that behaves well with respect to the labels. While generalization bounds for linear embeddings are known, the non-linear case is not well understood. In this work we fill this gap by providing uniform generalization guarantees for the case where the metric is induced by a neural network type embedding of the data. Specifically, we discover and analyze two regimes of behavior of the networks, which are roughly related to the sparsity of the last layer. The bounds corresponding to the first regime are based on the spectral and $(2,1)$-norms of the weight matrices, while the second regime bounds use the $(2,\infty)$-norm at the last layer, and are significantly stronger when the last layer is dense. In addition, we empirically evaluate the behaviour of the bounds for networks trained with SGD on the MNIST and 20newsgroups datasets. In particular, we demonstrate that both regimes occur naturally in realistic data.

1 Introduction

Metric Learning, Bellet et al. (2015), is the problem of finding a metric ρ on the space of features, such that ρ reflects some semantic properties of a given task. Generally, the input can be thought of as a set of labeled pairs $\{((x_i, x_i'), y_i)\}_{i=1}^{n}$, where $x_i, x_i' \in \mathbb{R}^d$ are the features, and y_i is the label, indicating whether x_i and x_i' should be close in the metric or far apart. For instance, in face identification, Schroff et al. (2015), features x_i and x_i' corresponding to the same face should be close in ρ, while different faces should be far apart.

Note that the above metric learning formulation is fairly general and one can convert supervised clustering, or even standard classification problems into metric learning simply by setting $y_i = 1$ if x_i and x_i' have the same original label and $y_i = 0$ otherwise Davis et al. (2007); Weinberger & Saul (2009); Cao et al. (2016); Khosla et al. (2020); Chicco (2020).

The metric ρ is typically assumed to be the Euclidean metric taken after a linear or non-linear *embedding* of the features. That is, we consider a parametric family $\mathcal{G}$ of embeddings into a k-dimensional space, $g : \mathbb{R}^d \to \mathbb{R}^k$, and set [1]

$$\rho(x, x') = \rho_g(x, x') = \|g(x) - g(x')\|_2^2. \tag{1}$$

As an example, Figure 1 shows tSNE plots, Maaten & Hinton (2008), of the classical 20newsgroups dataset. Figure 1a was generated using the regular bag of words representation of the data (see Sections 5 and D for additional details) while Figure 1b was generated by applying tSNE to an embedding of the data (of the form (3) below) learned by minimizing a loss based on labels, as described above. Clearly, there is no discernible relation between the label and the metric in the raw representation, but there is a strong relation in the learned metric. One can also obtain similar conclusions, and quantify them, by, for instance, replacing tSNE with spectral clustering.

The uniform *generalization* problem for metric learning is the following: Given a family of embeddings $\mathcal{G}$, provide bounds, that hold uniformly for all $g \in \mathcal{G}$, on the difference between the expected value of the loss and the average value of the loss on the train set; see Section 2 for a formal definition. Such bounds guarantee, for instance, that the train set would not be overfitted. Given a family

---

[1]Note that ρ in (1) is not strictly a metric. Nevertheless, this terminology is common.

(a) Raw features, after normalization and 500 dim. PCA.
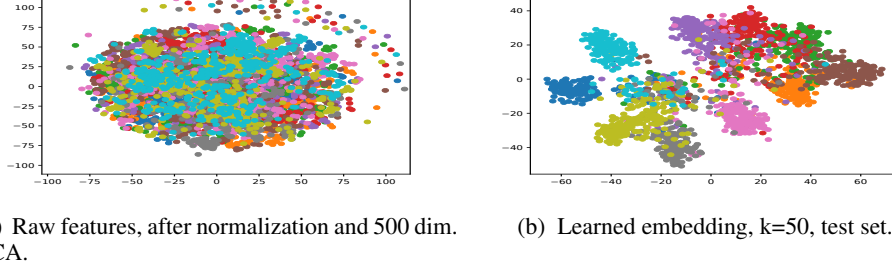
(b) Learned embedding, k=50, test set.

Figure 1: tSNE plot of the 20newsgroup data, restricted to the first 10 labels. Points are colored according to the label.

of embeddings $\mathcal{G}$, consider the family $\mathcal{F}$ of functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of the form

$$\mathcal{F} = \{ f(x, x') = \rho_g(x, x') \mid g \in \mathcal{G} \} . \tag{2}$$

We refer to these as the *distance functions*, which map a pair of features into their distance after the embedding. Well known arguments imply that one can obtain generalization bounds for $\mathcal{F}$ by providing upper bounds on the *Rademacher complexity* $\mathfrak{R}_n(\mathcal{F})$ of the family of the scalar functions $\mathcal{F}$. Therefore in what follows, we discuss directly the upper bounds on $\mathfrak{R}_n(\mathcal{F})$ for various families $\mathcal{G}$. We refer to Section 2 for formal definitions and details on the relation between $\mathcal{G}, \mathcal{F}, \mathfrak{R}_n(\mathcal{F})$, and generalization bounds.

1.1 OVERVIEW OF THE RESULTS

In this paper our goal is to study generalization bounds for neural network type non-linear embeddings $g : \mathbb{R}^d \to \mathbb{R}^k$ given by neural networks of depth $L \geq 1$. Specifically, we consider embeddings of the form

$$g(x) = g_A(x) = \phi_L \left(A_L^t \cdot \left(\phi_{L-1} \left(A_{L-1}^t \ldots \phi_1 \left(A_1^t x \right) \ldots \right) \right) \right) , \tag{3}$$

where $A = (A_1, \ldots, A_L)$ is a tuple of matrices such that $A_i \in \mathbb{R}^{k_{i-1} \times k_i}$ for $i = 1, \ldots, L$, and k_i are the layer widths, with $k_0 = d$ and $k_L = k$. The activations $\phi_i : \mathbb{R} \to \mathbb{R}$ are assumed to be Lipschitz, with $\|\phi\|_{Lip} \leq \rho_i$, and act on vectors in $\mathbb{R}^{k_i}$ coordinatewise. A family of matrix tuples will be generally denoted by $\mathcal{A}$, while the associated embeddings family (3) will be denoted by $\mathcal{G}$ and the associated family of distance functions, (2), will be denoted by $\mathcal{F}$.

In the context of deep learning generalization guarantees for classification, there exists a large body of work on various complexity bounds related to the family of mappings $\mathcal{G}$. The current strongest uniform bounds were given in Bartlett et al. (2017) (see also Neyshabur et al. (2018)). Our goal here is to translate these bounds to bounds on the Rademacher complexity of the family $\mathcal{F}$, derived from $\mathcal{G}$. That is, we study the aspects of the problem that are specific to the metric learning setting.

To state the results, we require a few norm definitions: For matrix $A \in \mathbb{R}^{s \times t}$, $\|A\|_{op}$ is the spectral norm, and set $\|A\|_{2,1} = \sum_{i=1}^{t} \|A_{\cdot i}\|_2$, where $A_{\cdot i}$ is the i-th column of A. For a family $\mathcal{A}$ of matrix tuples, $A = (A_1, \ldots, A_L) \in \mathcal{A}$, for $i \leq L$ we set

$$\|\mathcal{A}_i\|_{op} = \sup_{A \in \mathcal{A}} \|A_i\|_{op} \text{ and } \|\mathcal{A}_i\|_{2,1} = \sup_{A \in \mathcal{A}} \|A_i\|_{2,1} . \tag{4}$$

Thus $\|\mathcal{A}_i\|_{op}, \|\mathcal{A}_i\|_{op}$ are the largest respective norms of the component A_i in $\mathcal{A}$.

With this notation, our first result is the following bound (up to logarithmic factors):

$$\mathfrak{R}_n(\mathcal{F}) \leq \overline{O} \left(\frac{1}{\sqrt{n}} b^2 \left(\prod_{i=1}^{L} \rho_i \|\mathcal{A}_i\|_{op} \right)^2 \left(\sum_{i=1}^{L} \frac{\|\mathcal{A}_i\|_{2,1}^{\frac{2}{3}}}{\|\mathcal{A}_i\|_{op}^{\frac{2}{3}}} \right)^{\frac{3}{2}} \right) . \tag{5}$$

Here b is an upper bound on inputs, $\|x_i\|_2, \|x_i'\|_2 \leq b$ for $i \leq n$. The full statement is given in Section 4, Theorem 1. This bound can be regarded as a natural extension of the bounds of Bartlett

et al. (2017) to the metric learning setting. Indeed, it uses the same parameters – the $\|A\|_{op}$ and $\|A\|_{2,1}$ norms. The quantity $\left(\prod_{i=1}^{L} \rho_i \|\mathcal{A}_i\|_{op} \right)$ is the maximal Lipschitz constant of the mappings in $\mathcal{G}$, and it will typically be the dominant term in (5). One of the appealing properties of (5), inherited from the bounds in Bartlett et al. (2017), is that it is *dimension free*, in the sense that the depth L and the the layer widths k_i do not enter the bound explicitly.

As noted above, for the case of classification, the bounds in Bartlett et al. (2017) are currently the strongest known uniform bounds. However, we now show that in the metric learning setting, one can improve the bound in some situations, by using the $\|A\|_{2,\infty} = \max_{i \leq k} \|A_{\cdot i}\|_2$ norm at the last layer. In Theorem 2, Section 4, we show the following:

$$\mathfrak{R}_n \left(\mathcal{F} \right) \leq \overline{O} \left(\frac{k}{\sqrt{n}} b^2 \left(\rho_L \|\mathcal{A}_L\|_{2,\infty} \prod_{i=1}^{L-1} \rho_i \|\mathcal{A}_i\|_{op} \right)^2 \left(\sum_{i=1}^{L-1} \frac{\|\mathcal{A}_i\|_{2,1}^{\frac{2}{3}}}{\|\mathcal{A}_i\|_{op}^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right). \tag{6}$$

We refer to (5) as the *sparse* bound, and to (6) as the *non-sparse* bound. To compare the bounds, and to see the relation to sparsity, let us first consider the simpler single layer setting, $L = 1$. In this case, the bounds read

$$\mathfrak{R}_n \left(\mathcal{F} \right) \leq O \left(\frac{b^2 \|\phi\|_{Lip}^2 \|\mathcal{A}\|_{op} \|\mathcal{A}\|_{2,1}}{\sqrt{n}} \right), \tag{7}$$

and

$$\mathfrak{R}_n \left(\mathcal{F} \right) \leq O \left(\frac{k b^2 \|\phi\|_{Lip}^2 \|\mathcal{A}\|_{2,\infty}^2}{\sqrt{n}} \right), \tag{8}$$

where $\mathcal{A}$ is the family of weights matrices, $A \in \mathbb{R}^{d \times k}$, ϕ an activation, and $\|\mathcal{A}\|_* = \sup_{A \in \mathcal{A}} \|A\|_*$ where $\|\cdot\|_*$ is one of $\|\cdot\|_{2,1}, \|\cdot\|_{2,\infty}, \|\cdot\|_{op}$. Consider first the case where $\mathcal{A}$ contains non sparse matrices $A \in \mathcal{A}$. These are the weights A where all k neurons have roughly the same norm, or equivalently,

$$\|A\|_{2,1} \sim k \|A\|_{2,\infty}. \tag{9}$$

We refer to this condition as the *dense regime*. Note that (9) holds true for neural networks at initialization, and in Section 5 we observe empirically that this also holds for SGD trained networks on MNIST data. Next, when (9) holds, we also have that

$$\|A\|_{op} \sim \sqrt{k} \|A\|_{2,\infty}. \tag{10}$$

Indeed, one has $\|A\|_{op} \leq \sqrt{k} \|A\|_{2,\infty}$ for any A, and (9) implies $\|A\|_{op} \geq const \cdot \sqrt{\frac{k}{d}} \|A\|_{2,\infty}$ (see Lemma 3 in supplementary material A). The condition (10) will be also experimentally verified for MNIST data. Now, substituting (9),(10) into (7) and (8), we obtain that in the non-sparse regime, (8) is stronger than (7) by a factor of $\sqrt{k}$, a significant improvement. On the other hand, when A is sparse, with most output neurons zeroed out (e.x. $\|A\|_{2,1} = \|A\|_{op} = \|A\|_{2,\infty}$), (7) will be much stronger. Such sparse networks can be achieved by adding explicit sparsity regularization terms to the cost. This will be the case for networks trained on the 20newsgroups dataset.

Finally, as evident from (5) and (6), for the multilayer case $L > 1$, the relation between the sparse and non-sparse bounds is more involved. However, the asymptotics of the dependence on k is similar: With all other k_i fixed, (6) will be better by a factor of $\sqrt{k}$ for non-sparse last layer A_L, and (5) will be better otherwise.

We now discuss an additional direction in which the bounds may be improved: Note that the dependence of the bound on the coefficients of the weights in (5) and (6) is *quadratic*. As we show in Section 4, this is in general unavoidable, due to the quadratic form of the metric (1). However, if the highest activation, ϕ_L is *bounded*, such as for instance the sigmoid $\phi(x) = (1 + e^{-x})^{-1}$, then one can obtain a linearly homogeneous dependence on the layer weights, thus potentially significantly improving the bounds. These bounds are given in (18) and (21) in Theorems 1 and 2 respectively. For comparison, in the single layer case the sparse and non-sparse bounded ϕ_L bounds are given by

$$\mathfrak{R}_n \left(\mathcal{F} \right) \leq O \left(\frac{\sqrt{k} b \|\phi\|_\infty \|\phi\|_{Lip} \|\mathcal{A}\|_{2,1}}{\sqrt{n}} \right) \text{ and } \mathfrak{R}_n \left(\mathcal{F} \right) \leq O \left(\frac{k b \|\phi\|_\infty \|\phi\|_{Lip} \|\mathcal{A}\|_{2,\infty}}{\sqrt{n}} \right), \tag{11}$$

where $\|\phi\|_\infty$ is the upper bound on the values of ϕ.

1.2 METHODS

The general proof strategy we take in this paper is to combine the covering number bounds for the family $\mathcal{G}$, due to Bartlett et al. (2017), Zhang (2002), with the analysis of the specific structure induced by the metric learning problem, i.e. the structure of the family $\mathcal{F}(\mathcal{G})$, when $\mathcal{G}$ is fixed.

Theorems 1 and 2 exploit this structure differently. In Theorem 1 we directly estimate the covering numbers of $\mathcal{F}$ in terms of those of $\mathcal{G}$. Note that, for instance, $\mathcal{F}$ and $\mathcal{G}$ are function families on different domains, and thus such estimates require some care. Once the estimates are obtained, however, we bound $\mathfrak{R}_n(\mathcal{F})$ using a standard Dudley entropy integral argument (Vershynin (2018)).

On the other hand, to prove Theorem 2, we use the special structure of the metric (1) as a *sum*. This allows a decomposition of the problem into k problems, each of which has $k = 1$, and yields bounds which depend on $\|A\|_{2,\infty}$ and k, rather than on $\|A\|_{op}, \|A\|_{2,1}$. While this may seem as a rougher bound at first glance, as discussed above it is in fact much stronger than Theorem 1 in some situations.

Finally, we note that up to now generalization for metric learning was only studied in the linear case (see Section 3 for a detailed discussion). In particular, uniform bounds were derived in Verma & Branson (2015) and in Cao et al. (2016). Interestingly, already in the basic situation where $L = 1$ but $\phi(x) \neq x$ (i.e. a single layer but non-linear activation), one can not deduce our bounds (7) or (8), for instance, from the bounds in Verma & Branson (2015) or Cao et al. (2016). Nor one can use the methods employed in these works to obtain such bounds. The reason for this is that there seems to be no simple principle that would "remove the linearity" (such as the Contraction Lemma, Mohri et al. (2018)) in the case of the cost (1). As a result, somewhat surprisingly, even to obtain the linear instead of quadratic homogenity, as in (11) for $L = 1$, it appears that one already requires the methods used in this paper.

To summarize, we have obtained the first uniform generalization guarantees for multilayer non-linear metric learning. In particular, we introduced two types of bounds, which are appropriate for sparse and non-sparse weights A_L. We have also shown that by using a bounded last layer activation, one may avoid the quadratic dependence of the bound on the parameters, and we have empirically verified that both sparse and dense regimes may occur in SGD optimized networks.

The rest of this paper is organized as follows: In Section 2 we overview the necessary background on metric learning and generalization. Literature and related work are discussed in Section 3. In Section 4 we give the full formal statements of the results, and overview the main ideas involved in the proofs. Full proofs are deferred to the supplementary material due to space constraints. Experiments are described in Section 5 and concluding remarks are given in Section 6.

2 METRIC LEARNING BACKGROUND

In this paper we are assuming that the training data is given as a set $\{((x_i, x_i'), y_i)\}_{i=1}^n$ of labeled feature pairs, which we assume to be sampled independently form some distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}^d \times Y$, where Y is some set of label values. It is usually sufficeint to take $Y = \{0, 1\}$. Note that there may be dependence *within* the pair, i.e. x_i' may depend on x_i.

The quality of the embedding g on a data point $((x, x'), y)$ is measured via a *loss function* ℓ, which usually depends on $((x, x'), y)$ only through the metric. That is, given an embedding $g \in \mathcal{G}$, let $f \in \mathcal{F}$ be the corresponding distance function, $f(x, x') = \rho_g(x, x')$ (see (2)). We assume that there is a fixed function $\ell : \mathbb{R} \times Y \to [0, 1]$, such that the loss of the distance function $f \in \mathcal{F}$ on a data point $((x, x'), y)$ is given by $L_f((x, x'), y) = \ell(f(x_i, x_i'), y_i) = \ell(\rho_g(x_i, x_i'), y_i)$. A typical example of a loss ℓ is the following version of the margin loss:

$$\ell_{S,D}(a, y) = \begin{cases} ReLU(a - S) & \text{if } y = 1 \\ ReLU(D - a) & \text{otherwise,} \end{cases} \tag{12}$$

where $ReLU(x) = \max(0, x)$. As discussed in Section 1, this loss embodies the principle that (x, x') should be close iff $y = 1$, by penalizing distances above S when $y = 1$ and penalizing distances below D otherwise.

The overall loss on the data, or the empirical risk, is given by

$$\hat{R}_f = \frac{1}{n} \sum_{i=1}^{n} L_f((x_i, x_i'), y_i). \tag{13}$$

The expected risk is given by $R_f = \mathbb{E}_{((x,x'),y) \sim \mathcal{D}} L_f((x, x'), y)$. The uniform generalization problem of metric learning is similar to the generalization problem of classification: One is interested in conditions on the family $\mathcal{F}$ under which the gap between expected and empirical risks, $R_f - \hat{R}_f$, is small for all $f \in \mathcal{F}$.

Finally, since $\{((x_i, x_i'), y_i)\}_{i=1}^{n}$ are independent, standard results imply that to control the uniform generalization bounds of the risk, it is sufficient to control the Rademacher complexity of the family $\mathcal{F}$. Specifically, we have that (see Mohri et al. (2018) Theorem 3.3 and Lemma 5.7) ,

$$\sup_{f \in \mathcal{F}} R_f - \hat{R}_f \leq 2 \|\ell\|_{Lip} \mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log \delta^{-1}}{2n}}$$

holds with probability at least $1 - \delta$. Here $\mathfrak{R}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$ and $\|\ell\|_{Lip}$ is the Lipschitz constant of ℓ as a function of its first coordinate. In particular, we have $\|\ell\|_{Lip} = 1$ for losses defined by (12).

3 LITERATURE

General surveys of the field of metric learning can be found in Bellet et al. (2015) and Kulis (2012). See also Chicco (2020) for a survey of recent applications in deep learning contexts. In these situations, the metric learning loss is sometimes referred to as a Siamese network.

Up to now, generalization guarantees in metric learning were only studied in the *linear* setting, i.e. for embeddings of the form (3) where $L = 1$ and $\phi(x) = x$. In particular, all literature cited in this section deals with the linear case.

As discussed in Sections 1 and 2, in this paper we use the formal metric learning setting introduced in Verma & Branson (2015), where we assume that the data comes as a set of iid feature pairs with a label per pair, $((x_i, x_i'), y_i)_{i=1}^{n}$. In this setting, the empirical risk is given by (13). In the special case where the data comes with a label per feature, $(x_i, l_i)_{i=1}^{n}$, one can use an alternative notion of empirical risk, given by

$$\tilde{R}_f = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_f\left((x_i, x_j), \mathbb{1}_{\{l_i \neq l_j\}}\right), \tag{14}$$

which is viewed as a second order U-statistic, see Cao et al. (2016). That is, instead of considering the input as a set of independently sampled pairs, which can be obtained from the $(x_i, l_i)_{i=1}^{n}$ by, for instance, creating pairs out of consecutive samples, in (14) one considers all possible pairs. On one hand, compared to the risk $\hat{R}_f$ defined by (13), for small datasets $\tilde{R}_f$ might make a somewhat better use of the data. On the other hand, $\tilde{R}_f$ is less general, since the data does not necessarily have to be generated by the label-per-feature setting. Indeed, consider the celebrated word2vec text embedding technique, Mikolov et al. (2013b), Mikolov et al. (2013a). In word2vec, tokens x and x' should be mapped to similar vectors if x and x' tend to appear as contexts of each other, and should be mapped to distant vectors otherwise. Equivalently, similar token pairs are extracted from "windows" of the text, while dissimilar pairs are sampled "contrastively", i.e. i.i.d from a (version of) the marginal distribution. Clearly one can model this using the general setting described in Section 2, adopted in this paper. However, note that there are no labels l one can attach to the tokens x such that the distribution $((x, x'), y)$ can be described via the label-per-feature as above, with i.i.d samples (x, l). Moreover, even when the label-per-feature setting is applicable, evaluating (14) may be computationally difficult since the number of terms in the sum is quadratic in dataset size.

The generalization framework considered in this paper is that of the *uniform generalization bounds*, see Mohri et al. (2018). Alternatively, in Wang et al. (2019), Lei et al. (2020), the linear case of metric learning was studied in the framework of algorithmic stability (Bousquet & Elisseeff (2002), Feldman & Vondrak (2019), Bousquet et al. (2020)). In these works, stability, and consequently

generalization bounds, were obtained for an appropriately regularized empirical risk minimization (ERM) procedure. We note that these methods rely strongly on the (uniform) convexity of the underlying problem, and are unlikely to be generalized to non linear settings. In Bellet & Habrard (2015),Christmann & Zhou (2016), linear metric learning was studied in the framework of algorithmic robustness, Xu & Mannor (2012).

Finally, uniform bounds for the linear case were studied in Verma & Branson (2015), and similar results for the version of risk as in (14) were obtained in Cao et al. (2016). In particular it was shown in Verma & Branson (2015) that

$$\Re_n(\mathcal{F}) \leq \frac{b^2 \sup_{(A_1) \in \mathcal{A}} \|A_1 A_1^t\|_F}{\sqrt{n}}, \tag{15}$$

where $\|\cdot\|_F$ is the Frobenius norm. Interestingly, when $\phi(x) = x$, one can derive the single layer inequalities (7) and (8) from (15). Indeed, since $\|\cdot\|_F$ is a *matrix norm*, Bhatia (1997), we have $\|AA^t\|_F \leq \|A\|_{op} \|A^t\|_F = \|A\|_{op} \|A\|_F$ for any A. This in turn can be bounded either as $\|A\|_{op} \|A\|_F \leq \|A\|_{op} \|A\|_{2,1}$ or as $\|A\|_{op} \|A\|_F \leq \sqrt{k} \|A\|_{2,\infty} \cdot \sqrt{k} \|A\|_{2,\infty}$. [2] However, as discussed in Section 1.2, there likely is no way to obtain the general (i.e. non-linear) statements (7) and (8) directly from the purely linear (15).

The bound (15) itself is derived in Verma & Branson (2015) using a relatively short elegant argument involving only the Cauchy Schwartz inequality. However, again, this argument can only be applied when ϕ is the identity. Similarly to the situation in classification, for the non-linear settings other arguments are required.

4 RESULTS

In this Section we introduce the full statements of the main results, Theorems 1 and 2, and overview the main ideas of the proofs. The full proofs are given in supplementary material Sections B and C.

4.1 NOTATION

We begin with some necessary notation. For a vector $v = (v_1, \ldots, v_m) \in \mathbb{R}^m$, the ℓ_p norm is denoted by $\|v\|_p = \left(\sum_{j=1}^m |v_j|^p \right)^{1/p}$. For a matrix $A \in R^{d \times k}$, and $1 \leq p, s \leq \infty$, denote $\|A\|_{p,s} = \left\| \left(\|A_{\cdot 1}\|_p, \ldots, \|A_{\cdot k}\|_p \right) \right\|_s$. That is, one first computes the p-th norm of the columns and then the s-th norm of the vector of these norms. Note that $\|A\|_{2,2} = \|A\|_2$ is the Frobenius norm. Denote by $\|A\|_{op}$ the spectral norm of A. Throughout $\phi : \mathbb{R} \to \mathbb{R}$ will denote an activation/non-linearity, with Lipschitz constant $\|\phi\|_{Lip}$ and such that $\phi(0) = 0$.

The input features $\{(x_i, x_i')\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d$, (see Sections 1,2) may be alternatively organized as two matrices, $X, X' \in \mathbb{R}^{n \times d}$, with rows x_i and x_i' respectively. If $\mathcal{F}$ is a family of functions from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}$, the Rademacher complexity of $\mathcal{F}$ *given the inputs* X, X' is defined by

$$\Re_n(\mathcal{F}) = \Re_n(\mathcal{F}, X, X') = \frac{1}{n} \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(x_i, x_i'), \tag{16}$$

where σ_i are independent Bernoulli variables with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.

Finally, recall that if $\mathcal{A}$ is a family of matrix tuples, $A = (A_1, \ldots, A_L)$, and $\|\cdot\|_*$ some norm on matrices, then for $i \leq L$, $\|\mathcal{A}_i\|_*$ is defined as $\|\mathcal{A}_i\|_* = \sup_{A \in \mathcal{A}} \|A_i\|_*$.

4.2 STATEMENTS AND DISCUSSION

We have the following bound:

---

[2] $\|A\|_{op} \leq \sqrt{k} \|A\|_{2,\infty}$ is shown in supplementary material Section A

Theorem 1. *Let $\mathcal{A} = \{(A_1, \ldots, A_L)\}$ be a family of matrix tuples, $A_i \in \mathbb{R}^{k_{i-1} \times k_i}$, such that $k_0 = d$ and $k_L = k$. Denote $W = \max_{1 \leq i \leq L} k_i$. Let $\mathcal{G}$ be the associated family of embeddings (3), and $\mathcal{F}$ the associated family of distance functions, (2). Let $\{(x_i, x_i')\}_{i=1}^n$ be inputs, such that $\|x_i\|_2, \|x_i'\|_2 \leq b$ for all $i \leq n$. Set $\alpha_i = \|\mathcal{A}_i\|_{op}$, $\beta_i = \|\mathcal{A}_i\|_{2,1}$ and $\rho_i = \|\phi_i\|_{Lip}$ for $i \leq L$. Then*

$$\mathfrak{R}_n(\mathcal{F}) \leq O \left(\frac{1}{n} + \frac{\log(2W) \left[\log n + 2\log\left(4b\left(\prod_{i=1}^L \rho_i \alpha_i\right)\right) \right]}{\sqrt{n}} b^2 \left(\prod_{i=1}^L \rho_i \alpha_i\right)^2 \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^{\frac{3}{2}} \right).$$

(17)

Alternatively,

$$\mathfrak{R}_n(\mathcal{F}) \leq O \left(\frac{1}{n} + \frac{\sqrt{k} \|\phi_L\|_\infty b \log(2W) \left[\log n + 2\log(4\sqrt{k} \|\phi_L\|_\infty) \right]}{\sqrt{n}} \left(\prod_{i=1}^L \rho_i \alpha_i\right) \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^{\frac{3}{2}} \right).$$

(18)

The general structure of this bound is inherited from the covering number bounds of $\mathcal{G}$ in Bartlett et al. (2017). In contrast to the bounds in classification, note that the dependence on the coefficients of A_i and on b is *quadratic*. This can not in general be improved upon. Indeed, consider the special case of RelU activations, $\phi_i(x) = \max\{0, x\}$. Then the family $\mathcal{G}$ is positively homogeneous with respect to the matrix weights: Let $u = (u_1, \ldots, u_L)$ be a tuple of scalars, $u_i \geq 0$, and if $A = (A_1, \ldots, A_L)$ denote by uA the matrix tuple $(u_1 A_1, \ldots, u_L A_L)$. Recall that g_A is the embedding induced by A, given by (3). Then we have

$$g_{uA} = \left(\prod_{i=1}^L u_i\right) g_A.$$

(19)

Since the definition of $\mathcal{F}$ involves the squared norm, it follows from (19) that $f_{uA} = \left(\prod_{i=1}^L u_i\right)^2 f_A$ and thus $\mathfrak{R}_n(\mathcal{F})$ must be quadratically homogeneous in each of u_i.

On the other hand, as discussed in Section 1.1, when ϕ_L is bounded, it is possible to avoid the quadratic dependence. In this case we have the inequality (18) in Theorem 1.

We now briefly sketch the proof of Theorem 1. The argument is based on bounding the covering numbers $\mathcal{N}(\mathcal{F}_{X,X'}, \varepsilon)$ of the set $\mathcal{F}_{X,X'} \subset \mathbb{R}^n$ – the restriction of $\mathcal{F}$ to the input. As mentioned in Section 1.2, given such bounds, we bound $\mathfrak{R}_n(\mathcal{F})$ by the Dudley entropy integral. To bound $\mathcal{N}(\mathcal{F}_{X,X'}, \varepsilon)$, we represent $\mathcal{F}_{X,X'}$ as (a subset of) a Lipschitz image of the Cartesian product $\mathcal{G}_X \times \mathcal{G}_{X'}$, i.e. $\mathcal{G}_X \times \mathcal{G}_{X'} \overset{\Psi}{\longmapsto} \mathcal{F}_{X,X'}$ for an appropriate mapping Ψ. Here $\mathcal{G}_X$ and $\mathcal{G}_{X'}$ are the restrictions of $\mathcal{G}$ to the inputs. The covering numbers of $\mathcal{G}_X$ where estimated in Bartlett et al. (2017), and using these estimates we derive bounds on the coverings of $\mathcal{G}_X \times \mathcal{G}_{X'}$ and consequently of $\mathcal{F}_{X,X'}$. The full details are given in Section B. Here we note that the Lipschitz constant of the mapping $\mathcal{G}_X \times \mathcal{G}_{X'} \mapsto \mathcal{F}_{X,X'}$ can be estimated in two different ways, depending on whether ϕ_L is bounded, which result in the bounds (17) and (18).

We now state our second main result.

Theorem 2. *Let $\mathcal{A} = \{(A_1, \ldots, A_L)\}$ be a family of matrix tuples, $A_i \in \mathbb{R}^{k_{i-1} \times k_i}$, such that $k_0 = d$ and $k_L = k$. Denote $W = \max_{1 \leq i \leq L} k_i$. Let $\mathcal{G}$ be the associated family of embeddings (3), and $\mathcal{F}$ the associated family of distance functions, (2). Let $\{(x_i, x_i')\}_{i=1}^n$ be inputs, such that $\|x_i\|_2, \|x_i'\|_2 \leq b$ for all $i \leq n$. Set $\alpha_i = \|\mathcal{A}_i\|_{op}$, $\beta_i = \|\mathcal{A}_i\|_{2,1}$ and $\rho_i = \|\phi_i\|_{Lip}$ for $i \leq L$. In addition, set $a = \|\mathcal{A}_L\|_{2,\infty}$. Then*

$$\mathfrak{R}_n(\mathcal{F}) \leq O \left(\frac{8bk}{n} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) + \frac{8kb^2 \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right)^2 \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right).$$

(20)

$$\mathfrak{R}_n(\mathcal{F}) \leq O \left(\frac{8 \|\phi_L\|_\infty k}{n} + \frac{8 \|\phi_L\|_\infty kb \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right).$$

(21)

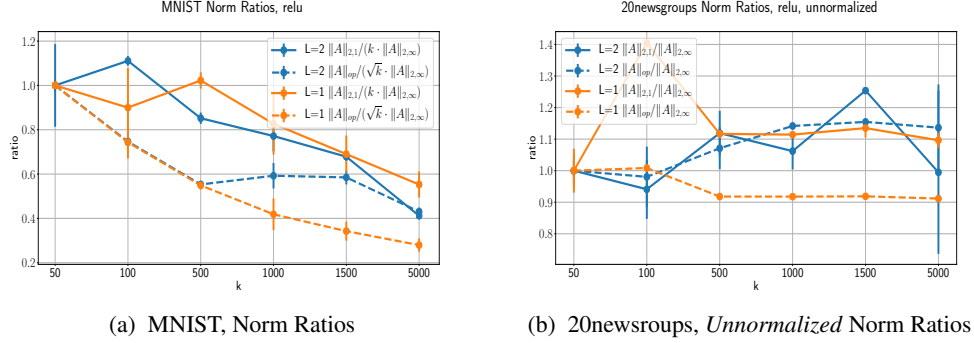(a) MNIST, Norm Ratios

(b) 20newsroups, *Unnormalized* Norm Ratios

Figure 2: Dense and Sparse Regimes Demonstrated on MNIST and 20newsgroups Data

The approach we take here in Theroem 2 is as follows: Note that every element $f \in \mathcal{F}$ can be written as a sum

$$f_A(x, x') = \|g_A(x) - g_A(x')\|_2^2 = \sum_{j=1}^{k} f_{j,A}(x, x'), \qquad (22)$$

where $f_j(x, x') = (g(x)_j - g(x')_j)^2 \in \mathbb{R}$ is the squared difference on the j-th coordinate. For a fixed $j \leq k$, denote by $\mathcal{G}'_j$ the family of the coordinate functions, $g_j : \mathbb{R}^d \to \mathbb{R}$, where $g \in \mathcal{G}$, and by $\mathcal{F}'_j$ the corresponding family of differences $f_j(x, x')$. Note that $\mathcal{G}'_j$ is induced by the same matrix tuples $\mathcal{A}$ as the full $\mathcal{G}$, with the exception of the last layer. Similarly to the arguments in Theorem 1, we can obtain bounds on $\mathfrak{R}_n \left(\mathcal{G}'_j \right)$ from the results in Bartlett et al. (2017). However, then we proceed differently. We first bound $\mathfrak{R}_n \left(\mathcal{F}'_j \right)$ in terms of $\mathfrak{R}_n \left(\mathcal{G}'_j \right)$ (Lemma 9, supplementary material), and then use the decomposition (22) to obtain that $\mathfrak{R}_n \left(\mathcal{F} \right) \leq \sum_{j=1}^{k} \mathfrak{R}_n \left(\mathcal{F}'_j \right)$. The full details are given in supplementary material Section C.

5 EXPERIMENTS

In this section we are interested in the following question: How do the matrices A_L – the weight matrices of the last layer of the network – look for networks trained on standard datasets, with standard training procedures. In particular, do we obtain sparse or dense weights?

Recall that the notion of "dense" in this paper refers the specific condition that

$$\|A\|_{2,1} \geq const \cdot k \cdot \|A\|_{2,\infty} \text{ and } \|A\|_{op} \geq const' \cdot \sqrt{k} \cdot \|A\|_{2,\infty} \qquad (23)$$

holds over a range of values k. That is, we look at several neural networks trained with the metric learning cost (12) on a fixed dataset, where all networks have the same architecture except the size of the output layer k. If the condition (23) holds[3] for a range of values k, we say that the problem is in the *dense regime*. For such a problem, the bounds of Theorem 2 would asymptotically be stronger than the bounds of Theorem 1. If the condition (23) fails, that is, for instance, the ratio $\|A\|_{2,1} / (k \cdot \|A\|_{2,\infty})$ exhibits strong decay with k, then we say that the problem is in the *sparse regime*, and the bounds of Theorem 1 would be better.

We demonstrate that for MNIST data, trained with SGD and no regularization, i.e. the standard deep learning training procedure, the weights are dense to a good approximation. On the other hand, for the 20newsgroup dataset, trained with ℓ_2 regularization, the weights are sparse.

MNIST (mnist Dataset (2005)) is the standard dataset of handwritten digits and the 20newsgroups (20newsgoups Dataset (2010)) consists of newsgroups emails, labeled according to the group. To make computation and illustrations simpler, 20newsgroups is restricted to the first 10 labels.

To perform the optimization, we sample feature/label pairs (x, l), (x', l') independently from the train set, and minimize the loss using SGD on batches of such pairs, until convergence. The precise

---

[3] As discussed in Section 1.1, we have shown that the first condition in (23) implies the second, for large k. Nevertheless, as a sanity check, in what follows we evaluate both conditions.

loss that we use is

$$\ell(x, x', l, l') = 9 \cdot \rho(x, x') \cdot \mathbb{1}_{\{l=l'\}} + ReLU(1 - \rho(x, x')) \cdot \mathbb{1}_{\{l \neq l'\}}, \tag{24}$$

where $\rho(x, x')$ is the distance after the embedding, given by (1). This loss is a version of the loss $\ell_{S,D}(a, y)$ introduced in Section 2, with $S = 0, D = 1$, where the same class case is weighted by 9, to compensate for the fact that $\{l \neq l'\}$ pairs appear 9 time more frequently in the data. Additional details on the setting and the experiments, as well as the implementation source code, are given in supplementary material Section D.

We now describe the results in more detail. Consider first the single layer architecture, $L = 1$, with relu activation. That is, we train single layer networks of size $d \times k$, where d is the problem's original feature dimension, and k varies in the range 50 to 5000. For each trained network, we evaluate the quantities of interest – the ratios $\|A\|_{2,1} / (k \cdot \|A\|_{2,\infty})$ and $\|A\|_{op} / (\sqrt{k} \cdot \|A\|_{2,\infty})$ for MNIST, and unnormalized ratios $\|A\|_{2,1} / \|A\|_{2,\infty}$ and $\|A\|_{op} / \|A\|_{2,\infty}$ for 20newsgroups (see details below). These experiments are represented by orange lines in Figures 2a,2b.

Figure 2a represents the results for the MNIST dataset. In particular, the solid orange line shows the ratio $\|A\|_{2,1} / (k \cdot \|A\|_{2,\infty})$ and the punctuated orange line shows the ratio $\|A\|_{op} / (\sqrt{k} \cdot \|A\|_{2,\infty})$. To simplify the comparison, since we are only interested in change w.r.t k, all curves are normalized to have value 1 at $k = 50$. In such plot, the perfect dense regime behaviour would look as a straight horizontal line with value 1 for all k. While the actual lines exhibit some decay, this decay is slow. Compared to $k = 50$, at $k = 5000$ the orange lines drop by a factor of 2.5, while k grows by a factor of 100. Thus we conclude that in this case the norms are well described by the dense regime.

Next, Figure 2b represents the results for the 20newsgroups dataset. While MNIST is generally nicely behaved, the 20newgroups is not. In particular, it consists of about $n = 7500$ samples in dimension $d = 15000 > n$, and unregularized SGD training would severely overfit the train data. Thus in this case we use the simplest possible regularization term, the ℓ_2 regularization $\lambda \cdot \|A_1\|_{2,2} / k$, with $\lambda = 0.01$ in all experiments.

The orange lines in Figure 2b describe the ratios $\|A\|_{2,1} / \|A\|_{2,\infty}$ (solid) and $\|A\|_{op} / \|A\|_{2,\infty}$ (punctuated). Note that these ratios are *not* normalized by k. The curves are still normalized to be 1 at $k = 50$. Thus, the approximately straight lines we observe in Figure 2b indicate that the above norm ratios practically do not change as k grows, implying that the problem is firmly in the sparse regime.

The blue lines in Figure 2 describe similar experiments for a two layer architecture, $L = 2$. In these experiments, the size of the first layer A_1 was fixed, $d \times 500$, with leaky $relu(0.2)$ activation. The second layer, A_2, was of size $500 \times k$, varying with k. The regualrization term for 2-newsgroups was $\lambda \left(\|A_1\|_{2,2} / 500 + \|A_2\|_{2,2} / k \right)$, again with $\lambda = 0.01$. The ratios in Figure 2 were this case computed for the second layer, A_2. One can see that the conclusions for $L = 2$ are similar: MNIST is in the dense regime, while 20newsgroups is sparse.

Each experiment was repeated 6 times, and every value in Figure 2 is a mean over 6 outcome values. Every value also has an error bar which indicates the magnitude of the standard deviation around the mean. However, in most cases, these bars are small compared to the magnitude of the mean are not visible in the figures.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have obtained the first generalization guarantees for multilayer non-linear metric learning. We have introduced two types of bounds, which are appropriate for sparse and non-sparse regimes of the weights A_L, and we have empirically shown that both regimes may occur in common SGD training settings.

We conclude with an open question: When there are two different bounds on the same quantity, it is natural to ask whether there is a third bound, that interpolates both bounds (i.e. is as strong as both of them) and has a simple form. As discussed in Section 3, for the linear case there indeed is such a bound. It thus would be of interest to understand whether one can obtain a similar strengthening for the non linear multilayer case.

REFERENCES

20newsgoups Dataset. 20newsgoups dataset. https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset, 2010.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *Neuro-computing*, 151:259–267, 2015.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

R. Bhatia. *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York, 1997.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, pp. 73–94, 2020.

Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, 2007.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *COLT*, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Brian Kulis. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper Generalization Bounds for Pairwise Learning. In *Advances in Neural Information Processing Systems*, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.

mnist Dataset. Mnist dataset. https://keras.io/api/datasets/mnist/, 2005.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. The MIT Press, second edition edition, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Nakul Verma and Kristin Branson. Sample complexity of learning mahalanobis distance metrics. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 1 edition, 2018.

Boyu Wang, Hejia Zhang, Peng Liu, Zebang Shen, and Joelle Pineau. Multitask metric learning: Theory and algorithm. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*. PMLR, 2019.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10, 2009.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Two Regimes of Generalization for Non-Linear Metric Learning

Supplementary Material

A $\|A\|_{op}$ BOUNDS

In this Section we prove the lower bound on $\|A\|_{op}$ for dense matrices,(27), that was discussed in Section 1.1. For completeness we prove also the simpler upper bound, (25).

Lemma 3. *(i) For any $A \in \mathbb{R}^{d \times k}$,*

$$\|A\|_{op} \leq \sqrt{k} \, \|A\|_{2,\infty} \, . \tag{25}$$

(ii) For any $A \in \mathbb{R}^{d \times k}$ and $\gamma > 0$ such that

$$\|A\|_{2,1} \geq \gamma k \, \|A\|_{2,\infty} \tag{26}$$

we have

$$\|A\|_{op} \geq \sqrt{\frac{c_\gamma \cdot k}{d}} \, \|A\|_{2,\infty} \, , \tag{27}$$

where $c_\gamma > 0$ is a constant depending solely on γ. In particular, c_γ does not depend on k, d or A.

Proof. We will show the claims for $\|A^t\|_{op}$. Since $\|A^t\|_{op} = \|A\|_{op}$, this will imply the result. Denote by $a_i \in \mathbb{R}^d, i \leq k$, the columns of A. Let $e_1, \dots, e_d$ be the standard basis in $\mathbb{R}^d$. Then

$$\left\|A^t\right\|_{op}^2 = \sup_{\|\theta\|_2 = 1} \left\|A^t \theta\right\|_2^2 = \sup_{\|\theta\|_2 = 1} \sum_{i=1}^k \langle a_i, \theta \rangle^2 \leq \sum_{i=1}^k \|a_i\|_2^2 \leq k \, \|A\|_{2,\infty}^2 \, . \tag{28}$$

This completes the proof of (25).

To prove (27), let us first treat the simple case where $\|A\|_{2,1} = k \, \|A\|_{2,\infty}$, i.e. $\gamma = 1$ (note that for (26) to hold we must have $\gamma \leq 1$). In that case, we have $\|a_i\| = \|A\|_{2,\infty}$ for all $i \leq k$. Using this,

$$k \, \|A\|_{2,\infty}^2 = \sum_{i=1}^k \|a_i\|_2^2 = \sum_{i=1}^k \sum_{j=1}^d \langle e_j, a_i \rangle^2 = \sum_{j=1}^d \sum_{i=1}^k \langle e_j, a_i \rangle^2 = \sum_{j=1}^d \left\|A^t e_j\right\|_2^2 . \tag{29}$$

This in particular implies that there is at least one j such that $\left\|A^t e_j\right\|_2^2 \geq \frac{k}{d} \, \|A\|_{2,\infty}^2$, which is equivalent to (27) with $c_\gamma = 1$.

To prove the general $\gamma \leq 1$ case, we will show that while one does not necessarily have the equality $k \, \|A\|_{2,\infty}^2 = \sum_{i=1}^k \|a_i\|_2^2$, the condition (26) still implies that $\sum_{i=1}^k \|a_i\|_2^2$ is large enough. We have the following:

Claim 4. *There is a constant c_γ such that for any $A \in \mathbb{R}^{d \times k}$ satisfying (26) we have*

$$\sum_{i=1}^k \|a_i\|_2^2 \geq c_\gamma k \, \|A\|_{2,\infty}^2 \, . \tag{30}$$

Clearly, given the claim, the same argument as with (29) yields (27). Thus to complete the proof of the Lemma, it remains to prove the claim.

To prove (30), first note that by homogenity, without loss of generality we can assume $\|A\|_{2,\infty} = 1$. Next, we reformulate (30) using a probabilistic language. Let X be a random variable on the set $\{1, \dots, k\}$, with uniform measure, such that $X(i) = \|a_i\|_2$ for $i \leq k$. Then (26) is equivalent to

$$\mathbb{E}X \geq \gamma \text{ and } \|X\|_\infty = 1, \tag{31}$$

while (30) is equivalent to

$$\mathbb{E}X^2 \geq c_\gamma, \tag{32}$$

for some $c_\gamma > 0$ independent of X.

We now show that (31) implies that there are many "large" values of X. That in turn will imply that $\mathbb{E}X^2$ must also be large. To this end, denote $\alpha_t = \mathbb{P}(X \leq t)$. Then, since $X \leq 1$ we have

$$\mathbb{E}X \leq \alpha_t \cdot t + (1 - \alpha_t) \cdot 1, \tag{33}$$

since t is the maximal value of X on the set $\{X \leq t\}$, and 1 is the maximal value on the complement. Combining this with $\mathbb{E}X \geq \gamma$ we get

$$\gamma \leq t\alpha_t + (1 - \alpha_t). \tag{34}$$

This implies

$$\alpha_t \leq \frac{1 - \gamma}{1 - t}, \tag{35}$$

which is a non-trivial bound for $t \leq \gamma$. To obtain (32) it remains to observe that

$$\mathbb{E}X^2 = 2\int_0^1 t\mathbb{P}(X > t)\,dt = 2\int_0^1 t(1 - \alpha_t)\,dt \geq 2\int_0^\gamma t\left(1 - \frac{1 - \gamma}{1 - t}\right)dt = 2\int_0^\gamma \frac{t(\gamma - t)}{1 - t}\,dt. \tag{36}$$

Indeed, the above expression is (32) with $c_\gamma = 2\int_0^\gamma \frac{t(\gamma - t)}{1 - t}\,dt$. $\qquad\square$

B PROOF OF THEOREM 1

In this Section we prove Theorem 1. Theorem 6 below gives a slightly more general statement, and Corollary 7 is equivalent to Theorem 1.

The Rademacher complexity of a set $\mathcal{V} \subset \mathbb{R}^n$ is defined by

$$\mathfrak{R}_n(\mathcal{V}) = \frac{1}{n}\mathbb{E}_\sigma \sup_{v \in \mathcal{V}} \sum_{i=1}^n \sigma_i v_i \tag{37}$$

where σ_i are independent Bernoulli variables with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. For $\varepsilon > 0$, the covering number of a set $\mathcal{V} \subset \mathbb{R}^n$, denoted $\mathcal{N}(\mathcal{V}, \varepsilon)$ is the minimal size of a set $\mathcal{U} \subset \mathcal{V}$ such that for every $x \in \mathcal{V}$ there is $y \in \mathcal{U}$ s.t. $\|x - y\|_2 \leq \varepsilon$.

If $\mathcal{H}$ is a family of functions $h : \mathcal{X} \to \mathbb{R}$ on a domain $\mathcal{X}$, and $X = \{x_1, \ldots, x_n\}$ is an ordered set of points in the domain, we denote by

$$\mathcal{H}_X = \{(h(x_1,), \ldots, h(x_n)) \mid h \in \mathcal{H}\} \subset \mathbb{R}^n \tag{38}$$

the restriction of $\mathcal{H}$ to X.

Recall that we assume that we are given some fixed input pairs $\{(x_i, x_i')\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d$, and we write $X, X' \in \mathbb{R}^{n \times d}$ for the corresponding matrices with raws x_i and x_i' respectively.

Given a family of matrix tuples $\mathcal{A}$ and the associated embeddings $\mathcal{G}$, we will be interested in the family of vectors

$$\mathcal{F}_{X,X'} = \left\{\left(\|g(x_i) - g(x_i')\|_2^2\right)_{i=1}^n \mid g \in \mathcal{G}\right\} \subset \mathbb{R}^n,$$

the restriction of $\mathcal{F}$ to the input X, X'. Note that the definition of $\mathfrak{R}_n(\mathcal{F}_{X,X'})$ given by (37) coinsides with the definition given in Section 4.1.

The proof of Theorem 6 will follow the general outline given in Section 4.2. The following Lemma will be used to bound the Lipschitz constant of the mapping Ψ discussed in the outline.

For a a matrix $U \in \mathbb{R}^{n \times k}$ denote by U^i, $i \leq n$, the i-th raw of U. Such matrices will be used to represent the set of vectors $g(X)$, where $g \in \mathcal{G}$ is a fixed embedding, acting on raws of X. For two matrices $U, \tilde{U} \in \mathbb{R}^{n \times k}$, the distance $\left\|U - \tilde{U}\right\|_2$ is the standard Euclidean distance where $U, \tilde{U}$ are considered as $n \cdot k$ dimensional vectors. The meaning of the quantities in the Lemma will become clear in the proof of Theorem 6.

Lemma 5 (Metric Header Lipschitzity). *Consider matrices $U, V, \tilde{U}, \tilde{V} \in \mathbb{R}^{n \times k}$, such that*

$$\left\| U - \tilde{U} \right\|_2 \leq \varepsilon, \quad \left\| V - \tilde{V} \right\|_2 \leq \varepsilon. \tag{39}$$

Set $D = \max_{i \leq n} \left\{ \left\| U^i \right\|_2, \left\| \tilde{U}^i \right\|_2, \left\| V^i \right\|_2, \left\| \tilde{V}^i \right\|_2 \right\}$ and define $f_{U,V} \in \mathbb{R}^n$ by

$$f_{U,V}(i) = \left\| U^i - V^i \right\|_2^2 \quad \text{for } i \leq n. \tag{40}$$

Then

$$\left\| f_{U,V} - f_{\tilde{U}, \tilde{V}} \right\|_2 \leq 8 D \varepsilon. \tag{41}$$

Proof. For $i \leq n$,

$$\left(f_{U,V} - f_{\tilde{U}, \tilde{V}} \right)(i) = \left\| U^i - V^i \right\|_2^2 - \left\| \tilde{U}^i - \tilde{V}^i \right\|_2^2 \tag{42}$$

$$= \left\langle U^i - \tilde{U}^i - (V^i - \tilde{V}^i), U_i + \tilde{U}^i - V^i - \tilde{V}^i \right\rangle \tag{43}$$

$$\leq 4D \left(\left\| U^i - \tilde{U}^i \right\|_2 + \left\| V^i - \tilde{V}^i \right\|_2 \right), \tag{44}$$

by the Cauchy-Schwartz and triangle inequalities. Therefore

$$\left\| f_{U,V} - f_{\tilde{U}, \tilde{V}} \right\|_2^2 = \sum_{i=1}^{n} \left(f_{U,V} - f_{\tilde{U}, \tilde{V}} \right)^2 (i) \tag{45}$$

$$\leq (4D)^2 \sum_{i=1}^{n} \left(\left\| U^i - \tilde{U}^i \right\|_2 + \left\| V^i - \tilde{V}^i \right\|_2 \right)^2 \tag{46}$$

$$\leq (4D)^2 \cdot 2 \cdot \sum_{i=1}^{n} \left(\left\| U^i - \tilde{U}^i \right\|_2^2 + \left\| V^i - \tilde{V}^i \right\|_2^2 \right) \tag{47}$$

$$= (4D)^2 \cdot 2 \cdot \left(\left\| U - \tilde{U} \right\|_2^2 + \left\| V - \tilde{V} \right\|_2^2 \right) \tag{48}$$

$$\leq (4D)^2 \cdot 4 \cdot \varepsilon^2. \tag{49}$$

$\square$

Theorem 6. *Let $\mathcal{A}$ be a family of matrix tuples such that for every $A = (A_1, \ldots, A_L) \in \mathcal{A}$ and $i \leq L$,*

$$\|A_i\|_{op} \leq \alpha_i \quad \text{and} \quad \|A_i\|_{2,1} \leq \beta_i. \tag{50}$$

Assume that $\|x_j\|_2, \|x'_j\|_2 \leq b$ for all $j \leq n$. Set $W = \max_{1 \leq i \leq L} k_i$ and

$$D = \max_{i \leq n, g \in \mathcal{G}} \max \left\{ \|g(x_i)\|_2, \|g(x'_i)\|_2 \right\}. \tag{51}$$

Then

$$\mathfrak{R}_n \left(\mathcal{F}_{X, X'} \right) \leq O \left(\frac{1}{n} + \frac{bD \log(2W)(\log n + 2\log(4D))}{\sqrt{n}} \left(\prod_{i=1}^{L} \rho_i \alpha_i \right) \left(\sum_{i=1}^{L} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} \right)^{\frac{3}{2}} \right). \tag{52}$$

Proof. For the input matrices $X, X' \in \mathbb{R}^{n \times k}$, set

$$\mathcal{H}(\mathcal{G}, X) = \left\{ (g(x_i)_j)_{i \leq n, j \leq k} \mid g \in \mathcal{G} \right\} \subset \mathbb{R}^{n \times k} \tag{53}$$

to be the set of images of X under mappings in $\mathcal{G}$, and set

$$\mathcal{H}(\mathcal{G}, X, X') = \left\{ (g(x_i)_1, \ldots, g(x_i)_k, g(x'_i)_1, \ldots, g(x'_i)_k)_{i \leq n} \mid g \in \mathcal{G} \right\} \subset \mathbb{R}^{n \times 2k} \tag{54}$$

to be the set of images of X and X' under mappings in $\mathcal{G}$, concatenated along the second dimension. Note that $\mathcal{F}_{X,X'}$ is by definition the image of $\mathcal{H}(\mathcal{G}, X, X')$ under the map $\hat{\zeta} : \mathbb{R}^{n \times 2k} \to \mathbb{R}^n$ given by

$$\hat{\zeta}\left((u_i, u_i')_{i=1}^n\right) = \left(\|u_i - u_i'\|_2^2\right)_{i=1}^n, \tag{55}$$

where $u_i, u_i' \in \mathbb{R}^k$.

To bound the covering numbers of $\mathcal{F}_{X,X'}$, first note that clearly

$$\mathcal{H}(\mathcal{G}, X, X') \subset \mathcal{H}(\mathcal{G}, X) \times \mathcal{H}(\mathcal{G}, X'). \tag{56}$$

Next, by Theorem 3.3 in Bartlett et al. (2017), we have

$$\log \mathcal{N}(\mathcal{H}(\mathcal{G}, X), \varepsilon) \le \frac{nb^2 \log^2(2W)}{\varepsilon^2} \left(\prod_{i=1}^L \rho_i \alpha_i\right)^2 \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^3. \tag{57}$$

Let $\left\{\tilde{U}\right\}$ and $\left\{\tilde{V}\right\}$ be ε-nets in $\mathcal{H}(\mathcal{G}, X)$ and $\mathcal{H}(\mathcal{G}, X')$ respectively, with sizes bounded by (57). Then by Lemma 5 and by (56), the vectors $\hat{\zeta}\left(\tilde{U}, \tilde{V}\right)$ form an $8D\varepsilon$-net in $\mathcal{F}_{X,X'} = \hat{\zeta}\left(\mathcal{H}(\mathcal{G}, X, X')\right)$. Equivalently,

$$\log \mathcal{N}(\mathcal{F}_{X,X'}, 8D\varepsilon) \le 2\frac{nb^2 \log^2(2W)}{\varepsilon^2} \left(\prod_{i=1}^L \rho_i \alpha_i\right)^2 \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^3, \tag{58}$$

which can also be rewritten as

$$\log \mathcal{N}(\mathcal{F}_{X,X'}, \varepsilon) \le 128\frac{D^2 nb^2 \log^2(2W)}{\varepsilon^2} \left(\prod_{i=1}^L \rho_i \alpha_i\right)^2 \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^3. \tag{59}$$

Denote

$$R = 12Db \log(2W) \left(\prod_{i=1}^L \rho_i \alpha_i\right) \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^{\frac{3}{2}}. \tag{60}$$

Set $D' = \sup_{v \in \mathcal{F}_{X,X'}} \|v\|_2 \le 4D^2 \sqrt{n}$. By Dudley's entropy bound, Vershynin (2018), and using (59), for any $\delta > 0$,

$$\begin{aligned}
\Re(\mathcal{F}(\mathcal{G})) &\le \frac{4\delta}{\sqrt{n}} + \frac{12}{n} \int_\delta^{D'} \sqrt{\log \mathcal{N}(\mathcal{F}, \varepsilon)} d\varepsilon \\
&\le \frac{4\delta}{\sqrt{n}} + \frac{144R}{\sqrt{n}} \log\left(4D^2 \sqrt{n}/\delta\right).
\end{aligned}$$

Taking $\delta = \frac{1}{\sqrt{n}}$ completes the proof.

$\square$

We now instantiate the bounds of Theorem 6 with bounds on the maximal norm D. The first bound uses the Lipschitzity of the full map g, while the second holds under the assumption that the output non-linearity ϕ_L is bounded.

Corollary 7. *Under the assumptions of Theorem 6, we have*

$$\Re_n(\mathcal{F}_{X,X'}) \le O\left(\frac{1}{n} + \frac{\log(2W)\left[\log n + 2\log(4b\left(\prod_{i=1}^L \rho_i \alpha_i\right))\right]}{\sqrt{n}} b^2 \left(\prod_{i=1}^L \rho_i \alpha_i\right)^2 \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^{\frac{3}{2}}\right). \tag{61}$$

Alternatively,

$$\Re(\mathcal{F}_{X,X'}) \le O\left(\frac{1}{n} + \frac{\sqrt{k}\|\phi_L\|_\infty b \log(2W)\left[\log n + 2\log(4\sqrt{k}\|\phi_L\|_\infty)\right]}{\sqrt{n}} \left(\prod_{i=1}^L \rho_i \alpha_i\right) \left(\sum_{i=1}^L \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}}\right)^{\frac{3}{2}}\right). \tag{62}$$

Proof. Due to the form (3), the Lipschitz norm of any $g\mathcal{G}$ as a function from $\mathbb{R}^d$ to $\mathbb{R}^k$ satisfies $\|g\|_{Lip} \le \left(\prod_{i=1}^{L} \rho_i \alpha_i\right)$. This in turn implies $D \le b\|g\|_{Lip} \le b\left(\prod_{i=1}^{L} \rho_i \alpha_i\right)$, which yields (61). To obtain (62), we observe that one also has $D \le \sqrt{k}\|\phi_L\|_\infty$. $\qquad\square$

C   PROOF OF THEOREM 2

In this Section we use the notation from Section B, and from Theorem 6. Similarly to Theorem 6, in Theorem 10 we prove a statement that uses a general diameter (but not the same diameter as in Theorem 6), and we instantiate the different bounds in Corollary 11, which is equivalent to Theorem 2.

We first obtain bounds for one dimensional embeddings. Let $\mathcal{G}'$ be a family of embeddings induced by tuples $\mathcal{A}'$ such that $k = k_L = 1$. $\mathcal{G}'$ may be equivalently described as a family of functions, and we can bound $\Re_n(\mathcal{G}'_X)$. Note that for such families $\|A_L\|_{2,1} = \|A_L\|_{op} = \|A_L\|_2$.

Claim 8. *Let $\mathcal{A}'$ be a family of matrix tuples such that $k_L = 1$ and such that for every $A = (A_1, \dots, A_L) \in \mathcal{A}$ and $i \le L$,*

$$\|A_i\|_{op} \le \alpha_i \quad and \quad \|A_i\|_{2,1} \le \beta_i. \tag{63}$$

Assume that $\|x_j\|_2 \le b$ for all $j \le n$ and set $W = \max_{1 \le i \le L} k_i$. Then

$$\Re(\mathcal{G}'_X) \le O\left(\frac{1}{n} + \frac{b\log(2W)\log n}{\sqrt{n}}\left(\prod_{i=1}^{L} \rho_i \alpha_i\right)\left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1\right)^{\frac{3}{2}}\right). \tag{64}$$

Proof. Note that in the $k_L = 1$ case the restrictions $\mathcal{G}'_X$ and $\mathcal{H}(\mathcal{G}', X)$ defined in (53) coinside. The bound on the covering numbers $\mathcal{N}(\mathcal{H}(\mathcal{G}', X), \varepsilon)$ was given in (57) in the proof of Theorem 6. The bound (64) now follows by combining the Dudley entropy integral inequality with the covering number bound, similarly to the argument in Theorem 6. $\qquad\square$

Let $\mathcal{H}$ be a family of functions $h : \mathcal{X} \to \mathbb{R}$ on some domain $\mathcal{X}$. Then we can construct a family $\mathcal{H}_*$ of functions on $\mathcal{X} \times \mathcal{X}$ as follows:

$$\mathcal{H}_* = \left\{h_*(x, x') = (h(x) - h(x'))^2 \ \mid\ h \in \mathcal{H}\right\}. \tag{65}$$

In the following Lemma we obtain a bound on the Rademacher complexity of $\mathcal{H}_{*,X,X}$, the restriction of $\mathcal{H}_*$ to X, X', in terms of the complexities of $\mathcal{H}_X, \mathcal{H}_{X'}$.

Lemma 9. *Given a family H, assume that there is $b_\infty > 0$ s.t. $\|h\|_\infty \le b_\infty$ for all $h \in \mathcal{H}$. Then*

$$\Re_n(\mathcal{H}_{*,X,X'}) \le 4b_\infty(\Re(\mathcal{H}_X) + \Re(\mathcal{H}_{X'})). \tag{66}$$

Proof. The function $\gamma(x) = x^2$ is $2a$-Lipschitz when $x \in [-a, a]$. By the boundedness assumptions, we have $|h(x) - h(x')| \le 2b_\infty$. Therefore by the Contraction Lemma for Rademacher complexities, Mohri et al. (2018), we have

$$\Re_n(\mathcal{H}_{*,X,X'}) = \frac{1}{n}\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (h(x_i) - h(x_i'))^2 \tag{67}$$

$$\le \frac{4b_\infty}{n}\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i (h(x_i) - h(x_i')) \tag{68}$$

$$\le \frac{4b_\infty}{n}\left(\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(x_i) + \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(x_i')\right). \tag{69}$$

$\qquad\square$

For any set of families of functions $\{\mathcal{H}_i\}_{i \leq m}$ on a common domain $\mathcal{X}$, define the direct sum $\oplus_{i=1}^m \mathcal{H}_i$ as the family

$$\oplus_{i=1}^m \mathcal{H}_i = \left\{ g = \sum_{i=1}^m h_i \mid h_i \in \mathcal{H}_i \right\}. \tag{70}$$

For a scalar a set also $a\mathcal{H} = \{ah \mid h \in \mathcal{H}\}$.

It is well known, Bartlett & Mendelson (2002), that the Rademacher complexity is additive with respect to the direct sum. Indeed,

$$\mathfrak{R}_n \left(\left(\oplus_{j=1}^m \mathcal{H}_j \right)_X \right) = \frac{1}{n} \mathbb{E}_\sigma \sup_{g \in \oplus_{j=1}^m \mathcal{H}_j} \sum_{i=1}^n \sigma_i g(x_i) \tag{71}$$

$$= \frac{1}{n} \mathbb{E}_\sigma \sup_{h_j \in \mathcal{H}_j, j \leq m} \sum_{i=1}^n \sigma_i \left(\sum_{j=1}^m h_j(x_i) \right) \tag{72}$$

$$= \frac{1}{n} \mathbb{E}_\sigma \sup_{h_j \in \mathcal{H}_j, j \leq m} \sum_{j=1}^m \left(\sum_{i=1}^n \sigma_i h_j(x_i) \right) \tag{73}$$

$$= \frac{1}{n} \mathbb{E}_\sigma \sum_{j=1}^m \sup_{h_j \in \mathcal{H}_j} \sum_{i=1}^n \sigma_i h_j(x_i) \tag{74}$$

$$= \sum_{j=1}^m \mathfrak{R}\left(H_{j,X} \right). \tag{75}$$

Theorem 10. *Let $\mathcal{A}$ be a family of matrix tuples such that for every $A = (A_1, \ldots, A_L) \in \mathcal{A}$, with $k_L = k$. For $i \leq L - 1$, let $\alpha_i, \beta_i > 0$ be such that*

$$\|A_i\|_{op} \leq \alpha_i \quad and \quad \|A_i\|_{2,1} \leq \beta_i, \tag{76}$$

and let $a > 0$ be such that

$$\|A_L\|_{2,\infty} \leq a. \tag{77}$$

Assume that $\|x_j\|_2, \|x'_j\|_2 \leq b$ for all $j \leq n$. Set $W = \max_{1 \leq i \leq L-1} k_i$ and

$$D = \max_{i \leq n, g \in \mathcal{G}, j \leq k} \max \left\{ |g(x_i)_j|, |g(x'_i)_j| \right\}, \tag{78}$$

where $g(x)_j$ is the j-th coordinate of $g(x) \in \mathbb{R}^k$. Then

$$\mathfrak{R}(\mathcal{F}_{X,X'}) \leq O \left(\frac{8Dk}{n} + \frac{8Dkb \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right). \tag{79}$$

Proof. Given the family $\mathcal{A}$, construct the family $\mathcal{A}'$ via:

$$\mathcal{A}' = \left\{ (A_1, \ldots, A_{L-1}, Z) \mid A = (A_1, \ldots, A_{L-1}, A_L) \in \mathcal{A}, Z \in \mathbb{R}^{1 \times k_{L-1}}, \|Z\|_2 \leq a \right\}. \tag{80}$$

That is, $\mathcal{A}'$ consists of tuples $A' = (A'_1, \ldots, A'_L)$ such that $A'_1, \ldots, A'_{L-1}$ belong to some tuple $A \in \mathcal{A}$, and A'_L can be an arbitrary vector, as long as $\|A'_L\|_2 \leq a$. Note in particular that $\mathcal{A}'$ induces one-dimensional embeddings (i.e. $k_L(\mathcal{A}') = 1$), and we denote the embeddings induced by $\mathcal{A}'$ by $\mathcal{G}'$.

For any $j \leq k$, denote by $\mathcal{G}_j$ the coordinate restriction of $\mathcal{G}$:

$$\mathcal{G}_j = \{ g(x)_j \mid g \in \mathcal{G} \}. \tag{81}$$

That is, $\mathcal{G}_j$ is the family of functions $\mathbb{R}^d \mapsto \mathbb{R}$ obtained by looking at the j-th coordinate of the embeddings in $\mathcal{G}$.

Note that due to the condition (77), we have $\mathcal{G}_j \subset \mathcal{G}'$ for any $j \leq k$.

Next, recall that by definition we have

$$\mathcal{F} = \left\{ f(x, x') = \|g(x) - g(x')\|_2^2 \ \mid \ g \in \mathcal{G} \right\} = \left\{ \sum_{j=1}^{k} (g(x)_j - g(x')_j)_2^2 \ \mid \ g \in \mathcal{G} \right\}. \quad (82)$$

This implies that

$$\mathcal{F} \subset \oplus_{j=1}^{k} (\mathcal{G}_j)_*, \quad (83)$$

where the "product-squared-difference" operation $\mathcal{H}_*$ is as defined in (65).

Now, using the above inclusions, the Rademacher additivity (75), Lemma 9, and Claim 8, we obtain

$$\mathfrak{R}_n (\mathcal{F}_{X,X'}) \leq \mathfrak{R}_n \left(\left(\oplus_{j=1}^{k} (\mathcal{G}_j)_*\right)_{X,X'} \right) \quad (84)$$

$$\leq \mathfrak{R}_n \left(\left(\oplus_{j=1}^{k} (\mathcal{G}')_*\right)_{X,X'} \right) \quad (85)$$

$$= k \mathfrak{R}_n \left((\mathcal{G}'_*)_{X,X'} \right) \quad (86)$$

$$\leq 4Dk \left(\mathfrak{R}_n (\mathcal{G}'_X) + \mathfrak{R}_n (\mathcal{G}'_{X'}) \right) \quad (87)$$

$$\leq O \left(\frac{8Dk}{n} + \frac{8Dkb \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right). \quad (88)$$

$$\square$$

Corollary 11. *Under the assumptions of Theorem 10, we have*

$$\mathfrak{R}_n (\mathcal{F}_{X,X'}) \leq O \left(\frac{8bk}{n} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) + \frac{8kb^2 \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right)^2 \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right), \quad (89)$$

and

$$\mathfrak{R}_n (\mathcal{F}_{X,X'}) \leq O \left(\frac{8 \|\phi_L\|_\infty k}{n} + \frac{8 \|\phi_L\|_\infty kb \log(2W) \log n}{\sqrt{n}} \left(\rho_L a \prod_{i=1}^{L-1} \rho_i \alpha_i \right) \left(\sum_{i=1}^{L-1} \frac{\beta_i^{\frac{2}{3}}}{\alpha_i^{\frac{2}{3}}} + 1 \right)^{\frac{3}{2}} \right). \quad (90)$$

Proof. The bounds follow from (79) by using either $D \leq b \sup_{g \in \mathcal{G}} \|g\|_{Lip}$ similarly to the argument in Corollary 7, or by using $D \leq \|\phi_L\|_\infty$. $\square$

D ADDITIONAL DETAILS ON EXPERIMENTS

- The full code of the experiments is provided as a supplementary material. The code is based on the Tensorflow framework. Details on the invocation can be found in the README.md file.

- To perform the optimization, we sample feature/label pairs (x, l), (x', l') independently from the train set, and minimize the loss using SGD on batches of such pairs, until convergence. The precise loss that we use is

$$\ell(x, x', l, l') = 9 \cdot \rho(x, x') \cdot \mathbb{1}_{\{l=l'\}} + ReLU(1 - \rho(x, x')) \cdot \mathbb{1}_{\{l \neq l'\}}, \quad (91)$$

  where $\rho(x, x')$ is the distance after the embedding, given by (1). This loss is a version of the loss $\ell_{S,D}(a, y)$ introduced in Section 2, with $S = 0, D = 1$, where the same class case is weighted by 9. Our datasets have 10 roughly balanced labels, and therefore the probability of the event $l = l'$ is about $1/10$. Thus the multiplier 9 helps in giving similar weights to the separation $l \neq l'$ and the compression $l = l'$ conditions, enforced by the respective terms in (91).

- The precise process of sampling the pairs (x, x') was as follows: We fix two batch sizes Bx, By, typically $Bx = By = 500$. We take a batch S_1 of size Bx and another, independent batch, S_2, of size B_y. Then the set S of all feature pairs (x, x') such that $x \in S_1$ and $x' \in S_2$ is created, with corresponding labels y as described in Section 1. This S is fed to the optimizer as a single batch.

- An *epoch* is when first batch iterator, S_1, finishes a single iteration over the whole data. The data is permuted at the end of the epoch.

- All experiments were run for 180 epochs. Convergence typically occurred much earlier, around 30 to 50 epochs, depending on the value of k.

- All experiments were run on a GTX 1080 GPU, with Tensorflow 2.0. A single run of the longest experiment, MNIST with $k = 5000$, took under 150 minutes to complete. All experiments combined, with repetitions, finish running in under 48 hours.

E LOWER BOUNDS

For $r > 0$, let

$$\mathcal{G}_r = \left\{ g_A(x) = \phi \left(A_L^t \cdot \left(\phi \left(A_{L-1}^t \ldots \phi \left(A_1^t x \right) \ldots \right) \right) \right) \ \mid \ \prod_{i=1}^L \|A_i\|_{op} \leq r \right\}, \quad (92)$$

be the family of neural net embeddings $\mathbb{R}^d \to \mathbb{R}^k$, satisfying $\prod_{i=1}^L \|A_i\|_{op} \leq r$. In what follows we assume throughout that $\phi(x) = \max(0, x)$ is the ReLU activation. The layer widths are given by $k_i, i = 0, \ldots, L$, with $k_0 = d$ and $k_L = k$. We have $A_i \in \mathbb{R}^{k_{i-1} \times k_i}$.

Recall that we denote by

$$\mathcal{F}(\mathcal{G}) = \left\{ f(x, x') = \|g(x) - g(x')\|^2 \ \mid \ g \in \mathcal{G} \right\} \quad (93)$$

the family of distance functions induced by the family of embeddings $\mathcal{G}$. Set $\mathcal{F}_r = \mathcal{F}(\mathcal{G}_r)$.

Recall also that the Rademacher complexity $\mathfrak{R}_n(\mathcal{F}, X, X')$ of a family of distance functions $\mathcal{F}$ on inputs $X, X' \in \mathbb{R}^{n \times d}$ was defined in (16).

We assume that the layer widths satisfy

$$k_i \geq 2 \text{ for } i = 1, \ldots, L. \quad (94)$$

Theorem 12. *There is a function of the features $J(X, X')$, such that for every $r > 0$ and features $X, X' \in \mathbb{R}^{n \times d}$,*

$$\mathfrak{R}_n(\mathcal{F}_r, X, X') \geq r^2 \cdot J(X, X'). \quad (95)$$

The function J does not depend on L or the layer widths $k_1, \ldots, k_L$, and satisfies $J(X, X') \geq 0$ for all X, X', with $J(X, X') = 0$ iff $X = X'$.

Proof. As discussed in Section 4.2, due to the positive homogeneity of ReLU activations and quadratic homogeneity of the cost (see eq. (19) and related discussion), it is sufficient to prove (95) for $r = 1$. That is, we shall prove that

$$\mathfrak{R}_n(\mathcal{F}_1, X, X') \geq J(X, X'). \quad (96)$$

Note that this statement is non-trivial since the left handside may depend on the layer widths, while the right handside does not. That is, the lower bound $J(X, X')$ holds for any choice of widths, as long as $\prod_{i=1}^L \|A_i\|_{op} \leq 1$ is satisfied.

To show (96), denote by

$$\mathcal{G}_1' = \left\{ g_A(x) = A_L^t \cdot \left(\phi \left(A_{L-1}^t \ldots \phi \left(A_1^t x \right) \ldots \right) \right) \ \mid \ \prod_{i=1}^L \|A_i\|_{op} \leq 1 \right\} \quad (97)$$

the family of embeddings without the top activation. Also set

$$\mathcal{H}_1' = \left\{ g_a(x) = \langle x, a \rangle e_1 \ \mid \ g_a(x) : \mathbb{R}^d \to \mathbb{R}^k, a \in \mathbb{R}^d, \|a\| \leq 1 \right\} \quad (98)$$

and

$$\mathcal{H}_1 = \left\{ g_a(x) : \mathbb{R}^d \to \mathbb{R}^k, g_a(x) = \langle x, a \rangle \, e_1 \ \mid \ a \in \mathbb{R}^d, \|a\| \leq 1 \right\}. \tag{99}$$

That is, $\mathcal{H}_1'$ is the set of *linear* mappings $\mathbb{R}^d \to \mathbb{R}$, viewed as mappings $\mathbb{R}^d \to \mathbb{R}^k$ where only the first column is nonzero and of norm at most 1. $\mathcal{H}_1$ is obtained by composing $\mathcal{H}_1'$ with ϕ.

As shown in Bartlett et al. (2017), for any set of widths k_i satisfying (94), we have $\frac{1}{2} \cdot \mathcal{H}_1' \subset \mathcal{G}_1'$.

Consequently we have $\frac{1}{2} \cdot \mathcal{H}_1 \subset \mathcal{G}_1$, which implies $\frac{1}{2} \cdot \mathcal{F}(\mathcal{H}_1) \subset \mathcal{F}(\mathcal{G}_1)$, which in turn implies

$$\mathfrak{R}_n \left(\mathcal{F}(\mathcal{G}_1), X, X' \right) \geq \frac{1}{2} \mathfrak{R} \left(\mathcal{F}(\mathcal{H}_1), X, X' \right) \tag{100}$$

by the monotinicity of Rademacher complexity w.r.t set inclusion. Since H_1, and thus $\mathfrak{R}(\mathcal{F}(\mathcal{H}_1), X, X')$ does not depend on $k_1, \ldots, k_L$, setting

$$J(X, X') = \frac{1}{2} \mathfrak{R} \left(\mathcal{F}(\mathcal{H}_1), X, X' \right) \tag{101}$$

completes the proof. $\qquad\square$