Compressive Learning of Multi-layer Perceptrons: An Error Analysis

Ata Kabán
School of Computer Science
University of Birmingham
B15 2TT, Birmingham, UK
A.Kaban@cs.bham.ac.uk

Abstract—We consider the class of 2-layer feed-forward neural networks with sigmoidal activations – one of the oldest black-box learning machines – and ask the question: Under what conditions can it successfully learn from a random linear projection of the data? Part of this question has been previously attempted in the literature: A high probability bound has been given on the absolute difference between the outputs of the network on the sample before and after random projection – provided that the size of the hidden layer, and $N$ is the number of training points. By contrast, in this paper we prove that a lower target dimension independent of both $N$ and $M$ suffices, not only to guarantee low distortion of the outputs but also to ensure good generalisation for learning the network on randomly projected data. We do not require a sparse representation of the data, instead our target dimension bound depends on the regularity of the problem expressed as norms of the weights. These are uncovered in our analysis by the use of random projection, which fulfils a regularisation role on the input layer weights.

Index Terms—Error analysis, Random projection, Multi-layer perceptron

I. INTRODUCTION

The intriguing prospect to estimate a – possibly nonlinear – predictive model from low dimensional random linear projections of high dimensional data sets instead of the original data has attracted much research interest in the past decade [7]. Due to the speedy increase of dimensionality of modern data sets, and the development of novel data acquisition techniques in the area of compressed sensing, this approach appears to have potential towards conquering the curse of dimensionality, and is also a door-opener for privacy-preserving data processing [6].

In fact, the idea to learn from randomly projected data dates back to seminal works in theoretical computer science, and theories of learning [3], [9]. A more recent and practical motivation is the prospect of making use of novel data acquisition techniques from compressed sensing, such as CCD and CMOS cameras [24]. These devices bypass the need to store and process large data sets and instead collect a random linear projection of the data directly.

Here we consider the class of feed-forward neural networks with a hidden layer and Lipschitz continuous activation functions. We seek conditions required for it to solve learning problems to good-enough approximation from random projections of the data. This class of networks is a classic and well-weathered workhorse in the predictive data analytics practitioners’ toolbox, and precursor of a variety of recent ‘deep network’ extensions. The same type of network was previously analysed under random projections in [24], but our findings (and approaches) differ.

Let $X \subset \mathbb{R}^d$ be an input domain. We denote by $H = \{x \rightarrow h(x) : x \in X\}$ the function class that implements neural networks of the following parametric form:

$$h(x) = u + \sum_{i=1}^{M} v_i \sigma(w_i^T x) \quad (1)$$

where $\sigma : \mathbb{R}^d \rightarrow [-b, b]$ is a Lipschitz continuous bounded activation function – traditionally a sigmoidal function, such as $\sigma(u) = \tanh(u)$, or the logistic function $\sigma(u) = \frac{1}{1+e^{-u}}$. Further, $w_i \in \mathbb{R}^d$, $u, v_i \in \mathbb{R}$ are the weights or parameters of the network. We assume that $\|v\|_1 = C_v$ for some constant $C_v > 0$, but do not constrain the first layer weights a-priori.

In practice, these parameters are estimated from a finite set of labelled training points denoted by $T_N = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \overset{i.i.d}{\sim} \mathcal{D}$, and $\mathcal{D}$ is an unknown distribution over $X \times Y$ with $X \subset \mathbb{R}^d$ and $Y = [-b, b] \subset \mathbb{R}$. Let $\ell : \mathbb{R} \times \mathbb{R} \rightarrow [0, \bar{\ell}]$ be a bounded loss function, assumed to be $L_1$-Lipschitz in its first argument. It measures the mismatch between the true and the predicted labels of a labelled point. Typically the loss function depends on its arguments only through their product $y \cdot h(x)$ (for classification), or through their difference, $y - h(x)$ (for regression).

It is well known that this class of neural networks is capable of approximating any smooth target function arbitrarily well when provided with a sufficient number of hidden units [8] – that is, when the size of the network, $M$, is large enough. It is also known that, for good generalisation, the size of the weights matters more than the size of the network [4].

The quantity of ultimate interest that quantifies the success of learning is the generalisation error (or risk). For an $h \in H$, this is defined as $E[\ell \circ h] := E_{(x,y) \overset{i.i.d}{\sim} \mathcal{D}}[\ell(h(x), y)]$. However, since $\mathcal{D}$ is unknown, we only have access to the empirical error, defined as $\hat{E}_{T_N}[\ell \circ h] = \frac{1}{N} \sum_{n=1}^{N} \ell(h(x_n), y_n)$. The optimal learner within $H$ will be denoted as $h^* = \arg\min_{h \in H} E[\ell \circ h]$. This work is funded by EPSRC Fellowship EP/P004245/1, “FORGING: Fortuitous Geometries and Compressed Learning”.
arg inf $E[h \circ h]$. The sample error minimiser of the loss is
\[ h = \arg \min_{h \in H} E_{\mathcal{T}^N}[\ell \circ h]. \]

We are interested in the setting where $d$ is too large to work with directly, as indeed this is the case in many modern data sets. We shall employ random projection (RP) to reduce dimension before feeding the data to the neural network. RPs represent a data-oblivious universal dimensionality reduction approach whose theoretical properties allow us to gain insight into its effects on learning and generalisation. In addition, in this work it will also fulfil a regularisation role – this will become apparent as one of the insights that our analysis yields.

Denote the random projection (RP) matrix by $R \in \mathbb{R}^k \times d$, $k < d$, with independent and identically distributed (i.i.d.) entries, or i.i.d. rows drawn independently of $\mathcal{T}^N$, from a suitable 0-mean 1/$k$-variance distribution, and the compressed training set is $\mathcal{T}^N_R = \{(Rx_n, y_n)\}_{n=1}^N$. The distribution of the entries of $R$ is usually chosen so as to satisfy the Johnson-Lindenstrauss property [13]. In line with this, we shall assume i.i.d. sub-Gaussian entries.

In the reduced $k$-dimensional space we have analogous definitions, and will use a subscript to refer to this reduced space. The functions in the reduced space have the form:
\[ h_R(Rx) = u_R + \sum_{i=1}^M (v_R)_i \cdot \sigma((w_R)_i^T Rx) \]
where $(w_R)_i \in \mathbb{R}^k$, $u_R, (v_R)_i \in \mathbb{R}$ are the parameters that are estimated from $\mathcal{T}^N_R$. Thus, the compressed function class of our interest is $\mathcal{H}_R = \{h_R : h_R(Rx) : x \in \mathcal{X}, \|v_R\|_1 = C_v\}$, where $C_v > 0$ is a constant. We will not restrict the norms of the nonlinear layer’s parameter vectors $(w_R)_i$, because the complexity on this layer is already reduced by the RP. This can be done if further complexity reduction is desired, as discussed in a subsequent section.

Let us denote the sample error minimiser in this reduced class as $\hat{h}_R = \arg \min_{h \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^N (h(Rx_n), y_n)$ is the empirical error in the reduced space. Likewise, the optimal learner within $\mathcal{H}_R$ is denoted as $\hat{h}_R = \arg \min_{h \in \mathcal{H}_R} E[\ell \circ h]$.\[ \]

**II. MAIN RESULT**

We will make use of Rademacher complexities, and it will be convenient to assume that $\mathcal{H}$ is closed under negation. This implies the same for $\mathcal{H}_R$, that is $\mathcal{H}_R = -\mathcal{H}_R$. Therefore the Rademacher complexity [5], [20] of the function class of our interest, $\mathcal{H}_R$, is the following:
\[ \mathcal{R}_N(\mathcal{H}_R) = E_{\gamma} \left[ \sup_{h \in \mathcal{H}_R} \frac{1}{N} \sum_{n=1}^N \gamma_n h(Rx_n) \right] \]
where $\gamma = (\gamma_1, ..., \gamma_N)$ and $\gamma_n$ takes values in $\{-1, 1\}$ with equal probability.

The following theorem bounds the generalisation error of the network $\hat{h}_R$ trained on randomly projected data under fairly standard assumptions.

**Theorem 2.1:** Let $\mathcal{H}$ be the function class of feed-forward neural networks of the form defined in eq.(1), with $L_\alpha$-Lipschitz continuous activation functions $f : \mathbb{R} \to [-b, b]$, and assume $\mathcal{H} = -\mathcal{H}$. Let $h^* = \arg \min_{h \in \mathcal{H}} E[\ell \circ h]$ be the optimal network in this class, with parameters $(W^* = [w^*_1, ..., w^*_M] \in \mathbb{R}^d \times M, v^* = (v^*_1, ..., v^*_M) \in \mathbb{R}^M, u^* \in \mathbb{R})$, where $\|v^*_i\|_1 = C_v$, and $C_v > 0$ is a constant. Denote by $\ell : Y \times Y \to [0, \tilde{\ell}]$ a loss function assumed to be $L_\ell$-Lipschitz in its first argument, and let $\hat{h}_R$ be the sample error minimiser of this loss with respect to $\mathcal{T}^N_R$. Let $R \in \mathcal{M}_{k \times d}, k \leq d$ be a random matrix with i.i.d. sub-Gaussian entries Let $\mathcal{T}^N_R$ be the RP-ed training set of size $N$, where the original sample $\mathcal{T}^N \doteq \mathcal{D}$. We assume that $E[Trace(xx^T)] < \infty$, but $D$ is unknown. Then, for any $\delta \in (0, 1)$, with probability at least $1 - 2\delta$, the generalisation error of $\hat{h}_R$ is upper bounded as the following:
\[ E_{x,y}[\ell \circ \hat{h}_R] \leq E_{x,y}[\ell \circ h^*] + cL_\ell b(1 + C_v) \cdot \sqrt{\frac{k}{N}} + 4\tilde{\ell} \frac{\log 2}{2N} + f_k(W^*, v^*, \delta) \]
where $c$ an absolute constant, and
\[ f_k(W^*, v^*, \delta) = \min \left\{ g_k(W^*, v^*) + \tilde{\ell} \frac{1}{2} \log \left( \frac{1}{7} \right), \frac{g_k(W^*, v^*)}{\delta} \right\} \cdot 1(k < \text{rank}(E[xx^T])) \]
where $g_k(W^*, v^*) = L_\ell L_\alpha \|v^*\|_2 \cdot \|W^*\|_{F \circ \alpha} \cdot \sqrt{\frac{2}{k} \cdot E\|x\|_2^2}$ which $1(\cdot)$ takes the value 1 if its argument is true and 0 otherwise.

Theorem 2.1 bounds the generalisation error of the sample error minimising 2-layer perceptron on randomly projected data, relative to that of the optimal network in the original uncompressed class. It shows that this error is controlled independently of the number of hidden units, by structural characteristics of the uncompressed optimal network, and structural characteristics of the input distribution.

In the main bound, eq. (4), the first term on the r.h.s. is the best achievable generalisation error in the original function class. The next two terms represent the complexity of the function class on dimension-reduced inputs. Under the conditions of the theorem, for the uncompressed function class the complexity would have been of order $\sqrt{d}$, and it is now reduced to $\sqrt{\tilde{\ell}}$. The last term is the price to pay for this reduction; a bias that, when nonzero, it cannot be eliminated by a larger sample size. As expected, the choice of $k$ balances between this bias and the reduced complexity of the function class.

This last term, $f_k(W^*, v^*, \delta)$ can be regarded as a new complexity term as it tells us how compressible is the problem – in other words, it contains information about what makes
it possible to learn the network successfully from randomly projected inputs. Therefore it is interesting to inspect the factors that influence this term. Also interesting to note that it is bounded independently of any label information.

We observe that the compressibility term \( f_k(W^*, v^*, \delta) \) is governed by the compression dimension \( k \), and the regularity of the problem – in this analysis, the degree of regularity is reflected by the size of norms of the weights of the optimal uncompressed network, and the Lipschitz constants of nonlinear functions involved.

Some comments are in order about our choice of network architecture for this analysis. As mentioned in the introduction, for this analysis we opted not to constrain the first layer weights in order to let the random projection exert its regularisation effect. It is well known in the classical literature, that the function class complexity can be reduced by regularisation of the weights [4], [18] – this is of course applicable here too, but the effect of regularising \( w_i, i = 1, \ldots , M \) would be to further reduce (the already reduced) function class complexity, at the expense if a likely increase in the magnitudes of the bias term (while its algebraic expression remains unchanged) – this may be undesirable if \( k \) is small.

We should point out that Theorem 2.1 does not require the input domain \( \mathcal{X} \) to be bounded either. Both the boundedness of \( \mathcal{X} \) and boundedness of all weights are restrictions often made for technical convenience, to ease the analysis. The reason we were able to avoid them is that our analysis exploits the fact that the activation function is bonded so we do not need to worry about the scale of inputs or that of the weights.

A. Bounding the required target dimension

To find out what target dimension \( k \) ensures good generalisation of the compressive learner, we require that the compressibility term is below a given threshold and solving for \( k \).

We shall pursue this from the starting point of eq. (5), since if the threshold is small then the variability is necessarily small as well. From the proof we shall see that the term \( g_k(W^*, v^*) \) is an upper bound on \( E_{R}E_{x,y}[(h^*(R^TT^T x, y) - (h^*(x))] \), it decreases with increasing \( k \), so we require that this term is below some threshold \( \eta > 0 \), yielding the following corollary.

Corollary 2.2: With the notations and conditions of Theorem 2.1, for any \( \eta > 0 \), the required target dimension of the compressed space to ensure \( g_k(W^*, v^*) \leq \eta \) is lower bounded as the following:

\[
k \geq \eta^{-2} \cdot L_{\sigma}^2 \cdot L_{\delta}^2 \cdot \|v^*\|^2_2 \cdot \|W^*\|^2_{F_{v_R}} \cdot 2 \cdot (E\|x\|_2)^2
\]

We may interpret the required \( k \), eq. (6) as the degree of compressibility of the particular learning problem. Our bound suggests that this depends on the degree of regularity of the optimal learner in the original uncompressed function class, as expressed through the norms of the network’s weights, and the Lipschitz constants of the activation function and that of the loss function.

In practice, of course the quantities involved in the bound are unknown. There are other means to set \( k \) in practice, in particular existing model selection methods may be used, such as cross-validation, structural risk minimisation or others. The value of \( k \) obtained may give us a hint about the amount of favourable structure for the problem at hand. However, the main role of our analysis is to gain insights into the characteristics of the data and the problem responsible when observing successful vs. unsuccessful learning of the network from the compressed data. Previous bounds that depended on the number of hidden units completely miss such explanation.

III. Comparison with previous work

The work in [24] derived a lower bound on the required dimension, \( k \), of a linear random projection for the same type of neural network as considered here. They did not consider generalisation analysis, instead their goal was just to bound the absolute difference between the network’s output when fed with the original sample and when fed with the random projection of the same sample. This goal is a subset of ours, as we also bound this as part of our proof.

We summarise the result of [24] below. To avoid confusion, we should point out that the quadratic term in \( M \) is actually missing from the original statement of theorem 3.1 in [24] due to a typo in their proof (see [16] for details), the below is our corrected version of their result.

\[
\text{Theorem 3.1 (} [24]\text{): Consider the class of feedforward neural networks with sigmoid activation function } \sigma, \text{ and } M \text{ hidden units, as defined in eq.(1). Let } R \text{ be a } k \times d, k < d \text{ matrix with entries } R_{ij} \text{ drawn i.i.d. from a Gaussian } N(0,1/k). \text{ For any } \eta > 0, \left( \frac{1}{N} \sum_{i=1}^{N} \sum_{v=1}^{M} v_i \left( \sigma(w_i^T R^T x_n) - \sigma(w_i^T x_n) \right) \right) \leq \eta \text{ with probability at least } 1 - \delta, \text{ provided that,}
\]

\[
k \geq \Omega \left( \eta^{-2} M^2 (\log MN) \max_{i=1}^{M} \left| v_i \cdot \|w_i\|_2 \cdot \sup_{x \in \mathcal{X}} \|x\|_2 \right) \right)
\]

Comparing eq. (7) with our result in eq. (6), the most striking difference is that our lower bound on \( k \) does not depend on \( M \) or \( N \). Hence we can let the number of hidden units \( M \to \infty \), and have as many training points as we can get \( (N \to \infty) \), without necessarily blowing up the compression dimension \( k \). Instead, the various norms of the weight parameters are sufficient to determine the required target dimension \( k \) – these capture the regularity and benign geometry of the problem that makes it solvable on a random subspace.

Our result implies that compressed learning of the network has similar behaviour as the original in the sense that, for good generalisation, the size of the weights matters more than the size of the network. In turn, the lower bound on \( k \) from [24] blows up if \( M \) or/and \( N \) grows without bounds.

We find it insightful to understand the reasons for this difference, which is indeed an artefact of the proof technique used in [24]. The analytic tool used in the analysis of [24] was the Johnson-Lindenstrauss lemma combined with union bounds. The spurious logarithmic factors are due to the use of union bounds, and the explicit dependence on \( M^2 \) is due to the inability of that approach to take advantage of possible
couplings between the weights, or any favourable geometry that may be present in the problem that our bounds are able to exploit.

In addition, a further difference between our approach and that of [24] is that that [24] is restricted to Gaussian random projections. This is because their proof of inner product preservation relies heavily on the rotation invariance of the Gaussian. Sub-Gaussian RPs have better computational scaling while they enjoy similar guarantees [1], [19]. To allow these one could instead use newer results on dot product preservation from [14]. However, in order to eliminate the use of union bounds and be able to exploit the geometry of the problem, we shall pursue a different line of attack instead.

IV. PROOF OF THEOREM 2.1

For a fixed instance of $R$, the Radamacher complexity bound [5], [20] gives that, uniformly for all $h_R \in \mathcal{H}_R$, the following holds $\forall \delta \in (0, 1)$ w.p. $1 - \delta$ over $T^N$:

$$E[\ell \circ h_R] \leq \hat{E}_{T^N}[\ell \circ h_R] + 2\hat{R}_N(\ell \circ \mathcal{H}_R) + 3\delta \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (8)$$

The first two terms on the r.h.s. are the empirical error measured on the compressed training set, and the empirical Rademacher complexity of the function class over the reduced dimensional input space. Both are now functions of $R$.

For the empirical minimiser $\hat{h}_R$ of the loss in $\mathcal{H}_R$ we have:

$$E_{T^N}[\ell \circ \hat{h}_R] = \min_{h_R \in \mathcal{H}_R} E_{T^N}[\ell \circ h_R] \leq \hat{E}_{T^N}[\ell \circ h^*_R] \quad (9)$$

where $h^*_R = \inf_{h_R \in \mathcal{H}_R} E[\ell \circ h_R]$, and we have a Höffding bound in the last line, since $h^*_R$ is fixed. It follows that,

$$E[\ell \circ \hat{h}_R] \leq E[\ell \circ h^*_R] + 2\hat{R}_N(\ell \circ \mathcal{H}_R) + 4\delta \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (10)$$

The first part of the proof is to bound the risk of $h^*_R$ with high probability with respect to the random draw of $R$. The second part will bound the empirical Rademacher complexity.

A. Part I: Upper bound on $E[\ell \circ h^*_R]$

The following is immediate from the definition of $h^*_R$, and decompose the r.h.s. into a distortion term that we call this the compressive distortion, denoted $D$, and the risk of $h^*$:

$$E[\ell \circ h^*_R] = \min_{h_R \in \mathcal{H}_R} E_{x,y}[\ell \circ h_R] \leq \min_{h_R \in \mathcal{H}_R} E_{x,y}[\ell \circ h_R - \ell \circ h^*] + E_{x,y}[\ell \circ h^*] \quad (11)$$

Next, we shall upper bound the compressive distortion $D$.

Let us denote by $W^* \in \mathbb{R}^{d \times M}$ the matrix that holds the weight vectors $w^*_i, i = 1, ..., M$ in its columns, and $v^* = (v^*_1, ..., v^*_M)$. Likewise, $W_R \in \mathbb{R}^{k \times M}$ will be the matrix of weight vectors $(w_R)_i, i = 1, ..., M$ in the reduced space. Using the Lipschitz property of $\ell(\cdot)$ and $\sigma(\cdot)$ we have: $D \leq \ldots$

$$L_{\ell} \cdot \min_{R \sim R, h_R \in H} E_h \left[ \sum_{i=1}^M (w_R)_i \sigma((w_R)_i^T x) + u_R - \sum_{i=1}^M v_i \sigma((v_i)_i^T x) - u^* \right]$$

$$\leq L_{\ell} \cdot \min_{W_R} E_h \left[ \sum_{i=1}^M v_i \sigma((w_R)_i^T x) + u_R - \sum_{i=1}^M v_i \sigma((v_i)_i^T x) - u^* \right]$$

$$\leq L_{\ell} \|v^*\|_2 \min_{W_R} E_h \left[ \sum_{i=1}^M [\sigma((w_R)_i^T x) - \sigma((v_i)_i^T x)]^2 \right]$$

by the Cauchy-Schwarz inequality. This is further bounded by:

$$\leq L_{\ell} L_\sigma \|v^*\|_2 \min_{W_R} E_h \left[ \sum_{i=1}^M (w_R)_i^T x - (w^*_i)_i^T x \right]^2$$

by the Lipschitz property of $\sigma(\cdot)$. Furthermore, we upper bound this as:

$$\leq L_{\ell} L_\sigma \|v^*\|_2 \min_{W_R} E_h \left[ (w_R)_i^T x - (w^*_i)_i^T x \right]$$

by Jensen’s inequality.

Observe that the minimiser in the last line has a closed form:

$$W_R^* = (w^*_i)_i^T E_h[x x^T] R \Sigma_x [x x^T] R^T - 1$$

Hence, $(W_R^*)_i^T = (w^*_i)_i^T E_h[x x^T] R \Sigma_x [x x^T] R^T - 1$. Plugging this back, and denoting $\Sigma = E_h[x x^T]$, we have:

$$\min_{W_R} E_h \left[ (w_R)_i^T x - (w^*_i)_i^T x \right]$$

$$= \sum_{i=1}^M E_h [w^*_i] T \Sigma x - (w^*_i)_i^T x]$$

$$= \sum_{i=1}^M (w^*_i)_i^T \Sigma w^*_i - (w^*_i)_i^T \Sigma R \Sigma x R^T - 1 R \Sigma w^*_i \quad (15)$$

By SVD, $\Sigma = U A U^T$, where $U A U^T = I_d$, and $\Lambda$ is the diagonal matrix of $\rho = \text{rank}(\Sigma)$ positive and $d - \rho$ zero eigenvalues of $\Sigma$. Denote by $A$ the $\rho \times \rho$ non-zero diagonal sub-matrix of $\Lambda$. Denote by $P^T$ the sub-matrix of rows of $R U$ that correspond to the non-zeros in $\Lambda$. Denote by $w^*_i$ the entries of $U^T w^*_i$ that correspond to the non-zeros in $\Lambda$. Then we can rewrite eq.(15) as the following:

$$= \sum_{i=1}^M (w^*_i)_i^T U A U^T w^*_i$$

$$- (w^*_i)_i^T U A U^T R \Sigma x R^T - 1 R \Sigma w^*_i \quad (16)$$

$$= \sum_{i=1}^M (w^*_i)_i^T \Lambda w^*_i - (w^*_i)_i^T \Lambda P \Lambda^T - 1 P \Lambda w^*_i \quad (16)$$

$$= \sum_{i=1}^M (w^*_i)_i^T \Lambda^{1/2} \left[ I_d - \Lambda^{1/2} P \Lambda^T \Lambda^{1/2} \right] \Lambda^{1/2} w^*_i$$

Interesting to observe that the above can be bounded deterministically for any $R$ that is full row rank. Although such
bound is not very informative it nevertheless has some useful features.

Indeed, eq. (16) can be upper bounded irrespective of the random matrix $R$, using the Rayleigh quotient inequality, and noting that the matrix $I_p - \Delta^{1/2} P^T (P \Delta P^T)^{-1} \Delta^{1/2}$ is symmetric:

\[
eq \sum_{i=1}^{M} (w_i^*)^T \Sigma w_i^* \cdot \lambda_{\text{max}} \left(I_p - \Delta^{1/2} P^T (P \Delta P^T)^{-1} \Delta^{1/2}\right) \leq \left\{ \begin{array}{ll}
\frac{\lambda_{\text{max}}}{\sqrt{\text{rank}(\Sigma)}} & \text{if } k < \text{rank}(\Sigma) \\
0 & \text{if } k \geq \text{rank}(\Sigma)
\end{array} \right. \]

(17)

The last line follows from the fact that $\Delta^{1/2} P^T (P \Delta P^T)^{-1} \Delta^{1/2}$ has $k$ eigenvalues $1$, and $\rho - k$ eigenvalues $0$.

Plugging this back into eq. (14) it follows that:

\[
D \leq L_{\ell} \cdot L_\sigma \cdot \|v^*\|_2 \cdot \left\|W^*\right\|_{F \rightarrow 0} \cdot \sqrt{\lambda_{\text{max}}(\Sigma)} \cdot \mathbf{1}(k < \text{rank}(\Sigma))
\]

From this we observe that whenever $k \geq \text{rank}(\Sigma)$ then $D = 0$. However, it is not informative about the behaviour of the compressive distortion as a function of the target dimension $k$ when $k < \text{rank}(\Sigma)$.

So far we did not use the random nature of $R$, and we can obtain a more informative bound by doing so, as follows.

We return to eq. (13), and take a different route, defining the distortion $D$ in a slightly different manner. We have:

\[
E[\ell \circ h_R^*] = \min_{h_R \in \mathcal{H}_R} E_{x,y}[\ell \circ h_R] \leq E_{x,y}[\ell \circ h^* \circ R^T] = E_{x,y}[\ell \circ h^* \circ R^T - \ell \circ h^*] + E_{x,y}[\ell \circ h^*]
\]

(19)

(20)

(21)

The inequality in eq. (20) holds because $\ell \circ h^* \circ R^T \in \mathcal{H}_R$, since the first layer weights are unconstrained, and since both $\mathcal{H}_R$ and $\mathcal{H}$ contain networks whose second layer weight vector has $11$ norm equal to $C_v$.

From here onward, $D$ refers to the quantity in eq. (21). This abuse of notation is because it will play the same role as the previous form of compressive distortion. We have:

\[
D = E_{x,y}[\ell \circ h^* \circ R^T - \ell \circ h^*] = E_{x,y}[\ell(h^*(R^T x), y) - \ell(h^*(x))]
\]

(22)

Here we will take a deviation approach, by which we bound $D$ from its expectation, and then bound the expected distortion $E_R[D]$.

By Höfdding inequality we have for any $\epsilon > 0$, $\Pr\{D \geq E_R[D] + \epsilon\} \leq \exp\left(-2\epsilon^2/\delta^2\right)$. Hence, w.p. at least $1 - \delta$

\[
D \leq E_R[D] + \ell \sqrt{\frac{1}{2} \log \frac{1}{\delta}}
\]

(23)

Note that whenever $E_R[D]$ is small (close to 0) then so is $D - E_R[D]$, since $D$ is always positive. The Markov inequality will capture this. $\Pr\{D \geq \epsilon\} \leq E_R[D]/\epsilon$. Taking the minimum we have:

\[
D \leq \min \left\{ \frac{E_R[D]}{\delta}, E_R[D] + \ell \sqrt{\frac{1}{2} \log \frac{1}{\delta}} \right\}
\]

(24)

Now it remains to bound the expected distortion, $E_R[D]$. Using the Lipschitz property of $\ell(\cdot)$ and $\sigma(\cdot)$ we have:

\[
E_R[D] \leq L_{\ell} \cdot E_{x,R} \left[ \sum_{i=1}^{M} \sigma(w_i^T R^T R x - \sigma(w_i^T x)) \right] \leq L_{\ell} \cdot \|v^*\|_2 E_{x,R} \left[ \sum_{i=1}^{M} \sigma(w_i^T R^T R x - \sigma(w_i^T x))^2 \right]
\]

by Cauchy-Schwarz inequality; this is further upper bounded as:

\[
\leq L_{\ell} \cdot \|v^*\|_2 E_{x,R} \left[ \sum_{i=1}^{M} L_k^2 \cdot \|w_i^T R^T R x - w_i^T x\|^2 \right]
\]

by Lipschitzness of $\sigma(\cdot)$, which in matrix form is the following:

\[
= L_{\ell} \cdot L_\sigma \cdot \|v^*\|_2 \cdot E_{x,R} \left[ W^T R^T R x - W^T x \right]_2
\]

(25)

We can bound the expectation w.r.t. $R$ that appears above, since we know the distribution of the entries of $R$. Specifically, after some algebra it is not too difficult to show that:

\[
E_R \left[ W^T R^T R x - W^T x \right]_2 \leq \sqrt{2 \frac{\kappa}{k}} \cdot \|W\|_{F \rightarrow 0} \cdot \|x\|_2
\]

(26)

Applying this to eq. (25), we obtain:

\[
E_R[D] \leq L_{\ell} \cdot L_\sigma \cdot \|v^*\|_2 \cdot \sqrt{2 \frac{\kappa}{k}} \cdot \|W\|_{F \rightarrow 0} \cdot E_{x} \|x\|_2
\]

(27)

Substituting this back into eq.(24), and taking into account the earlier observation that $D = 0$ when $k \geq \text{rank}(\Sigma)$. Hence we obtain the expression stated in eq. (5).

B. Part II: Upper bound on $\mathcal{R}_N(\ell \circ \mathcal{H}_R)$

As a general comment before proceeding, recall that we did not constrain the first layer weights of the network, and also that we did not require the input domain $X$ to be bounded. These restrictions are often made for the technical ease of Rademacher analysis, but for us the purpose is to let the random projection exert its regularisation effect. Therefore our Rademacher analysis will proceed through the fat shattering bound for linear functions rather than the $\ell_2$ geometry. This is possible because the activation function is bonded so we do not need to worry about the scale of weights.

In turn, the regularising effect of random projections, does not affect the second layer, which is why we had imposed the condition $\|v_R\|_1 = C_v$. Using this, we can start by well
known properties of the empirical Rademacher complexity. As in [20],
\[ \hat{\mathcal{R}}_N(F_R) \leq L_\ell \hat{\mathcal{R}}_N(H_R) \]
were \( F_R \) is the class of Lipschitz continuous functions over the
projected \( k \)-dimensional input space:
\[ F_R = \{ Rx \rightarrow (a((w_R)^T x) + b)/\sqrt{2b} : \mathbb{R}^k \rightarrow [0, 1] \} \]
In eq. (28) we used Talagrand’s contraction lemma (Theorem 7 in [21]). In eq. (29) we used Rademacher identities that hold
for any function class \( H \) [5], [20], namely that \( \mathcal{R}_N(cH) = |c| \mathcal{R}_N(H), \forall c \in \mathbb{R} \), and \( \mathcal{R}_N(\text{conv}(H)) = \mathcal{R}_N(H) \) where conv(.) denotes the convex hull of the set in its argument.

Now, without constraints on \( (w_R) \), we bound the empirical
Rademacher complexity of the Lipschitz continuous function
class \( F_R \) using that this class has a bounded range of values.
By Dudley’s entropy integral inequality [10], the Rademacher
complexity of the Lipschitz continuous function
\[ \mathcal{R}_N(F_R) \leq \sqrt{\log N(\alpha, F_R, \| \cdot \|_2)/N} \]
where \( \| \cdot \|_2 \) is the \( L_2 \)-norm with respect to the empirical
measure i.e., for an \( f \in F_R \), \( \| f \|_2 = \sqrt{\frac{1}{N} \sum_{n=1}^N f^2(x_n)} \).

The covering number in the above can be further bounded in terms of the fat shattering dimension\(^1\) using the following result of [2] (Theorem 2.18 in [22]).

**Theorem 4.1:** [22] There is an absolute constant \( c' \) that satisfies that \( \forall H \in B(L_\infty(X)) \) every sample, every \( 1 \leq p \leq \infty, \forall \alpha \in (0, 1) \),
\[ \mathcal{N}(\alpha, H, \| \cdot \|_2) \leq \left( \frac{2}{\alpha} \right)^{c' \cdot \text{fat}(\alpha, H)} \]
where \( c, c' \) are absolute constants, and \( \text{fat}(\cdot) \) is the fat
shattering dimension of the function class.

Using this yields:
\[ \mathcal{N}(\alpha, F_R, \| \cdot \|_2) \leq \left( \frac{2}{\alpha} \right)^{c' \cdot \text{fat}(\alpha, F_R)} \]

The fat shattering dimension is a measure of the complexity of
a real valued function class. It is known that linear function
classes have fat shattering dimension upperbounded by their
input dimension [11] for any \( \gamma \). Furthermore, composition with
a Lipschitz function (such as a Lipschitz continuous function
in our case) does not change the fat shattering dimension by
more than a constant [4], [11].

Plugging this back, eq. (31) is bounded as:
\[ \hat{\mathcal{R}}_N(F_R) \leq 12 \int_0^1 \left( \frac{c' \log(\alpha/\sqrt{N})}{\sqrt{N}} \right) d\alpha = C \sqrt{\frac{k}{N}} \]

where \( C \) is an absolute constant. This completes the proof
of Theorem 2.1.

**V. CONCLUSIONS AND OUTLOOK**

We proved size-independent guarantees for compressive
learning of 2-layer feed-forward neural networks, that is, in
contrast to what has been previously proved, the required
dimension of the random projection does not explicitly depend
on the number of hidden units in the architecture. Instead,
compressive learning of the network behaves similarly to the
uncompressed one in that small weights, in terms of Frobenius
norm of the first layer weights, matters more than the number
of hidden units. Informally, the intuition from our analysis is
that, the least complex the task the more it can be learned
from drastically compressed data. This is appealing since it
implies that simple tasks demand less computation, and
random projection is a suitable tool that to implement this.

We should mention that our approach in this study exploited
the boundedness of sigmoidal activation functions, so one does
not need to worry about the scale of the first layer weights. It
would be interesting to investigate whether a similar conclusion
would hold in the case unbounded activation functions. This
remains for future work.

Finally, in recent work we have used random projection to
analyse uncompressed learning machines [15], [17] – it would
be interesting, in a similar vein, to leverage the insights gained
from the analysis herein to shed new light on the original
uncompressed problem.

**REFERENCES**

Lindenstrauss with binary coins. J. Computer and System Sciences,
dimensions, uniform convergence, and learnability. J. of the ACM, 4:
cepts and random projection. In 40th Annual Symposium on Foundations
[4] P.L. Bartlett. For valid generalization, the size of the weights is more
important than the size of the network. In Neural Information Processing
bounds and structural results, J. of Machine Learning Research 3:
463-482, 2002.
transform itself preserves differential privacy, Proceedings of the 2012
IEEE 53rd Annual Symposium on Foundations of Computer Science
the number of measurements for reliable compressive classification, IEEE
Transactions on Signal Processing 64(22), 5778-5793, 2016.
functions. Mathematics of Control, Signals, and Systems, 2(4): 303-314,
1989.
Randomness: Probabilistic Algorithms for Constructing Approximate


