Compressive Learning of Multi-layer Perceptrons

Kaban, Ata

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 provably learn from a random linear projection of the data? 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, an answer to this question is of both practical and theoretical relevance. 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 target dimension is at least $\Omega(M^2(\log MN))$, where M is the size of the hidden layer, and N is the number of training points. By contrast, in this paper we show that a target dimension independent of both N and M suffices and is able 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 $\mathcal{X} \subset \mathbb{R}^d$ be an input domain. We denote by $\mathcal{H} = \{x \to h(x) : x \in \mathcal{X}\}$ the function class that implements neural networks of the following parametric form:

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

where $\sigma : \mathbb{R}^d \to [-b, b]$ 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, $w_i \in R^d, u, v_i \in \mathbb{R}$ 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 $\mathcal{T}^N = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}$, and $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = [-b, b] \subset \mathbb{R}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ be a bounded loss function, assumed to be L_ℓ-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 $h \in \mathcal{H}$, this is defined as $\mathrm{E}[\ell \circ h] := \mathrm{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$. However, since $\mathcal{D}$ is unknown, we only have access to the empiri-

This work is funded by EPSRC Fellowship EP/P004245/1, "FORGING: Fortuitous Geometries and Compressed Learning".

cal error, defined as $\hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h] = \frac{1}{N}\sum_{n=1}^{N} \ell(h(x_n), y_n)$. The optimal learner within $\mathcal{H}$ will be denoted as $h^* = \arg\inf_{h \in \mathcal{H}} \mathrm{E}[\ell \circ h]$. The sample error minimiser of the loss is $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h]$.

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 $R \in R^{k \times d}$, $k < d$, 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 $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$. 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:

$$h_R(Rx) = u_R + \sum_{i=1}^{M} (v_R)_i \cdot \sigma((w_R)_i^T Rx) \tag{2}$$

where $(w_R)_i \in R^k, u_R, (v_R)_i \in \mathbb{R}$ are the parameters that are estimated from $\mathcal{T}_R^N$. Thus, the compressed function class of our interest is $\mathcal{H}_R = \{Rx \to h_R(Rx) : x \in \mathcal{X}, \|v_R\|_1 = C_v\}$ 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 $\hat{h}_R = \arg\min_{h_R \in \mathcal{H}_R} \frac{1}{N}\hat{\mathrm{E}}_{\mathcal{T}_R^N}[\ell \circ h_R]$, where $\hat{h}_R \in \mathcal{H}_R$ and $\hat{\mathrm{E}}_{\mathcal{T}_R^N}[\ell \circ h_R] = \frac{1}{N}\sum_{n=1}^{N} \ell(h_R(Rx_n), y_n)$ is the empirical error in the reduced space. Likewise, the optimal learner within $\mathcal{H}_R$ is denoted as $h_R^* = \arg\min_{h_R \in \mathcal{H}_R} \mathrm{E}[\ell \circ h_R]$.

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:

$$\hat{\mathcal{R}}_N(\mathcal{H}_R) = \mathrm{E}_\gamma[\sup_{h_R \in \mathcal{H}_R} \frac{1}{N} \sum_{n=1}^{N} \gamma_n h_R(Rx_n)] \tag{3}$$

where $\gamma = (\gamma_1, ..., \gamma_N)$ and γ_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_σ-Lipschitz continuous activation functions $\sigma : \mathbb{R} \to [-b, b]$, and assume $\mathcal{H} = -\mathcal{H}$. Let $h^* = \arg\inf_{h \in \mathcal{H}} \mathrm{E}[\ell \circ h]$ be the optimal network in this class, with parameters $(W^* = [w_1^*, ..., w_M^*] \in \mathbb{R}^{d \times M}, v^* = (v_1^*, \ldots, v_M^*) \in \mathbb{R}^M$, $u^* \in \mathbb{R})$, where $\|v^*\|_1 = C_v$, and $C_v > 0$ is a constant. Denote by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ a loss function assumed to be L_ℓ-Lipschitz in its first argument, and let $\hat{h}_R$ be the sample error minimiser of this loss with respect to $\mathcal{T}_R^N$. Let $R \in \mathcal{M}_{k \times d}, k \leq d$ be a random matrix with i.i.d. sub-Gaussian entries Let $\mathcal{T}_R^N$ be the RP-ed training set of size N, where the original sample $\mathcal{T}^N \overset{\text{i.i.d}}{\sim} \mathcal{D}$. We assume that $\mathrm{E}[Trace(xx^T)] < \infty$, but $\mathcal{D}$ is unknown. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta - \phi(\delta)$, the generalisation error of $\hat{h}_R$ is upper bounded as the following:

$$\begin{aligned} \mathrm{E}_{x,y}[\ell \circ \hat{h}_R] &\leq \mathrm{E}_{x,y}[\ell \circ h^*] + cL_\ell b(1 + C_v) \cdot \sqrt{\frac{k}{N}} \\ &+ 4\bar{\ell}\sqrt{\frac{\log \frac{2}{\delta}}{2N}} + f_k(W^*, v^*, \delta) \end{aligned} \tag{4}$$

where c an absolute constant, and

$$\begin{aligned} &f_k(W^*, v^*, \delta) = \\ &\min\left\{ g_k(W^*, v^*) + \bar{\ell}\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)} \ , \ \frac{g_k(W^*, v^*)}{\delta} \right\} \\ &\qquad \cdot \mathbf{1}(k < \mathrm{rank}(\mathrm{E}[xx^T])) \end{aligned}$$

where

$$g_k(W^*, v^*) = L_\ell L_\sigma \|v^*\|_2 \cdot \|W^*\|_{Fro} \cdot \sqrt{\frac{2}{k}} \cdot \mathrm{E}[\|x\|_2] \tag{5}$$

where $\mathbf{1}(\cdot)$ 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 $\sqrt{k}$. 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 $w_i, i = 1, \dots M$ 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 $\mathrm{E}_R\mathrm{E}_{x,y}|\ell(h^*(R^TRx), y) - \ell(h^*(x))|$, 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:

$$k \geq \eta^{-2} \cdot L_\ell^2 \cdot L_\sigma^2 \cdot \|v^*\|_2^2 \cdot \|W^*\|_{Fro}^2 \cdot 2 \cdot (\mathrm{E}\|x\|_2)^2 \tag{6}$$

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.

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

$$k \geq \Omega\left(\eta^{-2} M^2 (\log MN) \max_{i=1}^{M} |v_i| \cdot \|w_i\|_2 \cdot \sup_{x \in \mathcal{X}} \|x\|_2\right) \tag{7}$$

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 Rademacher 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 $\mathcal{T}^N$:

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

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:

$$\begin{aligned} \hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ \hat{h}_R] &= \min_{h_R \in \mathcal{H}_R} \hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h_R] \le \hat{\mathrm{E}}_{\mathcal{T}^N}[\ell \circ h_R^*] && (9) \\ &\le \mathrm{E}[\ell \circ h_R^*] + \bar{\ell}\sqrt{\frac{\log \frac{2}{\delta}}{2N}} && (10) \end{aligned}$$

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

$$\mathrm{E}[\ell \circ \hat{h}_R] \le \mathrm{E}[\ell \circ h_R^*] + 2\hat{\mathcal{R}}_N(\ell \circ \mathcal{H}_R) + 4\bar{\ell}\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \quad (11)$$

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^*:

$$\begin{aligned} \mathrm{E}[\ell \circ h_R^*] &= \min_{h_R \in \mathcal{H}_R} \mathrm{E}_{x,y}[\ell \circ h_R] && (12) \\ &\le \underbrace{\min_{h_R \in \mathcal{H}_R} \mathrm{E}_{x,y}|\ell \circ h_R - \ell \circ h^*|}_{D} + \mathrm{E}_{x,y}[\ell \circ h^*] && (13) \end{aligned}$$

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: $D \leqslant \ldots$

$$\begin{aligned} &L_\ell \cdot \min_{v_R, W_R, u_R} \mathrm{E}_x|\sum_{i=1}^M (v_R)_i \sigma((w_R)_i^T Rx) + u_R - \sum_{i=1}^M v_i^* \sigma((w_i^*)^T x) - u^*| \\ &\leqslant L_\ell \cdot \min_{W_R} \mathrm{E}_x|\sum_{i=1}^M v_i^* \sigma((w_R)_i^T Rx) + u^* - \sum_{i=1}^M v_i^* \sigma((w_i^*)^T x) - u^*| \\ &\leqslant L_\ell \|v^*\|_2 \min_{W_R} \mathrm{E}_x \sqrt{\sum_{i=1}^M [\sigma((w_R)_i^T Rx) - \sigma((w_i^*)^T x)]^2} \end{aligned}$$

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

$$\le \quad L_\ell L_\sigma \|v^*\|_2 \min_{W_R} \mathrm{E}_x \sqrt{\sum_{i=1}^M |(w_R)_i^T Rx - (w_i^*)^T x|^2}$$

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

$$\begin{aligned} &\le \quad L_\ell L_\sigma \|v^*\|_2 \min_{W_R} \mathrm{E}_x \|W_R^T Rx - (W^*)^T x\|_2 \\ &\le \quad L_\ell L_\sigma \|v^*\|_2 \sqrt{\min_{W_R} \mathrm{E}_x \|W_R^T Rx - (W^*)^T x\|_2^2} && (14) \end{aligned}$$

by Jensen's inequality.

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

$$W_R^T = (W^*)^T \mathrm{E}_x[xx^T] R^T (R\mathrm{E}_x[xx^T]R^T)^{-1}$$

Hence, $(W_R)_i^T = (w_i^*)^T \mathrm{E}_x[xx^T] R^T (R\mathrm{E}_x[xx^T]R^T)^{-1}$. Plugging this back, and denoting $\Sigma = \mathrm{E}_x[xx^T]$, we have:

$$\begin{aligned} &\min_{W_R} \mathrm{E}_x \|W_R^T Rx - (W^*)^T x\|_2^2 \\ = \quad &\sum_{i=1}^M \mathrm{E}_x |(w_i^*)^T \Sigma R^T (R\Sigma R^T)^{-1} Rx - (w_i^*)^T x|^2 \\ = \quad &\sum_{i=1}^M (w_i^*)^T \Sigma w_i^* - (w_i^*)^T \Sigma R^T (R\Sigma R^T)^{-1} R\Sigma w_i^* && (15) \end{aligned}$$

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

$$\begin{aligned} = \quad &\sum_{i=1}^M (w_i^*)^T U\Lambda U^T w_i^* \\ - \quad &(w_i^*)^T U\Lambda U' R^T (RU\Lambda U^T R^T)^{-1} RU\Lambda U^T w_i^* \\ = \quad &\sum_{i=1}^M (\underline{w}_i^*)^T \underline{\Lambda w}_i^* - (\underline{w}_i^*)^T \underline{\Lambda} P^T (P\underline{\Lambda}P^T)^{-1} P\underline{\Lambda w}_i^* \\ = \quad &\sum_{i=1}^M (\underline{w}_i^*)^T \underline{\Lambda}^{1/2} \left(I_\rho - \underline{\Lambda}^{1/2} P^T (P\underline{\Lambda}P^T)^{-1} P\underline{\Lambda}^{1/2}\right) \underline{\Lambda}^{1/2} \underline{w}_i^* \\ & && (16) \end{aligned}$$

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 $I_\rho - \underline{\Lambda}^{1/2}P^T(P\underline{\Lambda}P^T)^{-1}P\underline{\Lambda}^{1/2}$ is symmetric:

$$\begin{aligned}
& \text{eq. (16)} \\
\leq\ & \sum_{i=1}^{M}(w_i^*)^T\Sigma w_i^* \cdot \lambda_{\max}(I_\rho - \underline{\Lambda}^{1/2}P^T(P\underline{\Lambda}P^T)^{-1}P\underline{\Lambda}^{1/2}) \qquad (17) \\
=\ & \begin{cases} \sum_{i=1}^{M}(w_i^*)^T\Sigma w_i^*, \text{ if } k < \text{rank}(\Sigma) \\ 0, \text{ if } k \geq \text{rank}(\Sigma) \end{cases} \\
\leq\ & \begin{cases} \lambda_{\max}(\Sigma)\|W^*\|_{Fro}^2, \text{ if } k < \text{rank}(\Sigma) \\ 0, \text{ if } k \geq \text{rank}(\Sigma) \end{cases} \qquad (18)
\end{aligned}$$

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

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

$$D \leq L_\ell \cdot L_\sigma \cdot \|v^*\|_2 \cdot \|W^*\|_{Fro} \cdot \sqrt{\lambda_{\max}(\Sigma)} \cdot \mathbf{1}(k < \text{rank}(\Sigma))$$

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:

$$\begin{aligned}
& \text{E}[\ell \circ h_R^*] \\
=\ & \min_{h_R \in \mathcal{H}_R} \text{E}_{x,y}[\ell \circ h_R] \qquad (19) \\
\leq\ & \text{E}_{x,y}[\ell \circ h^* \circ R^T] \qquad (20) \\
\leq\ & \underbrace{\text{E}_{x,y}|\ell \circ h^* \circ R^T - \ell \circ h^*|}_{D} + \text{E}_{x,y}[\ell \circ h^*] \qquad (21)
\end{aligned}$$

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 l1 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:

$$\begin{aligned}
D &= \text{E}_{x,y}|\ell \circ h^* \circ R^T - \ell \circ h^*| \\
&= \text{E}_{x,y}|\ell(h^*(R^TRx), y) - \ell(h^*(x))| \qquad (22)
\end{aligned}$$

Here we will take a deviation approach, by which we bound D from its expectation, and then bound the expected distortion $\text{E}_R[D]$.

By Höffding inequality we have for any $\epsilon > 0$, $\Pr\{D \geqslant \text{E}_R[D] + \epsilon\} \leqslant \exp\left(-2\epsilon^2/\bar{\ell}^2\right)$. Hence, w.p. at least $1 - \delta$

$$D \leqslant \text{E}_R[D] + \bar{\ell}\sqrt{\frac{1}{2}\log\frac{1}{\delta}} \tag{23}$$

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

$$D \leq \min\left\{\frac{\text{E}_R[D]}{\delta}\ ,\ \text{E}_R[D] + \bar{\ell}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\} \tag{24}$$

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

$$\begin{aligned}
\text{E}_R[D] &\leq L_\ell \cdot \text{E}_{R,x}|\sum_{i=1}^{M} v_i^*\left(\sigma(w_i^{*T}R^TRx) - \sigma(w_i^{*T}x)\right)| \\
&\leq L_\ell \cdot \|v^*\|_2 \text{E}_{x,R}\sqrt{\sum_{i=1}^{M}[\sigma(w_i^{*T}R^TRx) - \sigma(w_i^{*T}x)]^2}
\end{aligned}$$

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

$$\leq L_\ell \cdot \|v^*\|_2 \text{E}_{x,R}\sqrt{\sum_{i=1}^{M} L_\sigma^2 \cdot [w_i^{*T}R^TRx - w_i^{*T}x]^2}$$

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

$$= L_\ell \cdot L_\sigma \cdot \|v^*\|_2 \cdot \text{E}_{x,R}\|W^{*T}R^TRx - W^{*T}x\|_2 \tag{25}$$

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:

$$\text{E}_R\|W^{*T}R^TRx - W^{*T}x\|_2 \leq \sqrt{\frac{2}{k}} \cdot \|W^*\|_{Fro} \cdot \|x\|_2. \tag{26}$$

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

$$\text{E}_R[D] \leq L_\ell \cdot L_\sigma \cdot \|v\|_2 \cdot \sqrt{\frac{2}{k}} \cdot \|W\|_{Fro} \cdot \text{E}_x\|x\|_2 \tag{27}$$

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 $\hat{\mathcal{R}}_N(\ell \circ \mathcal{H}_R)$

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 $\mathcal{X}$ 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 l2 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],

$$\begin{aligned} \hat{\mathcal{R}}_N(\ell \circ \mathcal{H}_R) &\leq L_\ell \hat{\mathcal{R}}_N(\mathcal{H}_R) && (28) \\ &= L_\ell(1 + C_v)2b\hat{\mathcal{R}}_N(\mathcal{F}_R) && (29) \end{aligned}$$

were $\mathcal{F}_R$ is the class of Lipschitz continuous functions over the projected k-dimensional input space:

$$\mathcal{F}_R = \{Rx \rightarrow (\sigma((w_R)_i^T Rx) + b)/(2b) : \mathbb{R}^k \rightarrow [0,1]\} \quad (30)$$

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 $\hat{\mathcal{R}}_N(cH) = |c|\hat{\mathcal{R}}_N(H), \forall c \in \mathbb{R}$, and $\hat{\mathcal{R}}_N(\text{conv}(H)) = \hat{\mathcal{R}}_N(H)$ where $\text{conv}(\cdot)$ denotes the convex hull of the set in its argument.

Now, without constraints on $(w_R)_i$, we bound the empirical Rademacher complexity of the Lipschitz continuous function class $\mathcal{F}_R$ using that this class has a bounded range of values. By Dudley's entropy integral inequality [10], the Rademacher complexity of any $[0,1]$-valued function class can be bounded in terms of covering numbers:

$$\hat{\mathcal{R}}_N(\mathcal{F}_R) \quad \leq \quad 12\int_0^1 \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}_R, \|\cdot\|_2)}{N}} d\alpha \quad (31)$$

where $\|\cdot\|_2$ is the $\mathcal{L}_2$-norm with respect to the empirical measure i.e. for an $f \in \mathcal{F}_R$, $\|f\|_2 = \sqrt{\frac{1}{N}\sum_{n=1}^N f^2(x_n)}$.

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'_p that satisfies that $\forall \mathcal{H} \in B(L_\infty(X))$ every sample, every $1 \leq p \leq \infty, \forall \alpha \in (0,1)$,

$$\mathcal{N}(\alpha, \mathcal{H}, \|\cdot\|_2) \leq \left(\frac{2}{\alpha}\right)^{c_p \cdot \text{fat}_{c'_p \alpha}(\mathcal{H})} \quad (32)$$

where c, c' are absolute constants, and $\text{fat}_\gamma(\cdot)$ is the fat shattering dimension of the function class.

Using this yields:

$$\mathcal{N}(\alpha, \mathcal{F}_R, \|\cdot\|_2) \leq \left(\frac{2}{\alpha}\right)^{c \cdot \text{fat}_{c'\alpha}(\mathcal{F}_R)} \quad (33)$$

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 γ. 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:

$$\hat{\mathcal{R}}_N(\mathcal{F}_R) \quad \leq \quad 12\int_0^1 \sqrt{\frac{c'' k \log(2/\alpha)}{N}} d\alpha = C\sqrt{\frac{k}{N}} \quad (34)$$

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 known previously, the required dimension of the random projection does not explicitly depend of 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 shednew light on the original uncompressed problem.

References

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Computer and System Sciences, 66(4): 671–687, 2003.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability. J. of the ACM, 4: 615-631, 1997.

[3] R.I. Arriaga, S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In 40th Annual Symposium on Foundations of Computer Science (FOCS 1999), pp. 616-623, 1999.

[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 Systems 9 (NIPS), pp. 134-140, 1997.

[5] P.L. Bartlett, S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results, J. of Machine Learning Research 3: 463-482, 2002.

[6] J. Blocki, A. Blum, A. Datta, O. Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy, Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), pp. 410-419, 2012.

[7] H. Reboredo, F. Renna, R. Calderbank, M.R.D. Rodrigues. Bounds on the number of measurements for reliable compressive classification, IEEE Transactions on Signal Processing 64(22), 5778-5793, 2016.

[8] G. Cybenko. Approximations by superpositions of Lipschitz continuous functions. Mathematics of Control, Signals, and Systems, 2(4): 303-314, 1989.

[9] S. Dasgupta. Learning mixtures of Gaussians, in Annual Symposium on Foundations of Computer Science (FOCS), pp. 634-644, 1999.

[10] R.M. Dudley. Uniform central limit theorems. Cambridge University Press, Cambridge, MA, 1999.

[1] Let $\gamma > 0$ be fixed, and let $\mathcal{F}$ be a function class. We say that F γ-shatters a set $A \subset X$ if $\exists s : A \rightarrow \mathbb{R}$ s.t. $\forall E \subseteq A, \exists f_E \in \mathcal{F}$ satisfying that $\forall x \in A \setminus E, f_E(x) \leq s(x) - \gamma$ and $\forall x \in E, f_E(x) \geq s(x) + \gamma$.
The maximum cardinality of $A \subseteq X$ that is γ-shattered by $\mathcal{F}$ is defined as the fat-shattering dimension of $\mathcal{F}$, denoted $\text{fat}_\gamma(\mathcal{F})$.

[11] L. Gurvits and P. Koiran. Approximation and learning of convex superpositions, in Computational Learning Theory: EUROCOLT95, 1995.

[12] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, SIAM Review, 53(2), pp. 217288, 2011.

[13] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space, Contemporary Mathematics 26, pp. 189-206, 1984.

[14] A. Kabán. Improved bounds on the dot product under random projection and random sign projection, in ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 487-496, 2015.

[15] A. Kabán and R.J. Durrant. Structure-aware error bounds for linear classification with the zero-one loss. arXiv:1709.09782, 2017.

[16] A.Kabán and Y. Thummanusarn. Tighter Guarantees for the Compressive Multi-layer Perceptron. 7th International Conference on the Theory and Practice of Natural Computing (TPNC18), Dublin, Ireland December 12-14, 2018. Lecture Notes in Computer Science (LNCS, vol. 11324), pp. 388-400.

[17] A. Kabán. Dimension-Free Error Bounds from Random Projections. 33rd AAAI Conference on Artificial Intelligence (AAAI-19), January 27 February 1, 2019, Hawaii, USA (to appear).

[18] S.M. Kakade, K. Sridharan, A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization, in Neural Information Processing Systems (NIPS), pp. 793-800, 2008.

[19] D.M. Kane, J. Nelson. Sparser Johnson-Lindenstrauss transforms, J. of the ACM 61(1), Article 4, 2014.

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

[21] R. Meir, T. Zhang. Generalization error bounds for Bayesian Mixture Algorithms, J. of Machine Learning Research 4: 839-860, 2003.

[22] S. Mendelson. A Few notes on statistical learning theory, in S. Mendelson and A. J. Smola (eds), Advanced Lectures in Machine Learning, vol. 2600 of Lecture Notes in Computer Science, pp. 1-40, Springer-Verlag, Berlin, 2003.

[23] M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of machine learning, MIT Press, 2012.

[24] E. Skubalska-Rafajlowicz. Neural networks with Lipschitz continuous activation functions: dimension reduction using normal random projection. Nonlinear Analysis, 71(12): e1255-e1263, 2009.