A RANDOM MATRIX AND CONCENTRATION INEQUALITIES FRAMEWORK FOR NEURAL NETWORKS ANALYSIS

Cosme Louart, Romain Couillet

CentraleSupélec, Gif-sur-Yvette, France.

ABSTRACT

This article provides a theoretical analysis of the asymptotic performance of a regression or classification task performed by a simple random neural network. This result is obtained by leveraging a new framework at the crossroads between random matrix theory and the concentration of measure theory. This approach is of utmost interest for neural network analysis at large in that it naturally dismisses the difficulty induced by the non-linear activation functions, so long that these are Lipschitz functions. As an application, we provide formulas for the limiting law of the random neural network output and compare them conclusively to those obtained practically on handwritten digits databases.

***Index Terms*—** Neural networks, random matrix theory, concentration inequalities, extreme learning machines.

1. INTRODUCTION

One of the main popularity features of deep neural networks lies in their (still barely fathomed) performance stability. That is, as the number n and size p of the training data grow large (and so does the network), independently of the random initialization point of the backpropagation learning algorithm, essentially the same performances are ultimately achieved. This characteristic is at the core of a current stream of research, based on tools from statistical physics and random matrix theory [1, 2, 3], aiming at theorizing these observations. And, indeed, [2] explored a model *akin to* deep neural networks and concludes that the local minima of the learning cost function become increasingly dense as data and network grow large, having essentially the same associated loss, and that the probability to escape these minima vanishes. However, the statistical physics model of [2] is in reality far from satisfying from a neural network perspective as it, for once, breaks all dependence induced by the non-linearities of the activation functions (ReLU non-linearities being replaced by products with independent Bernoulli random variables) and, most importantly, assumes data to be constituted of random independent entries; both conditions ensure that only random uncorrelated scalars propagate through the network, a highly criticizable model for deep nets. Alternative neural network analyses discard the non-linearities altogether, as in [4]; but in this case, convergence to global minima are studied, which are known not to be achieved by practical deep networks (fortunately so, as this avoids overfitting).

In this article, following our seminal works [5, 6], we propose a different angle of approach to neural network analysis. Rather than modelling a complete deep neural net, we focus here primarily on simple network structures, so far not considering backpropagation learning but accounting for non-linearities induced when traversing a hidden layer. The main technical driver to this endeavor is the *concentration of measure theory*, which has the key features of (i) extending many results holding for vectors of independent entries to the wider scope of *concentrated random vectors* (see definition in subsequent sections) and most importantly of (ii) being a theory "stable to Lipschitz mappings" in that Lipschitz functions $f : \mathbb{R}^p \to \mathbb{R}^q$ of concentrated random vectors in $\mathbb{R}^p$ are still concentrated vectors in $\mathbb{R}^q$. Feature (ii) notably allows one to accommodate with the non-linear activation functions, so long that these are Lipschitz (e.g., ReLU, sigmoid maps).

In [5], we merely exploited Feature (ii) as a technical means to study the asymptotic (as $n, p \to \infty$) performance of extreme learning machines (ELM) [7] (i.e., single hidden-layer regression networks with no backpropagation learning), assuming a model encompassing a random connectivity matrix (which induces the concentration of the output vectors) but *deterministic data*. Under this model, however, while the asymptotic network training performance was readily accessible, the asymptotic generalization performance remained out of technical grasp and only a conjecture under "reasonable" yet unclear assumptions on the deterministic dataset could be proposed. As an answer, the present study strongly suggests that a key property of neural network stability (and likely of many statistical learning methods) lies, not in the (initial) randomness of the inter-layer connections, but rather in a *concentration property of the dataset*. This property structurally appears when studying the orders of magnitude of the output of an ELM for concentrated versus deterministic data (in the former case the output has a controlled magnitude, while in the latter case the output may diverge as $n, p \to \infty$). Exploiting Features (i) and (ii) together, and therefore working on *concentrated input data* in the first place, brings us to a more

This work is supported by the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

generic analysis framework, where Lipschitz non-linearities need not be explicitly studied as they do not affect the concentration properties of the data.

Under this setting, in the present work, we generalize several results from random matrix theory, by providing notably a deterministic equivalent for the covariance matrix of a k-mixture of concentrated vectors along with the asymptotic statistical behavior of a ridge-regression on these vectors. As an immediate aftermath, the asymptotic performance analysis of ELMs, and in passing the spectral characterization of random feature maps, then reduce to mere corollaries by specifying the structural properties of the concentrated vectors.

Notation. In the remainder, $C, c > 0$ as well as $C_\ell, c_\ell > 0$ are constants independent of all other parameters, and $C', c' > 0$ constants dependent only on C, c.

2. SYSTEM SETTING

2.1. Basic notions of concentration of measure

We start our system modelling assumptions with a few definitions and essential notions of the concentration of measure theory [8] that will be used in this article.

Definition 1 (Concentration of measure)**.** *The random variable $Z \in \mathbb{R}$ is said to be concentrated and we denote $Z \propto \alpha(\cdot)$ if, for Z' an independent copy of Z,*

$$\mathbb{P}\left(|Z - Z'| \geq t\right) \leq \alpha(t).$$

We say that Z is normally concentrated, if $Z \propto Ce^{-(\cdot/c)^2}$.

Normal concentrations owe their names to the fact that normal random variables are normally concentrated ; when optimized, the parameter c of the definition has the same order as the standard deviation of the variable. Normally concentrated variables interestingly concentrate around their means in the sense that, $Z \propto Ce^{-(\cdot/c)^2} \Rightarrow \mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq C'e^{-(t/2c)^2}$. Also, these fast concentrations induce moment controls:

$$Z \propto Ce^{-(\cdot/c)^2} \Rightarrow \forall r > 0,\ \mathbb{E}[|Z - \mathbb{E}Z|^r] \leq C'(2r)^{r/2}c^r$$

As $f(|Z - Z'|) \leq \lambda|Z - Z'|$ for λ-Lipschitz functions f, we have the following structural property of the theory.

Property 1 (Lipschitz maps)**.** *For $f : \mathbb{R} \to \mathbb{R}$ a λ-Lipschitz function and $Z \in \mathbb{R}$,*

$$Z \propto \alpha(\cdot) \Rightarrow f(Z) \propto \alpha(\cdot/\lambda).$$

Similarly, linear combinations of concentrated random variables remain concentrated. Products of concentrated random variables are more difficult to handle, but we have the following lemma, of importance in this article.

Lemma 1 (Concentration of squared variables[1])**.** *If we assume that $Z \propto Ce^{-(\cdot/c)^2}$, then, with obvious notations,*

$$Z^2 \propto C'e^{-\cdot/2c^2} + C'e^{-(\cdot/4c|\mathbb{E}Z|)^2}.$$

The concentration of measure theory however finds its fullest significance when considering random vectors (rather than scalars) $Z \in \mathbb{R}^p$. As most random vectors of practical interest do not *localize* (e.g., large Gaussian vectors tend to spread along a sphere), the notion of concentration of measure for vectors is defined by means of *all their Lipschitz "observations"*.

Definition 2 (Concentration of random vectors)**.** *A vector $Z \in \mathbb{R}^p$ is concentrated, denoted $Z \propto \alpha(\cdot)$, if for every 1-Lipschitz map $f : \mathbb{R}^p \to \mathbb{R}$, $f(Z) \propto \alpha(\cdot)$.*

A non-trivial result is that standard normal random vectors $Z \in \mathbb{R}^p$ are indeed normally distributed, in the sense of Definition 2, with parameters C, c *independent of* p.

Theorem 1. *[8, Prop. 1.9.], $Z \sim \mathcal{N}(0, I_p) \Rightarrow Z \propto 2e^{-\cdot^2/2}$.*

With these notions at hand, we are in position to present our work setting.

2.2. Setting

Keeping in mind that neural networks are mostly used for regression or classification, we consider here a set of input-output data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^d$, p will be supposed to be at most of order $O(n)$ and d remains small irrespective of n. Our core assumption is that the vectors x_l are independent and Lipschitz transformations of canonical Gaussian vectors to obtain with Theorem 1 the concentration (for C, c independent of p and n):

$$X = [x_1, \ldots, x_n] \propto Ce^{-(\cdot/c)^2}.$$

Besides, we suppose that the vectors are distributed in only k distributions $\mu_1, \ldots, \mu_k$ (with k small compared to p and n), i.e., for each $l \leq n$, $x_l \sim \mu_\ell$ for some $\ell \leq k$. For $Z \sim \mu_\ell$, we denote $\bar{M}_\ell \equiv \mathbb{E}[Z]$ (and $\bar{M} = [\bar{M}_1, \ldots, \bar{M}_k]$) and $\Sigma_\ell \equiv \mathbb{E}[ZZ^\mathsf{T}]$ (not to be confused with the covariance matrix). To avoid technicalities, we assume that $C < \max_{1 \leq \ell \leq k}\{\frac{1}{p}\operatorname{tr}\Sigma_\ell\} < C'$ for $C, C' > 0$ independent of p. We finally denote n_ℓ the number of x_l's drawn from μ_ℓ.

This set of hypotheses is of interest as it notably encompasses the cases where the x_l's are the output of a *random feature map* $x_l = \sigma(Ws_l)$ for $W \in \mathbb{R}^{p \times q}$ deterministic with $\|W\| \leq 1$, $\sigma : \mathbb{R} \to \mathbb{R}$ a 1-Lipschitz map (operating here

[1] This result is similar to the Hanson–Wright inequality [9, Th. 6.2.1], but more adequate to our present setting; the proof, provided in an extended version of this article [10], is also structurally simpler as it relies on more elementary properties.

entry-wise), and $s_l \in \mathbb{R}^q$ such that $s_l \propto C_\ell e^{-(\cdot/c)^2}$ (for instance $s_l \sim \mathcal{N}(m_\ell, \Sigma_\ell)$). This justifies our claim that, under this setting, the crucial concentration property can propagate through the different layers of a feed-forward neural network disregarding the specificities of the non-linear activation function since $s_l \propto C_\ell \alpha_{\mathcal{N}}(\sqrt{c_\ell}\,\cdot) \Rightarrow x_l \propto C_\ell \alpha_{\mathcal{N}}(\sqrt{c_\ell}\,\cdot)$.

3. MAIN RESULTS

Under the assumptions of Section 2.2, extreme learning machines [7] may be merely seen as a linear ridge-regression with training set $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_l = \sigma(W s_l)$ for some input observations $s_1, \ldots, s_n$. Since linear ridge-regression involves as a core object the sample covariance matrix

$$C_X \equiv \frac{1}{n} X X^\mathsf{T} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\mathsf{T}$$

with $X \equiv [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$, our first objective is to characterize the eigenspectrum of this matrix for large p and n. This first analysis shall subsequently allow for a full characterization of functionals of the random matrix X, among which the performance of ELM regression and classification.

3.1. Sample covariance matrix analysis

To tackle the eigenspectrum analysis of C_X, random matrix theory [11, 12] provides a quite versatile tool: the resolvent

$$Q_X \equiv (C_X + z I_p)^{-1}$$

of C_X, for $z > 0$.[2] The matrix Q_X encapsulates much information about C_X; notably, the so-called Stieltjes transform $\frac{1}{p}\operatorname{tr} Q_X$ uniquely characterizes the empirical eigenvalue distribution $\frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(C_X)}$ of C_X [13, Th. B.9.], while quadratic forms of the type $a^\mathsf{T} Q_X a$ allow for a characterization of the projections $|a^\mathsf{T} u_i(C_X)|$ of deterministic vectors $a \in \mathbb{R}^p$ on the "isolated" eigenvectors $u_i(C_X)$ of C_X [14].

Besides, Q_X is a convenient tool for the present article as it naturally transfers concentrations. Indeed, since $\|Q_X\| \leq z^{-1}$ and $\|Q_X X\| \leq \sqrt{n}$, it is easily shown that the mapping $X \mapsto Q_X$ is $2/\sqrt{nz}$-Lipschitz so that $X \propto \alpha(\cdot) \Rightarrow Q_X \propto \alpha(\sqrt{nz}/2\,\cdot)$ (with X and Q_X respectively seen as vectors in $\mathbb{R}^{np}$ and $\mathbb{R}^{p^2}$).

Owing to this property, we have the following core result.

Theorem 2. *Let $\bar{Q}$ be defined as*

$$\bar{Q} \equiv \left(\sum_{\ell=1}^k \frac{n_\ell}{n} \frac{\Sigma_\ell}{1+\delta_\ell} + z I_p \right)^{-1}$$

where $(\delta_1, \ldots, \delta_k) \in \mathbb{R}_+^k$ is the unique solution with nonnegative elements of the system $\delta_\ell = \frac{1}{n}\operatorname{tr}\Sigma_\ell \bar{Q}$, $\ell = 1, \ldots, k$. Then, $\|\mathbb{E}[Q_X] - \bar{Q}\| \leq \frac{C'\sqrt{p}}{n}$, and for any matrix $A \in \mathcal{M}_{p,n}$:

$$\mathbb{P}\left(\left|\operatorname{tr} A(Q_X - \bar{Q})\right| \geq t\right) \leq C e^{-(\sqrt{n}t/c\|A\|_1)^2},$$

where $\|A\|_1 = \operatorname{tr}\left((AA^\mathsf{T})^{\frac{1}{2}}\right)$ is the nuclear norm of A.

Note that $\|I_p\|_1 = p$ and for any $u, v \in \mathbb{R}^p$, $\left\|uv^T\right\|_1 = \|u\|\,\|v\|$. We thus deduce from this theorem the normal concentration of $u^T Q v$ and $\frac{1}{p}\operatorname{tr} Q$ respectively around $u^T \tilde{Q} v$ and $\frac{1}{p}\operatorname{tr}\tilde{Q}$ with a standard deviation of order $O(\frac{1}{\sqrt{n}})$.

Sketch of Proof. Given the bound on $\|\mathbb{E}Q_X - \bar{Q}\|$, the second and third results follow from $Q_X - \bar{Q} = (Q_X - \mathbb{E}Q_X) + (\mathbb{E}Q_X - \bar{Q})$ and normal concentration inequalities based on the Lipschitz character of $X \mapsto Q_X$. We are then left to prove the first result. Let $X_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$, $Q_{-i} = Q_{X_{-i}}$ and $Q = Q_X$. By the identity $Q x_i (1 + \frac{1}{n} x_i^\mathsf{T} Q_{-i} x_i) = Q_{-i} x_i$, we have

$$\mathbb{E}[Q] - \bar{Q} = \sum_{\ell=1}^k \frac{n_\ell}{n} \mathbb{E}[\varepsilon_\ell^1 + \varepsilon_\ell^2]$$

where, for any $x_l \sim \mu_\ell$, we introduced the matrices

$$\varepsilon_\ell^1 \equiv \frac{Q_{-l} x_l x_l^\mathsf{T} \bar{Q}(\delta_\ell - \frac{1}{n} x_l^\mathsf{T} Q_{-l} x_l)}{(1 + \frac{1}{n} x_l^\mathsf{T} Q_{-l} x_l)(1+\delta_\ell)}, \quad \varepsilon_\ell^2 \equiv \frac{Q_{-l} x_l x_l^\mathsf{T} Q \Sigma_\ell \bar{Q}}{n(1+\delta_\ell)}.$$

Because of the leading factor $\frac{1}{n}$ in ε_ℓ^2, algebraic manipulations similar to [6] ensure that this term vanishes faster than ε_ℓ^1. As for ε_ℓ^1, the main technical part is to control

$$\begin{aligned} x_l^\mathsf{T} Q_{-l} x_l - n\delta_\ell &= (x_l^\mathsf{T} Q_{-l} x_l - \operatorname{tr}\Sigma_\ell Q_{-l}) \\ &\quad + \operatorname{tr}\Sigma_\ell (Q_{-l} - \mathbb{E}Q_{-l}) + (\operatorname{tr}\Sigma_\ell \mathbb{E}Q_{-l} - n\delta_\ell). \end{aligned}$$

Since $\|Q_{-l}\| \leq z^{-1}$ and $x_l \propto C_\ell e^{-(\cdot/c)^2}$, Lemma 1 applied to $x_l^\mathsf{T} Q_{-l} x_l = \|Q_{-l}^{\frac{1}{2}} x_l\|^2$ ensures that the first right-hand side difference is normally-exponentially concentrated. Since $X \mapsto Q_X$ is Lipschitz, the second difference is normally concentrated. As for the third term, its control follows from pre-established random matrix results (see e.g., [15]). Since normal and exponential concentrations convert to bounds on moments, $\mathbb{E}[\varepsilon_\ell^1]$ can be appropriately bounded, thereby completing the proof. □

3.2. Ridge-regression and ELM classification

Linear ridge-regression for the training data pairs (x_i, y_i) previously defined consists in determining the vector $\beta \in \mathbb{R}^{p \times d}$ that minimizes, for $z > 0$, the cost

$$E_{\text{train}}(\beta) \equiv \frac{1}{n} \sum_{i=1}^n \|y_i - \beta^\mathsf{T} x_i\|_F^2 + z\|\beta\|_F^2$$

[2] Usually one defines Q_X as $Q_X = (C_X - z I_p)^{-1}$ for $z \in \mathbb{C}$ with $\Im z > 0$ but analyticity arguments along with the nonnegative definiteness of C_X justify the equivalence of this definition here.

(with $\|\cdot\|_F$ the Frobenius norm). Letting $Y = [y_1, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^{n\times d}$, the solution is explicitly given by $\beta^\star = \frac{1}{n} Q_X XY$. For a test data pair $(x, y) \in \mathbb{R}^{p\times d}$, the regression output of x is then given by

$$S(x) = x^\mathsf{T}\beta^\star = \frac{1}{n} x^\mathsf{T} Q_X XY.$$

For k-class classification purposes, one naturally takes $y_l = e_\ell \in \mathbb{R}^k$ when $x_l \sim \mu_\ell$ (hence $d = k$), with $[e_\ell]_a = \delta_a^\ell$ the indicator vector of class ℓ. This procedure strongly relates to the kernel LS-SVM approach [16, 17].

The associated regression and classification performance measures are the mean-square error $\mathbb{E}[\|S(x) - y\|^2]$ and the misclassification rate, respectively. In both cases, these relate to the probability distribution of $S(x)$, which is then our object of present interest. Our main result, restricted for readability to the classical random matrix regime on n, p, reads:

Theorem 3. *Let $x \sim \mu_\ell$ and assume $\max_{1\le i\le n} \|y_i\| < C$. Then, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$,*

$$\mathcal{V}_\ell^{-\frac{1}{2}} \left(S(x) - \bar{S}_\ell\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_d)$$

where $\mathcal{V}_\ell = V_\ell - \bar{S}_\ell \bar{S}_\ell^\mathsf{T}$,

$$\bar{S}_\ell \equiv \frac{1}{n} \bar{M}_\ell^\mathsf{T} \bar{Q} \bar{M} \Delta J^\mathsf{T} Y$$

$$V_\ell \equiv \frac{1+\delta_\ell}{n^2} Y^\mathsf{T} J \Delta \left([\Phi]_{\ell,\cdot,\cdot} + D_\Psi^\ell - \frac{n}{n_\ell}(D^\ell \Theta + \Theta D^\ell) \right) \Delta J^\mathsf{T} Y$$

where $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n\times k}$ with $[j_\ell]_i = \delta_{x_i \sim \mu_\ell}$, $\Phi = (I_k - \tilde{\Psi})^{-1}\tilde{\Phi} \in \mathbb{R}^{k\times k\times k}$,[3] $\Psi = (I_k - \tilde{\Psi})^{-1}\tilde{\Psi} \in \mathbb{R}^{k\times k}$, $D^\ell = \operatorname{diag}(\Psi_{\cdot,\ell}) \in \mathbb{R}^{k\times k}$, $D_\Psi^\ell = \operatorname{diag}(\frac{n^2(1+\delta_\cdot)}{n_\cdot^2}\Psi_{\ell,\cdot}) \in \mathbb{R}^{k\times k}$, $\Delta = \operatorname{diag}((1+\delta_\cdot)^{-1})$, and

$$\Theta = \bar{M}^\mathsf{T} \bar{Q} \bar{M}$$

$$\tilde{\Phi} = \left\{ \frac{\bar{M}_i^\mathsf{T} \bar{Q} \Sigma_l \bar{Q} \bar{M}_j}{1+\delta_l} \right\}_{1\le l,i,j\le k}$$

$$\tilde{\Psi} = \left\{ \frac{n_j}{n^2} \frac{\operatorname{tr} \bar{Q}\Sigma_i \bar{Q}\Sigma_j}{(1+\delta_i)(1+\delta_j)} \right\}_{1\le i,j\le k}.$$

Proof. Asymptotic means and covariances follow from a concentration inequality-based analysis. The central limit is then obtained from a refined version of [18] adapted to our present setting. Details are provided in the extended article [10]. □

3.3. Description of our results

As discussed previously, letting $x_l = \sigma(Ws_l)$ for $s_l \in \mathbb{R}^q$ with $q \sim p$ such that $s_l \propto C_\ell e^{-(\cdot/c)^2}$, $\sigma(\cdot)$ 1-Lipschitz and $W \in \mathbb{R}^{p\times q}$ with $\|W\| \le 1$ (thereby ensuring that $x_l \propto C_\ell e^{-(\cdot/c)^2}$), $S(x)$ is the output of an ELM. The asymptotic statistics of $S(x)$ in Theorem 3 therefore directly translate in terms of simple neural network performances.

[3] Here we understand the product AB for $A \in \mathbb{R}^{k\times k}$ and $B \in \mathbb{R}^{k\times k\times k}$ by $[AB]_{abc} = \sum_d A_{ad} B_{dbc}$.

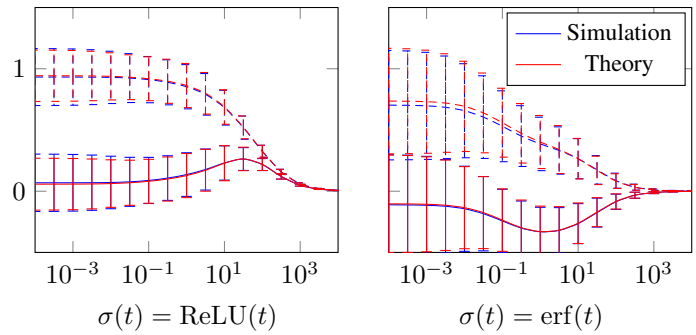

Fig. 1. Scores $[S(x)]_1$ (blue dashed) and $[S(x)]_2$ (blue solid) of 2-class MNIST ELM as a function of regularization $z \in [10^{-4}, 10^4]$ (digits 3 for $\mathcal{C}_1$ and 8 for $\mathcal{C}_2$) for $x \in \mathcal{C}_1$, versus theory (red dashed and solid). Based on $n = 2048$ samples, $p = q = 784$, W random unitary.

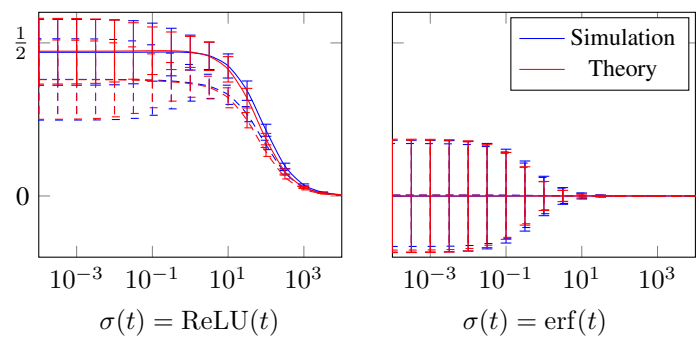

Fig. 2. Scores $[S(x)]_1$ (blue dashed) and $[S(x)]_2$ (blue solid) of Gaussian 2-mixture ELM as a function of $z \in [10^{-4}, 10^4]$ ($\mathcal{C}_1 \equiv \mathcal{N}(0, I_p)$ and $\mathcal{C}_2 \equiv \mathcal{N}(0, 2I_p)$) for $x \in \mathcal{C}_1$, versus theory (red dashed and solid). Based on $n = 4096$ samples, $p = q = 256$, W random unitary.

Figure 1 provides the simulated $S(x) \in \mathbb{R}^2$ versus theoretical average $\bar{S}_\ell$ output of an ELM for 2-class $(\mathcal{C}_1, \mathcal{C}_2)$ classification with $\sigma(t) = \mathrm{ReLU}(t) = \max(t, 0)$ and $\sigma(t) = \mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du$, W random unitary, and s_l extracted from the MNIST handwritten digits dataset [19]. Here $Y \in \mathbb{R}^{n\times k}$ is defined by $Y_{l\ell} = \delta_{s_l \in \mathcal{C}_\ell}$. Note the accurate fit between theory and practice, suggesting that the MNIST data are conveniently modelled as concentrated random vectors.

Figure 2 proceeds similarly with Gaussian $s_l \sim \mathcal{N}(0, \alpha_\ell I_p)$ with $\alpha_\ell \in \{1, 2\}$ according to the class. Being non-linearly separable classes, a straightforward application of Theorem 3 ensures that both classes are non-discriminable for $\sigma(t)$ such that $\sigma(-t) = -\sigma(t)$, which Figure 2 visually confirms; indeed, for such symmetric σ, for all ℓ, $\bar{M}_\ell = \mathbb{E}[\sigma(Ws_l)] = 0$ where $s_l \sim \mathcal{N}(0, \alpha_\ell I_p)$, and we thus find that $\bar{S}_\ell = 0$ indeed not allowing for class discrimination.

4. REFERENCES

[1] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in neural information processing systems*, 2014, pp. 2933–2941.

[2] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial Intelligence and Statistics*, 2015, pp. 192–204.

[3] J. Pennington and Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory," in *International Conference on Machine Learning*, 2017, pp. 2798–2806.

[4] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[5] C. Louart and R. Couillet, "Harnessing neural networks: a random matrix approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, New Orleans, USA, 2017.

[6] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *(in Press) Annals of Applied Probability*, 2017.

[7] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[8] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2005, no. 89.

[9] R. Vershynin, *High dimensional probability*, 2017.

[10] C. Louart, R. Couillet, and F. Benaych-Georges, "Sample covariance of concentrated random vectors," (in preparation).

[11] T. Tao, *Topics in random matrix theory*. American Mathematical Soc., 2012, vol. 132.

[12] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. NY, USA: Cambridge University Press, 2011.

[13] Z. D. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, 2nd ed. New York, NY, USA: Springer Series in Statistics, 2009.

[14] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[15] F. Benaych-Georges and R. Couillet, "Spectral analysis of the gram matrix of mixture models," *ESAIM: Probability and Statistics*, vol. 20, pp. 217–237, 2016. [Online]. Available: http://dx.doi.org/10.1051/ps/2016007

[16] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[17] R. C. Zhenyu Liao, "A large dimensional analysis of least squares support vector machines," *(submitted to) Journal of Machine Learning Research*, 2017.

[18] S. Chatterjee, "Fluctuations of eigenvalues and second order poincaré inequalities," *Probability Theory and Related Fields*, vol. 143, no. 1-2, pp. 1–40, 2009.

[19] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," 1998.