HARNESSING NEURAL NETWORKS: A RANDOM MATRIX APPROACH

Cosme Louart, Romain Couillet

CentraleSupélec, Gif-sur-Yvette, France.

ABSTRACT

This article proposes an original approach to the performance understanding of large dimensional neural networks. In this preliminary study, we study a single hidden layer feed-forward network with random input connections (also called *extreme learning machine*) which performs a simple regression task. By means of a new random matrix result, we prove that, as the size and cardinality of the input data and the number of neurons grow large, the network performance is asymptotically deterministic. This entails a better comprehension of the effects of the hyper-parameters (activation function, number of neurons, etc.) under this simple setting, thereby paving the path to the harnessing of more involved structures.

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

1. INTRODUCTION

Artificial neural networks have had a long history of successive crests and troughs, which may be summarized as the result of an arms race between the theoretically grounded signal processing (with methods such as support vector machines) and the sequential improvement of computational power and size of available datasets of modern computer science to run neural networks. The incompatibility of neural networks and signal processing primarily lies in the inherent intractability of neural network performances, which mainly originates from the non linearity of the neural activations (as well as from learning by back-propagation of the error).

With this observation in mind, we propose here a theoretical study of the performance of *large dimensional neural networks* (in the sense of large datasets and number of neurons) in the instrumental setting of a single hidden layer neural network with random input connections, sometimes referred to as extreme learning machines (ELM) [1]. Although a poor model of present deep learning structures, ELMs are simple networks that allow to focus on the present task in this article: harnessing the non-linearity riddle. We believe that the extension of our present findings to multiple layers and the introduction of several steps of learning by gradient descent is a merely secondary, more easily addressable, task.

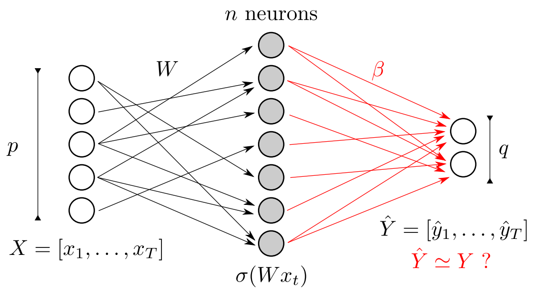

Fig. 1. Extreme learning machine neural network.

Precisely, assuming the ELM depicted in Figure 1, constituted of n neurons trained for T time steps by input data $x_1, \ldots, x_T \in \mathbb{R}^p$ to achieve a regression task, we shall show that, as $n, T, p \to \infty$ at proportional growth rate and under mild assumptions on the activation function σ at each neuron and on the random connectivity matrix $W \in \mathbb{R}^{p \times n}$, the (ridge) regression performance of the neural network is asymptotically deterministic. The obtained formulas notably unveil important performance properties, such as the relevance of the distribution of the entries of W (provably not the case in linear networks) as well as unexpected common features of several classical choices of σ (sign, ReLu, etc.).

2. SYSTEM SETTING

We start our study by considering an *extreme learning machine* neural network, as introduced in [1] and depicted in Figure 1. The network is fed by a set of T input data vectors $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$ and is trained to map corresponding vectors $Y = [y_1, \ldots, y_T] \in \mathbb{R}^{q \times T}$, through a single hidden layer of n neurons with non-linear activation function σ. The training phase in ELMs is particular in that only the hidden layer-to-sink connectivity matrix $\beta \in \mathbb{R}^{n \times q}$ is learnt while the input-to-hidden layer connectivity matrix $W \in \mathbb{R}^{p \times n}$ is static but randomly selected.

For a given β, the output of the ELM is thus simply given by $\hat{Y} = \beta^T \Sigma \in \mathbb{R}^{q \times T}$, where $\Sigma \equiv \sigma(WX)$, and the application of σ is understood entry-wise. During the training phase, the output weight matrix β is merely learnt by ridge

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

regression, by solving the quadratic minimization problem

$$\beta = \mathrm{argmin}_{\tilde{\beta}\in\mathbb{R}^{n\times q}} \left\{ \frac{1}{T} \left\| \tilde{\beta}^{\mathsf{T}}\Sigma - Y \right\|_F^2 + \gamma \|\tilde{\beta}\|_F^2 \right\}$$

with $\|\cdot\|_F$ the Frobenius matrix norm, $\gamma > 0$ some constant. The solution is explicit and given by

$$\beta = \frac{1}{T}\Sigma \left(\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma + \gamma I_T \right)^{-1} Y^{\mathsf{T}}$$

which entails a (per sample) training mean-square error

$$E_{\mathrm{train}}(\gamma) = \frac{1}{T} \left\| \beta^{\mathsf{T}}\Sigma - Y \right\|_F^2 = \frac{\gamma^2}{T} \mathrm{tr}\, \left(Y^T Y Q(\gamma)^2 \right)$$

where we defined $Q(\gamma) \equiv \left(\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma + \gamma I_T \right)^{-1}$, the so-called *resolvent* of the matrix $\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma$. It shall be particularly convenient in the following to observe that

$$E_{\mathrm{train}}(\gamma) = -\frac{\gamma^2}{T} \frac{\partial}{\partial\gamma} \mathrm{tr}\, \left(Y^T Y Q(\gamma) \right). \tag{1}$$

Once training is achieved, the performance of the ELM is assessed from its generalization capability, measured here as the mean square error $E_{\mathrm{test}}(\gamma)$ on test data $\tilde{X}$ of size $\tilde{T}$, with corresponding output $\tilde{Y}$, as

$$E_{\mathrm{test}}(\gamma) = \frac{1}{\tilde{T}} \left\| \beta^{\mathsf{T}}\sigma(W\tilde{X}) - \tilde{Y} \right\|_F^2.$$

The quantities E_{train} and E_{test} are inherently random, as the connectivity matrix W is chosen randomly (which is also true of more elaborate neural networks since a random weight initialization is often carried out). Our objective is to show that, under rather mild assumptions on the distribution of the entries of W and on the activation function σ, E_{train} and E_{test} converge to deterministic values when $n, T, \tilde{T}, p \to \infty$ at the same rate, while q is fixed. This result is presented next.

3. MAIN RESULTS

In the remainder, we shall assume that $n, T, \tilde{T}, p$ grow at the same rate in the sense that there exists $0 < m < M$ such that, as $T \to \infty$, $m < n/T < M$, $m < n/\tilde{T} < M$ and $m < p/T < M$. As for q, we assume it fixed for all values of $n, p, T, \tilde{T}$. In particular, the phrase $n \to \infty$ or $T \to \infty$ shall indicate that $n, p, T, \tilde{T} \to \infty$.

For simplicity in this article, we shall only focus on the simpler E_{train}. From (1) and classical random matrix considerations (see e.g., [2, 3]), to describe E_{train} as $T \to \infty$, one needs to exhibit a so-called *deterministic equivalent* to the resolvent $Q(\gamma)$ of $\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma$; that is, a matrix $\bar{Q}(\gamma)$ such that, for all bounded norm, deterministic vectors $a, b \in \mathbb{R}^T$ and matrix $A \in \mathbb{R}^{T\times T}$, $a^{\mathsf{T}}(Q(\gamma) - \bar{Q}(\gamma))b \to 0$ and $\frac{1}{T}\mathrm{tr}\, A(Q(\gamma) - \bar{Q}(\gamma)) \to 0$, almost surely. To this end, most random matrix methods fundamentally rely on the independence or linear dependence of the entries of Σ to retrieve $\bar{Q}(\gamma)$ [2, 4]. However, even when assuming the entries of W independent (as shall be done next), our present setting falls outside this standard scenario as $\Sigma = \sigma(WX)$ has independent rows but non-linearly dependent columns. As a work around, we shall exploit the fact that the rows of Σ satisfy a concentration of measure identity under mild assumptions on σ and W, which is a sufficiently strong property to recover a deterministic equivalent $\bar{Q}(\gamma)$, somewhat in the same spirit as in the work [5].

3.1. Concentration of measure preliminaries

Our first result is therefore a concentration of measure property, for which we need the following set of assumptions.

Assumption 1 (Connectivity Matrix W)**.** *The matrix $W \in \mathbb{R}^{n\times p}$ has independent and identically distributed rows having ($\mathbb{R}^p$-supported) measure μ_W satisfying the concentration*

$$\alpha_{\mu_W}(t) \leq \kappa e^{-Kpt^2} \tag{4}$$

for some $\kappa, K > 0$ and $\alpha_\mu(t) \equiv \sup\left\{1 - \mu(A_t) \mid \mu(A) \geq \frac{1}{2}\right\}$ with $A_t = \{x \in \mathbb{R}^p \mid \exists y \in A,\ \|x - y\| \leq t\}$.

Assumption 1 essentially states that the rows of W exhibit a concentration of measure phenomenon. An example is W with i.i.d. Gaussian or uniform entries with zero mean and variance $1/n$ but, more generally, W with rows having a Lebesgue density e^{-U} where the Hessian of U is greater than cI_p, $c > 0$, is appropriate [6, 5]. Since Lipschitz functions propagate concentration of measure, we subsequently need:

Assumption 2 (Lipschitz σ)**.** *The function σ is λ-Lipschitz for some $\lambda > 0$.*

Assumption 3 (Boundedness of X)**.** *There exists $\kappa > 0$ such that $\|X\| \leq \kappa\sqrt{p}$, with $\|\cdot\|$ the matrix operator norm.*

Assumptions 1–3 induce the necessary concentration of measure for Σ as follows:

Proposition 1 (Concentration of Measure for Σ)**.** *Let Assumptions 1–3 hold. Then the rows of $\frac{1}{\sqrt{T}}\Sigma = \frac{1}{\sqrt{T}}\sigma(WX)$ are i.i.d. with measure μ_Σ satisfying*

$$\alpha_{\mu_\Sigma}(t) \leq \kappa' e^{-K'pt^2} \tag{5}$$

for some $\kappa', K' > 0$ and α_μ defined in Assumption 1.

The arguments follow the classical results from [6] but are not detailed here in the interest of space. A more general assumption would be to directly request the conclusion of Proposition 1 as the base assumption, which in general is however not necessarily simple to ensure. Notably, we believe that the rows of $\frac{1}{\sqrt{T}}\Sigma$ may concentrate without requiring σ to be Lipschitz, e.g., for $\sigma(t) = 1_{\{t>0\}}$ or $\sigma(t) = \mathrm{sign}(t)$, possibly to the expense of more stringent assumptions on X. These considerations are left to future investigations.

$$\bar{E}_{\mathrm{train}}(\gamma) = \frac{\gamma^2}{T}\operatorname{tr}\left(Y^\mathsf{T} Y \bar{Q}(\gamma)\left[\frac{\frac{1}{n}\operatorname{tr}(\Psi_X \bar{Q}(\gamma)^2)}{1-\frac{1}{n}\operatorname{tr}(\Psi_X\bar{Q}(\gamma))^2}\Psi_X + I_T\right]\bar{Q}(\gamma)\right) \tag{2}$$

$$\begin{aligned}\bar{E}_{\mathrm{test}}(\gamma) &= \left\|\tilde{Y}^\mathsf{T} - \Psi_{X,\tilde{X}}^\mathsf{T}\bar{Q}(\gamma)Y^\mathsf{T}\right\|_F^2 \\ &\quad + \frac{\frac{1}{n}\operatorname{tr} Y\bar{Q}(\gamma)\Psi_X\bar{Q}(\gamma)Y^\mathsf{T}}{1-\frac{1}{n}\operatorname{tr}(\Psi_X\bar{Q}(\gamma))^2}\left[\frac{1}{\tilde{T}}\operatorname{tr}\left(\Psi_{\tilde{X}}\right) - \frac{\gamma}{\tilde{T}}\operatorname{tr}\left(\bar{Q}(\gamma)\Psi_{X,\tilde{X}}\Psi_{\tilde{X},X}\bar{Q}(\gamma)\right) - \frac{1}{\tilde{T}}\operatorname{tr}\left(\Psi_{\tilde{X},X}\bar{Q}(\gamma)\Psi_{X,\tilde{X}}\right)\right]\end{aligned} \tag{3}$$

where $\Phi_{A,B} = \mathrm{E}[\frac{1}{n}\sigma(WA)^\mathsf{T}\sigma(WB)]$, $\Psi_{A,B} = \frac{n}{T}\frac{1}{1+\delta}\Phi_{A,B}$, and $\Psi_A = \Psi_{A,A}$.

3.2. Deterministic Equivalent

We provide here a heuristic development of our main technical result. The detailed, mathematically thorough, arguments are deferred to an extended version of the article. Recall that our objective is to retrieve a matrix $\bar{Q}(\gamma)$ such that, for $a, b \in \mathbb{R}^T$ and $A \in \mathbb{R}^{T\times T}$ deterministic and bounded, $a^\mathsf{T}(Q(\gamma) - \bar{Q}(\gamma))b \to 0$ and $\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \to 0$, almost surely. Focusing only on the latter identity (and leaving the former to the extended version of the article), in the same spirit as in [7], let us write $\bar{Q}(\gamma) = (F + \gamma I_T)^{-1}$ for some deterministic $F \in \mathbb{R}^{T\times T}$ to be identified. Then, we have

$$\begin{aligned}&\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \\ &= \frac{1}{T}\operatorname{tr} AQ(\gamma)\left(F - \frac{1}{T}\Sigma^\mathsf{T}\Sigma\right)\bar{Q}(\gamma) \\ &= \frac{1}{T}\operatorname{tr} AQ(\gamma)F\bar{Q}(\gamma) - \frac{1}{T}\sum_{i=1}^n \frac{1}{T}\Sigma_{i,\cdot}\bar{Q}(\gamma)AQ(\gamma)\Sigma_{i,\cdot}^\mathsf{T}\end{aligned}$$

where the first equality uses $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, while the second equality uses $\Sigma^\mathsf{T}\Sigma = \sum_{i=1}^n \Sigma_{i,\cdot}^\mathsf{T}\Sigma_{i,\cdot}$. For a reason that will become evident next, we now write

$$Q(\gamma)\Sigma_{i,\cdot}^\mathsf{T} = \frac{Q_{-i}(\gamma)\Sigma_{i,\cdot}^\mathsf{T}}{1+\frac{1}{T}\Sigma_{i,\cdot}Q_{-i}(\gamma)\Sigma_{i,\cdot}^\mathsf{T}}$$

with $Q_{-i} = (\frac{1}{T}\Sigma^\mathsf{T}\Sigma - \frac{1}{T}\Sigma_{i,\cdot}^\mathsf{T}\Sigma_{i,\cdot} + \gamma I_T)^{-1}$, using here the identity $(A + vv^\mathsf{T})^{-1}v = A^{-1}v/(1 + v^\mathsf{T}A^{-1}v)$. This entails

$$\begin{aligned}&\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \\ &= \frac{1}{T}\operatorname{tr} AQ(\gamma)F\bar{Q}(\gamma) - \frac{1}{T}\sum_{i=1}^n \frac{\frac{1}{T}\Sigma_{i,\cdot}\bar{Q}(\gamma)AQ_{-i}(\gamma)\Sigma_{i,\cdot}^\mathsf{T}}{1+\frac{1}{T}\Sigma_{i,\cdot}Q_{-i}(\gamma)\Sigma_{i,\cdot}^\mathsf{T}}.\end{aligned}$$

This form has the advantage of isolating terms of the type $\frac{1}{T}\Sigma_{i,\cdot}A\Sigma_{i,\cdot}^\mathsf{T}$ for A a matrix *independent* of $\Sigma_{i,\cdot}$.

The main difficulty in extending the standard random matrix techniques (devised in e.g., [2, 4]) to the present study lies in the generalization of a key proof ingredient at this point of the derivation, often referred to as the *trace lemma* (e.g., [2, Lemma B.26]). This lemma states that, for $x \in \mathbb{R}^n$ with i.i.d. entries having zero mean and unit variance and $A \in \mathbb{R}^{n\times n}$ Hermitian independent of x and with bounded operator norm, $\frac{1}{n}x^\mathsf{T}Ax - \frac{1}{n}\operatorname{tr} A \to 0$, almost surely as $n \to \infty$. For $\sigma(t) = t$ and W having i.i.d. zero mean and variance $1/n$ entries, using this lemma, we would easily obtain that $\frac{1}{T}\Sigma_{i,\cdot}A\Sigma_{i,\cdot}^\mathsf{T} - \frac{1}{Tn}\operatorname{tr}(XAX^\mathsf{T}) \to 0$. However, for non-linear σ, $\Sigma_{i,\cdot}$ does not have linearly dependent entries and we therefore must resort to a generalization of the trace lemma. A direct consequence of Proposition 1 precisely provides this result as follows.

Proposition 2 (Trace lemma for Σ)**.** *Under Assumptions 1–3, with $A \in \mathbb{R}^{T\times T}$ deterministic of bounded spectral norm,*

$$\max_{1\le i\le n}\left|\frac{1}{T}\Sigma_{i,\cdot}A\Sigma_{i,\cdot}^\mathsf{T} - \frac{1}{T}\operatorname{tr}\Phi_X A\right| \to 0$$

almost surely, where $\Phi_X = \mathrm{E}\left[\frac{1}{n}\Sigma^\mathsf{T}\Sigma\right]$.

From Proposition 2, we now have that, as $T \to \infty$,

$$\begin{aligned}&\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \\ &\asymp \frac{1}{T}\operatorname{tr} AQ(\gamma)F\bar{Q}(\gamma) - \frac{1}{T}\sum_{i=1}^n \frac{\frac{1}{T}\operatorname{tr}\Phi_X\bar{Q}(\gamma)AQ_{-i}(\gamma)}{1+\frac{1}{T}\operatorname{tr}\Phi_X Q_{-i}(\gamma)} \\ &\asymp \frac{1}{T}\operatorname{tr} AQ(\gamma)F\bar{Q}(\gamma) - \frac{n}{T}\frac{\frac{1}{T}\operatorname{tr}\Phi_X\bar{Q}(\gamma)AQ(\gamma)}{1+\frac{1}{T}\operatorname{tr}\Phi_X Q(\gamma)}\end{aligned}$$

where by $x_n \asymp y_n$, we mean informally that $x_n - y_n \to 0$ as $n \to \infty$, and where the second equality uses the fact that $\frac{1}{T}\operatorname{tr}(A+vv^\mathsf{T})^{-1} - \frac{1}{T}\operatorname{tr} A^{-1} \to 0$ if A is positive definite with smallest eigenvalue uniformly away from zero (ensured here by $\gamma > 0$). Since we aim at having $\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \to 0$, a clear choice for F is to take it as the solution to the implicit equation $F = \frac{n}{T}\Phi_X/(1 + \frac{1}{T}\operatorname{tr}\Phi_X\bar{Q}(\gamma))$ (it is implicit since $\bar{Q}(\gamma) = (F + \gamma I_T)^{-1}$), and we obtain the desired result.

From these heuristic derivations, we have thus reached our main technical result.

Theorem 1 (Deterministic Equivalent)**.** *Under Assumptions 1–3, for $A \in \mathbb{R}^{T\times T}$ deterministic bounded, as $n \to \infty$,*

$$\frac{1}{T}\operatorname{tr} A(Q(\gamma)-\bar{Q}(\gamma)) \to 0, \quad \bar{Q}(\gamma) \equiv \left(\frac{n}{T}\frac{\Phi_X}{1+\delta} + \gamma I_T\right)^{-1}$$

almost surely, with δ the positive solution to $\delta = \frac{1}{T}\operatorname{tr}\Phi_X\bar{Q}(\gamma)$.

$\sigma(t)$	W_{ij}	$[\Phi_{A,B}]_{ij}$
t	any	$\frac{m_2}{n} a_i^\mathsf{T} b_j$
t^2	any	$\frac{m_2^2}{n^2}(\sigma(a_i^\mathsf{T} b_j) + 2\sigma(a_i)^\mathsf{T} 1_p 1_p^\mathsf{T} \sigma(b_j) + \frac{m_4 - 3m_2^2}{n^2}\sigma(a_i)^\mathsf{T}\sigma(b_j))$
$\max(t,0)$	$\mathcal{N}(0,\frac{1}{n})$	$\frac{1}{2\pi n}\Vert a_i\Vert \Vert b_j\Vert \left(Z_{ij} \operatorname{acos}(-Z_{ij}) + \sqrt{1 - Z_{ij}^2} \right)$
$\operatorname{erf}(t)$	$\mathcal{N}(0,\frac{1}{n})$	$\frac{2}{\pi} \operatorname{asin}\left(\frac{a_i^\mathsf{T} b_j}{\sqrt{(n+2\Vert a_i\Vert^2)(n+2\Vert b_j\Vert^2)}} \right)$
$1_{\{t>0\}}$	$\mathcal{N}(0,\frac{1}{n})$	$\frac{1}{2} - \frac{1}{2\pi}\operatorname{acos}(Z_{ij})$
$\operatorname{sign}(t)$	$\mathcal{N}(0,\frac{1}{n})$	$1 - \frac{2}{\pi}\operatorname{acos}(Z_{ij})$.

Fig. 2. Values of $\Phi_{A,B}$ for W_{ij} i.i.d. with zero mean and k-th order moments $m_k n^{-\frac{k}{2}}$, $Z_{ij} \equiv \frac{a_i^\mathsf{T} b_j}{\Vert a_i\Vert \Vert b_j\Vert}$.

3.3. Neural Network Performance

Theorem 1 is the key intermediary result to investigate E_{train} and E_{test}. The complete investigation requires more advanced but mostly classical random matrix tools and we only provide here the final results.

Proposition 3 (Network Performance)**.** *Under Assumptions 1–3, as $n \to \infty$, we have $E_{\text{train}} - \bar{E}_{\text{train}} \to 0$ and $E_{\text{test}} - \bar{E}_{\text{test}} \to 0$, almost surely, where $\bar{E}_{\text{train}}$ and $\bar{E}_{\text{test}}$ are deterministic, defined in* (2) *and* (3) *in the previous page.*

It is useful to obtain the exact value of $\Phi_{A,B}$ for various W and σ, as they are at the core of Theorem 1 and Proposition 3. Some examples, obtained through various integration tricks, are provided in Figure 2. Even though Theorem 1, as it stands, does not cover non-Lipschitz σ, we display here some other practically used σ. Note in particular the surprising closeness of the formulas for $1_{\{t>0\}}$, $\text{ReLu}(t) = \max(t, 0)$ and $\operatorname{sign}(t)$ which all revolve on the angles $\operatorname{acos}(Z_{ij})$.

Proposition 3 along with Figure 2 allow for a theoretical evaluation of numerous configurations of the neural network. A comparison between theory and practice is depicted in Figure 3; there, X is extracted from the MNIST handwritten digit database [8] with $x_1, \ldots, x_{T/2} \in \mathbb{R}^p$ vectorized images of zeros and $x_{T/2+1}, \ldots, x_T \in \mathbb{R}^p$ of ones, and associated $y_1, \ldots, y_{T/2} = -1$, $y_{T/2+1}, \ldots, y_T = 1$ (here $q = 1$). We took $n = 512$, $T = 1024$ and $p = 784$ and tested over another set of $\tilde{T} = 1024$ images, $W_{ij} \sim \mathcal{N}(0, 1/n)$. As expected, non-linear activation functions σ provide better performances in this highly non linear classification task.

4. CONCLUDING REMARKS

This study lays the first steps of a theoretical analysis of large dimensional neural networks, starting from the elementary case of ELMs. The training and testing performances were provided deterministic approximations featuring the main hyper-parameters of the network. While seemingly complex, these expressions greatly simplify in specific cases, such as when (i) either n, p or T grows faster than the other dimensions, or (ii) X has a simple structure. These considerations are discussed at length in the extended version of the article.

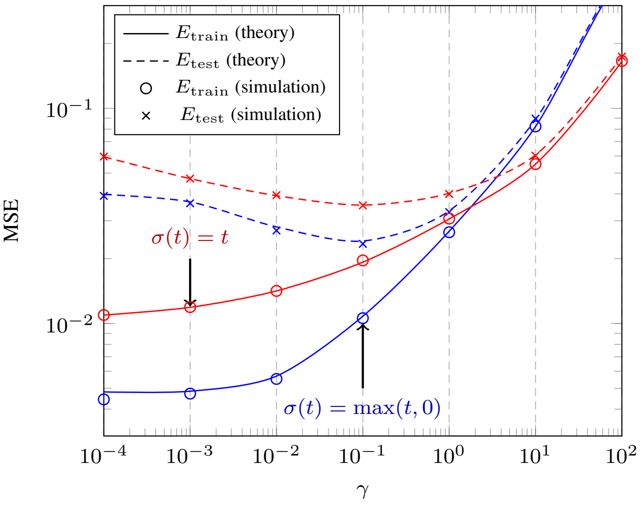

Fig. 3. Neural network performance for $\sigma(t) = t$ and $\sigma(t) = \max(t, 0)$, as a function of γ, for 2-class MNIST data (zeros, ones), $n = 512$, $T = 1024$, $p = 784$.

Since the main technical difficulty, related to the nonlinearity of σ, is now covered, it appears not a grand task to generalize our present random matrix framework to multiple hidden layers, thereby opening up the possibility to partially answer the open problem of network dimensioning. Subsequently, or rather in parallel, generalizing our results to incorporate a few steps of back-propagation of the error so to update the matrix W seems reachable and may notably provide insights within the underlying mechanisms of neural network learning. Further, the combination of the present findings with the study of linear echo-state networks (the recurrent version of ELMs) in [9] may allow for extensions to recursive network structures. These considerations, we believe, might open up the road to a new (random matrix-based) angle of investigation of neural networks.

5. REFERENCES

[1] 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.

[2] 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.

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

[4] L. Pastur and M. Ŝerbina, *Eigenvalue distribution of large random matrices*. American Mathematical Society, 2011.

[5] N. El Karoui, "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond," *The Annals of Applied Probability*, vol. 19, no. 6, pp. 2362–2405, 2009.

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

[7] J. W. Silverstein and Z. D. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 175–192, 1995.

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

[9] R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali, "The asymptotic performance of linear echo state neural networks," *(submitted to) Journal on Machine Learning Research*, 2016.