A Random Matrix and Concentration Inequalities framework for Neural Networks Analysis
Cosme Louart, Romain Couillet

To cite this version:
Cosme Louart, Romain Couillet. A Random Matrix and Concentration Inequalities framework for Neural Networks Analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Apr 2018, Calgary, Canada. 10.1109/ICASSP.2018.8462001. hal-01962077

HAL Id: hal-01962077
https://hal.archives-ouvertes.fr/hal-01962077
Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 \( f : \mathbb{R}^p \rightarrow \mathbb{R}^q \) 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 \( n, p \rightarrow \infty \)) 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 \( n, p \rightarrow \infty \)). Exploiting Features (i) and (ii) together, and therefore working on concentrated input data in the first place, brings us to a more...
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 $C_l, c_l > 0$ 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 modellings 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 $Z \in \alpha(\cdot)$ if, for $Z'$ an independent copy of $Z$,

$$P(|Z - Z'| \geq t) \leq \alpha(t).$$

In particular, $Z$ is normally concentrated, denoted $Z \in C_{\alpha_N}(\sqrt{c} \cdot)$, if $Z \in C e^{-ct^2}$, and $Z$ is exponentially concentrated, denoted $Z \in C_{\alpha_{exp}}(c \cdot)$, if $Z \in C e^{-c \cdot}$.

Normal and exponential concentrations owe their names to the fact that normal and exponential random variables are respectively normally and exponentially concentrated. The fast decay of their tails allows for additional properties. In particular, both concentrate around their means in the sense that, e.g., $Z \in C_{\alpha_N}(\sqrt{c} \cdot) \Rightarrow P(|Z - EZ| \geq t) \leq C' e^{-ct^2}$.

Also, these fast concentrations induce moment controls:

- $Z \in C_{\alpha_N}(\sqrt{c} \cdot) \Rightarrow \forall r > 0, E[|Z - EZ|^r] \leq C'(2r/c)^{r/2}$
- $Z \in C_{\alpha_{exp}}(c \cdot) \Rightarrow \forall r > 0, E[|Z - EZ|^r] \leq C'(2r/c)^r$.

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

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

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

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).** If $Z \in C_{\alpha_N}(\sqrt{c} \cdot)$, then, with obvious notations,

$$Z^2 \in C' \alpha_{exp}\left(\frac{c}{2} \cdot\right) + C'' \alpha_N\left(\frac{\sqrt{c}}{4|EZ|} \cdot\right).$$

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 $Z \in \alpha(\cdot)$, if for every $1$-Lipschitz map $f : \mathbb{R}^p \to \mathbb{R}$, $f(Z) \in \alpha(\cdot)$. Normally and exponentially distributed vectors are defined similarly.

A non-trivial result [8, Prop. 1.9.] 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 $p$. Precisely, $Z \sim N(0, I_p) \Rightarrow Z \in 2\alpha_N(\cdot / \sqrt{2})$.

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 large while $d$ remains small irrespective of $p$). Our core assumption is that there exists $k$ measures $\mu_1, \ldots, \mu_k$ such that, for each $l \leq n, x_l \sim \mu_l$ for some $\ell \leq k$; besides all $x_l$ are independent and

$$Z \sim \mu_l \Rightarrow Z \in C_{l} \alpha_{N}(\sqrt{c_l} \cdot).$$

For $Z \sim \mu_l$, we denote $\bar{M}_l \equiv E[Z]$ (and $\bar{M} = [\bar{M}_1, \ldots, \bar{M}_k]$) and $\bar{C}_l \equiv E[ZZ^T]$ (not to be confused with the covariance matrix). To avoid technicalities, we assume that $C < \max_{1 \leq \ell \leq k} \{\frac{1}{p} tr \bar{C}_l \} < C'$ for $C, C' > 0$ independent of $p$. We finally denote $n_l$ the number of $x_i$’s drawn from $\mu_l$.

This set of hypotheses is of interest as it notably encompasses the cases where:

1. the $x_i$’s arise from a Gaussian mixture model where $\mu_l = N(m_l, \Sigma_l)$;
2. the $x_i$’s are the output of a random feature map $x_i = \sigma(W s_l)$ 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 entry-wise), and $s_l \in \mathbb{R}^q$ such that $s_l \in C_{l} \alpha_{N}(\sqrt{c_l} \cdot)$ (for instance $s_l \sim N(m_l, \Sigma_l)$).

---

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.
Item 2 justifies our claim that, under this setting, simple random neural networks analysis reduces to the analysis of concentrated random vectors, disregarding the specificities of the non-linear activation function since $s_t \in C_{\Theta N}(\sqrt{c_t}) \Rightarrow x_t \in C_{\Theta N}(\sqrt{c_t})$ and that only this concentration property will be effectively used in the forthcoming analysis.

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 $x_t = \sigma(Ws_t)$ for some input observations $s_1, \ldots, s_n$. Since linear ridge-regression involves as a core object the sample covariance matrix

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

with $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$, 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 $X$, amongst 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

$$Q_X = (C_X + zI_p)^{-1}$$

of $C_X$, for $z > 0$. The matrix $Q_X$ encapsulates much information about $C_X$; notably, the so-called Stieljes transform $\frac{1}{p} \text{tr} Q_X$ 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 $a^T Q_X a$ allow for a characterization of the projections $|a^T u_i(C_X)|$ 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 $2/\sqrt{\pi z}$-Lipschitz so that $X \in \alpha(\cdot) \Rightarrow Q_X \in \alpha(\sqrt{\pi / 2 z})$ (with $X$ and $Q_X$ respectively seen as vectors in $\mathbb{R}^{np}$ and $\mathbb{R}^p$).

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

Theorem 1. Let $Q$ be defined as

$$Q = \left( \sum_{i=1}^{k} \frac{n_t}{n} \frac{\bar{C}_t}{1 + \delta_t} + zI_p \right)^{-1}$$

where $(\delta_1, \ldots, \delta_k) \in \mathbb{R}_+^k$ is the unique solution with nonnegative elements of the system $\delta_t = \frac{1}{n} \text{tr} \bar{C}_t Q$, $t = 1, \ldots, k$.

Then, for any unit-norm $u \in \mathbb{R}^p$, the following non-asymptotic inequalities hold

$$|\|E[QX] - \bar{Q}\| | \leq \frac{C' \sqrt{n} \max(1, \frac{p}{n})}{\sqrt{n}}$$

$$P \left( \frac{1}{n} \text{tr} (QX - \bar{Q}) \right) \geq t + \sqrt{\frac{C' \max(1, \frac{p}{n})}{n}} \leq C e^{-cnp t^2}$$

$$P \left( |u^T (QX - \bar{Q}) u| \right) \geq t + \sqrt{\frac{C' \max(1, \frac{p}{n})}{n}} \leq C e^{-cnp^2}. $$

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

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

where, for any $x_i \sim \mu$, we introduced the matrices

$$\varepsilon_{\ell}^1 \equiv \frac{Q_{X_{i-t}}x_i^T Q_{X_{i-t}}(x_i^T - \frac{1}{n} x_i^T Q_{X_{i-t}})}{(1 + \frac{1}{n} x_i^T Q_{X_{i-t}})(1 + \delta_t)}$$

$$\varepsilon_{\ell}^2 \equiv \frac{Q_{X_{i-t}}x_i^T \bar{C}_t Q_{X_{i-t}}}{n(1 + \delta_t)}$$

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

$$x_i^T Q_{X_{i-t}}x_i - n\delta_t = (x_i^T Q_{X_{i-t}} - \text{tr} \bar{C}_t Q_{X_{i-t}}) + \text{tr} \bar{C}_t (Q_{X_{i-t}} - E[Q_{X_{i-t}}]) + (\text{tr} \bar{C}_t Q_{X_{i-t}} - n\delta_t).$$

Since $\|Q_{X_{i-t}}\| \leq z^{-1}$ and $x_i \in C_{\Theta N}(\sqrt{c_t})$, Lema 1 applied to $x_i^T Q_{X_{i-t}}x_i = \|Q_{X_{i-t}}x_i\|^2$ 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 (deterministic) term, its control follows from pre-established random matrix results (see e.g., [15]). Since normal and exponential concentrations convert to bounds on moments, $\text{E}[\varepsilon_{\ell}^1]$ 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 $\beta \in \mathbb{R}^{n \times d}$ that minimizing, for $z > 0$, the cost

$$E_{\text{train}}(\beta) \equiv \frac{1}{n} \sum_{i=1}^{n} ||y_i - \beta^T x_i||^2_F + z||\beta||^2_F.$$
(with $\|\cdot\|_F$ the Frobenius norm). Letting $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times d}$, the solution is explicitly given by $\beta^* = \frac{1}{n} X^T Y$. For a test data pair $(x, y) \in \mathbb{R}^n \times d$, the regression output of $x$ is then given by

$$S(x) = x^T \beta^* = \frac{1}{n} x^T X Y.$$ 

For $k$-class classification purposes, one naturally takes $y_\ell = e_\ell \in \mathbb{R}^k$ when $x_t \sim \mu_\ell$ (hence $d = k$), with $[e_\ell]_a = \delta_{a\ell}$ the indicator vector of class $\ell$. 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 2.** Let $x \sim \mu_\ell$ and assume $\max_{1 \leq \ell \leq n} \|y_\ell\| < C$. Then, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$,

$$V_\ell^{-\frac{1}{2}} (S(x) - \bar{S}_\ell) \xrightarrow{d} N(0, I_d),$$

where $V_\ell = V_t - \bar{S}_\ell S_t^T$.

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

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

where $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$ with $[j_\ell]_k = \delta_{\ell \ell}$, $\Phi = (I_k - \bar{\Phi})^{-1} \bar{\Phi} \in \mathbb{R}^{k \times k}$, $\bar{M}_t = (I_k - \bar{\Phi})^{-1} \bar{\Phi} \in \mathbb{R}^{k \times k}$, $D^\ell = \text{diag}(\bar{\Phi}_{\ell,\ell}) \in \mathbb{R}^{k \times k}$, $D^\ell \bar{\Phi} = \text{diag}(\frac{n}{n^2} (1 + \delta_\ell)(1 + \delta_\ell)) \in \mathbb{R}^{k \times k}$, $\Delta = \text{diag}((1 + \delta_\ell)^{-1})$, and $\Theta = M_t^T \bar{Q} \bar{M}$.

$$\bar{\Phi} = \left. \left( \frac{M_t^T \bar{Q} \bar{M} \bar{M}}{1 + \delta_\ell} \right) \right|_{1 \leq i, j \leq k}$$

$$\bar{\Psi} = \left. \left( \frac{n_\ell \text{tr} \bar{Q} \bar{C}_\ell \bar{Q}_\ell}{n^2(1 + \delta_\ell)(1 + \delta_\ell)} \right) \right|_{1 \leq i, j \leq k}.$$

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

As discussed previously, letting $x_t = \sigma(W S_t)$ for $S_t \in \mathbb{R}^q$ with $q \sim p$ such that $S_t \in C_1 \mathcal{O} \mathcal{N}(\sqrt{c} \mathcal{E} \cdot), \sigma(\cdot)$ 1-Lipschitz and $W \in \mathbb{R}^{p \times q}$ with $\|W\| \leq 1$ (thereby ensuring that $x_t \in C_1 \mathcal{O} \mathcal{N}(\sqrt{c} \mathcal{E} \cdot), S(x)$ is the output of an ELM. The asymptotic statistics of $S(x)$ in Theorem 2 therefore directly translate in terms of simple neural network performances.

Fig. 1 provides the simulated $S(x) \in \mathbb{R}^2$ versus theoretical average $S_t$ output of an ELM for 2-class $(C_1, C_2)$ classification with $\sigma(t) = \text{ReLU}(t) = \max(t, 0)$ and $\sigma(t) = \text{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-u^2} du$, $W$ random unitary, and $S_t$ extracted from the MNIST handwritten digits dataset [19]. Here $Y \in \mathbb{R}^{n \times k}$ is defined by $Y_{t,\ell} = \delta_{s_t \in C_{\ell_\ell}}$. 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 $S_t \sim \mathcal{N}(0, \sigma_t I_p)$ with $\sigma_t \in \{1, 2\}$ according to the class. Being non-linearly separable classes, a straightforward application of Theorem 2 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 $\sigma$, for all $t, M_t = \mathbb{E}[\sigma(W S_t)] = 0$ where $S_t \sim \mathcal{N}(0, \sigma_t I_p)$, and we thus find that $S_t = 0$ indeed not allowing for class discrimination.
4. REFERENCES


