A Random Matrix and Concentration Inequalities framework for Neural Networks Analysis

Cosme Louart[1], Romain Couillet[1,2]

[1]LSS, CentraleSupélec, Université ParisSaclay, France.
[2]GSTATS DataScience Chair, GIPSA-lab, Université Grenoble Alpes, France.

Abstract

Context:
- Classification task performed by **non-linear** one-layer feed-forward neural net.
- **Theoretical asymptotic performances** for large dimensions (data size, number of data and neurons).

Objective:
- Formalize "data regularity".
- Introduce efficient framework for neural net understanding.

Results:
- **Concentration of measure** as solution to understand neural net output stability.
- **Theoretical formulas for asymptotic classification error**.

I – Concentration of measure basics

Definition (Concentration of a random vector)

$(E, \|\cdot\|)$ normed vector space ; $Z \in E$ random vector ; $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$; $f : E \to \mathbb{R}$:

With inequality $\mathcal{I}(X) \Leftrightarrow \mathbb{P}(|f(Z) - X| \geq t) \leq \alpha(t)$, define:

- ***Lipschitz* concentration** $Z \propto \alpha$: $\mathcal{I}(f(Z'))$ true for any f 1-Lipschitz with Z' independent copy of Z.
- ***Convex* concentration** $Z \propto_c \alpha$: $\mathcal{I}(f(Z'))$ true for any f 1-Lipschitz *and convex*.
- ***Linear* concentration** around **deterministic equivalent** $\tilde{Z} \in E$ $Z \in \tilde{Z} \pm \alpha$: $\mathcal{I}(f(\tilde{Z}))$ true for f linear s.t. $\|f\| = \sup_{\|x\| \leq 1} |f(x)| \leq 1$.

Theorem (Concentration of some random vectors $Z \in \mathbb{R}^p$)

- ***Gaussian vectors***: $Z \sim \mathcal{N}(0, I_p) \implies Z \propto Ce^{-(\cdot/c)^2}$
- ***Bounded vectors***: $Z_i \in [a, b]$ *i.i.d.* $\implies Z \propto_c Ce^{-(\cdot/c)^2}$.

Proposition (Operations on concentrated random $Z \in \mathbb{R}$)

For $Z_1, Z_2 \in \mathbb{R}$ such that $Z_1, Z_2 \propto Ce^{-(\cdot/c)^2}$:

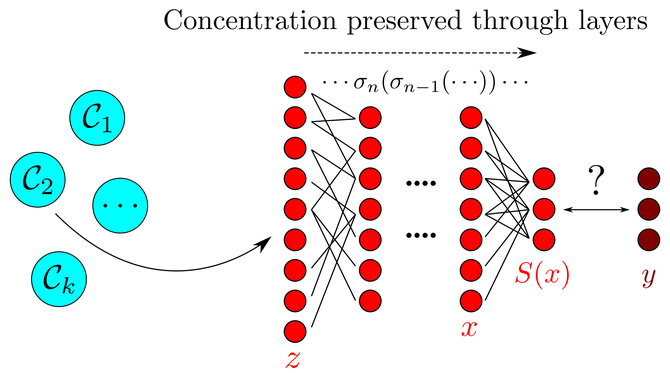
- $Z_1 + Z_2 \propto C'e^{-(\cdot/c')^2}$
- $Z_1^2 \propto C'e^{-(\cdot/c'\mathbb{E}Z_1)^2} + C'e^{-\cdot/c'^2}$
- *If* $Z_1 \leq K$: $Z_1 Z_2 \propto C'e^{-(\cdot/c'(K+\mathbb{E}Z_2))^2} + C'e^{-\cdot/c'^2}$

Implies **Hanson-Wright** inequality: for $x \propto_c Ce^{-(\cdot/c)^2}$:

$$x^T A x = \|A^{\frac{1}{2}}x\|^2 \propto C'e^{-(\cdot/c'\|A\|\mathbb{E}\|x\|)^2} + C'e^{-\cdot/c'^2\|A\|^2}$$

II – System Model

A feed-forward neural network



- $(z_1, y_1), \ldots, (z_n, y_n) \in \mathbb{R}^q \times \{0, 1\}^k$, data-label pairs from $\mathcal{C}_1, \ldots, \mathcal{C}_k$.
- $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ output of previous-to-last layer.
- Regression $\beta \in \mathbb{R}^{p \times k}$, min of $E_{\text{train}}(\beta) \equiv \frac{1}{n}\sum_{i=1}^n \|y_i - \beta^T x_i\|_F^2 + \eta\|\beta\|_F^2$,

$$\beta = \frac{1}{n} Q_X X Y^T, \quad Q_X \equiv \left(\frac{1}{n} X X^T + \eta I_p\right)^{-1}, \quad Y \equiv [y_1, \ldots, y_n] \in \mathbb{R}^{k \times n}.$$

- For $x \in \mathbb{R}^p$, neural net output $\boxed{S(x) \equiv \beta^T x = \frac{1}{n} Y X^T Q_X x \in \mathbb{R}^k}$.

Hypotheses

- $n \to \infty$ and $p = O(n)$.
- $x_i = f(\zeta_i) + g(\xi_i)$ where:
  - $\zeta_i \in \mathbb{R}^p$ Gaussian, $\xi_i \in \mathbb{R}^p$ with i.i.d. bounded entries; all independent
  - $f : \mathbb{R}^p \to \mathbb{R}^p$ Lipschitz, $g : \mathbb{R}^p \to \mathbb{R}^p$ affine.

$$\left(\implies X \propto_c Ce^{-\cdot^2/c} \quad \text{in } (\mathbb{R}^{p \times n}, \|\cdot\|_F) \right)$$

- For $\bar{x}_\ell = \mathbb{E}[x]$ with $x \in \mathcal{C}_\ell$, $\|\bar{x}_\ell\| = O(\sqrt{p})$.

III – Main Results

Proposition (Deterministic equivalent of Q)

Let $\Sigma_\ell = \mathbb{E}[xx^T]$, for $x \in \mathcal{C}_\ell$, and $\tilde{Q}_\gamma \equiv \left(\sum_{\ell=1}^k \frac{\#\mathcal{C}_\ell}{n} \frac{\Sigma_\ell}{1+\gamma_\ell} + \eta I_p\right)^{-1}$.

Then, the system $\left\{\delta_\ell = \frac{1}{n}\text{tr}(\Sigma_l \tilde{Q}_\delta)\right\}_{\ell=1}^k$ admits a unique solution and

$$Q_X \in \tilde{Q}_\delta \pm Ce^{-(\cdot)^2/c} \text{ in } (\mathbb{R}^{p \times p}, \|\cdot\|_{Sp}).$$

Theorem (Central limit theorem for $S(x)$)

$$S(x) \propto Ce^{-t^2/c} \quad \text{and} \quad \mathcal{V}_\ell^{-\frac{1}{2}}\left(S(x) - \tilde{S}_\ell\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_k)$$

where, for $\Delta = \text{diag}((1+\delta_\cdot)^{-1})$, $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$, $[j_\ell]_i = \delta_{x_i \in \mathcal{C}_\ell}$,

$$\tilde{S}_\ell \equiv \frac{1}{n} Y J \Delta \bar{X}^T \tilde{Q} \bar{x}_\ell,$$

$$\mathcal{V}_\ell = h(\delta, \Sigma_1, \ldots, \Sigma_k) \text{ for some } h \text{ (see article).}$$

Application: extreme learning machines $x = \sigma(Wz)$, W random

- **MNIST**: good **match with concentration predictions**:

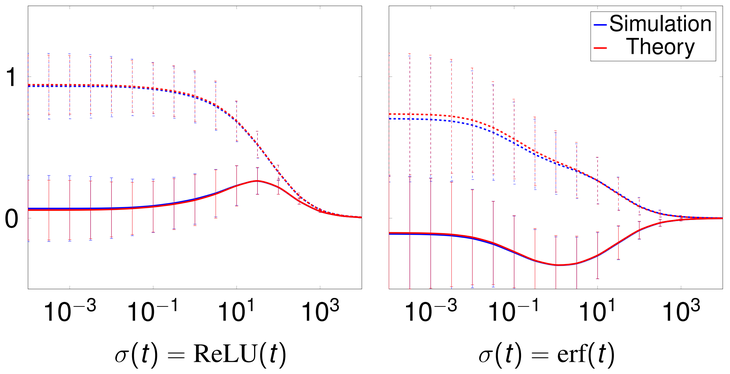


Figure: $[S(x)]_1$ (blue dotted) and $[S(x)]_2$ (blue solid) of 2-class MNIST ELM as a function of $\eta \in [10^{-4}, 10^4]$ (digits 3 for $\mathcal{C}_1$ and 8 for $\mathcal{C}_2$) for $x \in \mathcal{C}_1$, versus theory (red). Here $n = 2048$, $p = q = 784$, W random unitary.

- For symmetrical distribution, no classification if $\sigma(-z) = \sigma(z)$

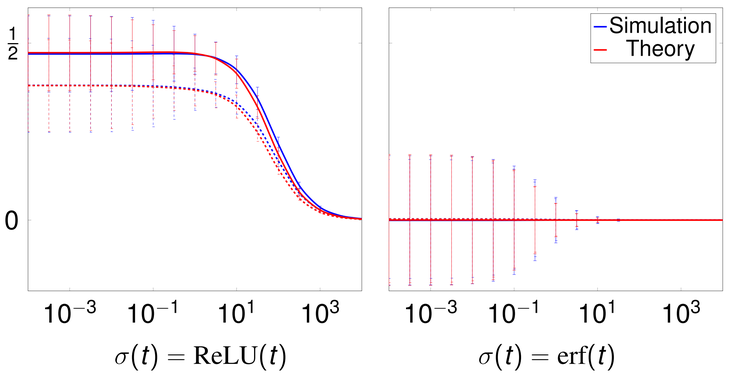


Figure: $[S(x)]_1$ (blue dotted) and $[S(x)]_2$ (blue solid) of Gaussian 2-mixture ELM ($\mathcal{C}_1 \equiv \mathcal{N}(0, I_p)$ and $\mathcal{C}_2 \equiv \mathcal{N}(0, 2I_p)$) for $x \in \mathcal{C}_1$, versus theory (red). Here $n = 4096$, $p = q = 256$, W random unitary.

Email: cosme.louart@ens.fr, romain.couillet@centralesupelec.fr