A Random Matrix and Concentration Inequalities framework for Neural Networks Analysis

Cosme Louart\(^1\), Romain Couillet\(^1,2\)

\(^1\)LSS, CentraleSupélec, Université Paris-Saclay, 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)\(^\star\)**

\([E, |||\cdot|||]\) normed vector space, \(Z \in E\) random vector, \(a : \mathbb{R} \to \mathbb{R}\), \(f : E \to \mathbb{R}\). With inequality \(I(X) \Leftrightarrow P(\{f(Z) - X \geq t\}) \leq o(t)\), define:

- **Lipschitz concentration** \(Z \sim \alpha \Rightarrow I(f(Z))\) for any \(f\) 1-Lipschitz with \(Z\) independent copy of \(Z\).
- **Convex concentration** \(Z \sim \alpha \Rightarrow I(f(Z))\) for any \(f\) 1-Lipschitz and convex.
- **Linear concentration** around deterministic equivalent \(Z' \in E\) \(Z \in Z + \alpha : I(f(Z'))\) true for \(f\) linear s.t. |||\(f(x)\)||| \leq 1.

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

- **Gaussian vectors:** \(Z \sim \mathcal{N}(0, I_p) \Rightarrow Z \sim Ce^{-\frac{1}{2}\tau^2}\).
- **Bounded vectors:** \(Z_i \in [a, b] \text{i.i.d.} \Rightarrow Z \sim Ce^{-\frac{1}{2}\tau^2}\).

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

For \(Z_1, Z_2 \in \mathbb{R}\) such that \(Z_1, Z_2 \sim Ce^{-\frac{1}{2}\tau^2}\):
- \(Z_1 + Z_2 \sim Ce^{-\frac{1}{2}\tau^2}\).
- \(Z_1 Z_2 \sim Ce^{-\frac{1}{2}\tau^2} + Ce^{-\frac{1}{2}\tau^2}\).
- If \(Z_1 \leq K\), \(Z_2 \sim Ce^{-\frac{1}{2}\tau^2} + Ce^{-\frac{1}{2}\tau^2}\).

Implies Hanson-Wright inequality: for \(x \sim Ce^{-\frac{1}{2}\tau^2}\), \(x^T A x \sim Ce^{-\frac{1}{2}\tau^2}\).

---

### II – System Model

**A feed-forward neural network**

Concentration preserved through layers.

**Proposition (Deterministic equivalent of \(O\))**

Let \(\Sigma_i = \mathbb{E}[x_i x_i^T]\), for \(x \in C_t\), and \(\hat{O}_t = \left(\sum_{i=1}^k \frac{1}{\beta_i} \frac{X_i}{\beta_i} + \eta_i \right)^{-1}\).

Then, the system \(\{\hat{S}_t = \frac{1}{n} tr(\hat{O}_t)\}^k\) admits a unique solution and \(Q_x \in \hat{O}_t \sim Ce^{-\frac{1}{2}\tau^2}\) in \((\mathbb{R}^{p, p}, ||\cdot||_F)\).

---

### III – Main Results

**Proposition (Deterministic equivalent of \(\Sigma\))**

Let \(\Sigma_i = \mathbb{E}[x_i x_i^T]\), for \(x \in C_t\), and \(\hat{O}_t = \left(\sum_{i=1}^k \frac{1}{\beta_i} \frac{X_i}{\beta_i} + \eta_i \right)^{-1}\).

Then, the system \(\{\hat{S}_t = \frac{1}{n} tr(\hat{O}_t)\}^k\) admits a unique solution and \(Q_x \in \hat{O}_t \sim Ce^{-\frac{1}{2}\tau^2}\) in \((\mathbb{R}^{p, p}, ||\cdot||_F)\).

---

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

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

---

**Theorem (Central limit theorem for \(S(x)\))**

\[ S(x) \sim Ce^{-\frac{1}{2}\tau^2} \text{ and } \mathcal{V}_i = \frac{1}{n} Y_i^X Q_{x_{i}}. \]

where, for \(\Delta = \text{diag}((1 + \delta)^{-1}), J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}, [b_j] = \delta_{c \in C_t}, \]

\[ \mathcal{S}_i = \frac{1}{n} Y_i^X Q_{x_{i}}. \]

\[ \mathcal{V}_i = h(\delta, \Sigma_1, \ldots, \Sigma_k) \text{ for some } h \text{ (see article).} \]

---

**Simulations**

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