A Random Matrix Analysis of Learning with α-Dropout

Artemiss Workshop ICML 2020

MEA. Seddik[12*], R. Couillet[23], M. Tamaazousti[1]

[1]CEA List, France
[2]CentraleSupélec, L2S, France
[3]GIPSA Lab Grenoble-Alpes University, France

[*]http://melaseddik.github.io/

July 15, 2020

Abstract

Context:

- Study of a one-hidden-layer network with **α-Dropout**.

Motivation:

- Classical Dropout[1] corresponds to *zero-imputation*.
- *Zero-imputation* alter neural networks performances[2].

Results:

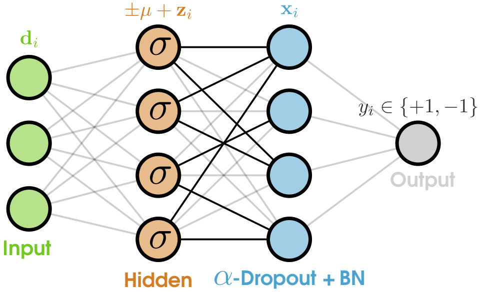
- **Asymptotic generalization performances** on a binary classification problem.
- An aftermath analysis exhibits $\alpha \neq 0$ which improves generalization.

[1] Srivastava et al., *Dropout: a simple way to prevent neural networks from overfitting*. JMLR 2014.
[2] Yi et al., *Why not to use zero imputation? correcting sparsity bias in training neural networks*. ICLR 2019.

Model and Problem Statement

Let $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n \in \mathbb{R}^q$ in two classes $\mathcal{C}_1$ and $\mathcal{C}_2$, and $\sigma : \mathbb{R}^q \to \mathbb{R}^p$ s.t. for $\boldsymbol{d}_i \in \mathcal{C}_a$

$$\mathbb{E}[\sigma(\boldsymbol{d}_i)] = (-1)^a \boldsymbol{\mu} \qquad \mathbb{E}[\sigma(\boldsymbol{d}_i)\sigma(\boldsymbol{d}_i)^\intercal] = \boldsymbol{I}_p + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$$



After the α-**Dropout layer and BN**, the features matrix $\boldsymbol{X}_{\alpha,\varepsilon} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathcal{M}_{p,n}$ is

$$\boldsymbol{X}_{\alpha,\varepsilon} = \frac{(\boldsymbol{B}_\varepsilon \odot (\boldsymbol{Z} + \boldsymbol{\mu}\boldsymbol{y}^\intercal))\, \boldsymbol{P}_n + \alpha \boldsymbol{B}_\varepsilon \boldsymbol{P}_n}{\sqrt{\varepsilon + \alpha^2 \varepsilon (1-\varepsilon)}}$$

with $[\boldsymbol{B}_\varepsilon]_{ij} \sim \mathrm{Ber}(\varepsilon)$, $Z_{ij} \sim \mathcal{N}(0,1)$ and $\boldsymbol{P}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal$.

Learning with α-Dropout

We consider the Ridge-classifier with ℓ_2-loss

$$\mathcal{E}(\boldsymbol{w}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}_{\alpha,\varepsilon}^{\intercal}\boldsymbol{w}\|^2 + \gamma\|\boldsymbol{w}\|^2$$

The solution of which is explicitly given by, for $z \in \mathbb{C} \setminus \mathbb{R}^-$

$$\boldsymbol{w} = \frac{1}{n}\boldsymbol{Q}(\gamma)\boldsymbol{X}_{\alpha,\varepsilon}\boldsymbol{y}, \qquad \boldsymbol{Q}(z) \equiv \left(\frac{1}{n}\boldsymbol{X}_{\alpha,\varepsilon}\boldsymbol{X}_{\alpha,\varepsilon}^{\intercal} + z\boldsymbol{I}_p\right)^{-1}$$

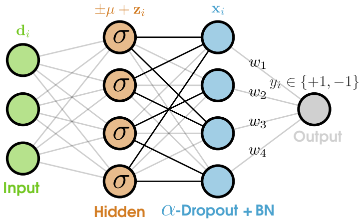
- The corresponding (hard) decision function is

$$g(\boldsymbol{x}) \equiv \boldsymbol{x}^{\intercal}\boldsymbol{w} = \frac{1}{n}\boldsymbol{x}^{\intercal}\boldsymbol{Q}(\gamma)\boldsymbol{X}_{\alpha,\varepsilon}\boldsymbol{y} \quad \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \quad 0$$

Assumptions (Growth rate)

As $n \to \infty$,

1. $\frac{q}{n} \to r \in (0,\infty)$ and $\frac{p}{n} \to c \in (0,\infty)$;
2. For $a \in \{1,2\}$, $\frac{n_a}{n} \to c_a \in (0,1)$;
3. $\|\boldsymbol{\mu}\| = \mathcal{O}(1)$.



Main Results

Deterministic equivalent of $\boldsymbol{Q}(z)$

Under the previous Assumptions,

$$\boldsymbol{Q}(z) \leftrightarrow \bar{\boldsymbol{Q}}(z) \equiv \mathcal{D}_z - \frac{\frac{\varepsilon}{1+\alpha^2(1-\varepsilon)} \mathcal{D}_z \boldsymbol{\mu}\boldsymbol{\mu}^\intercal \mathcal{D}_z}{1 + cq(z) + \frac{\varepsilon}{1+\alpha^2(1-\varepsilon)} \boldsymbol{\mu}^\intercal \mathcal{D}_z \boldsymbol{\mu}},$$

where $\mathcal{D}_z \equiv q(z)\mathrm{diag}\left\{ \frac{1+cq(z)}{1+cq(z)+\frac{(1-\varepsilon)q(z)}{1+\alpha^2(1-\varepsilon)}\mu_i^2} \right\}_{i=1}^{p}$ with $q(z) \equiv \frac{c-z-1+\sqrt{(c-z-1)^2+4\,z\,c}}{2\,z\,c}$.

Gaussian Approximation of $g(\boldsymbol{x})$

Under the previous Assumptions, for $\boldsymbol{x} \in \mathcal{C}_a$ with $a \in \{1, 2\}$,

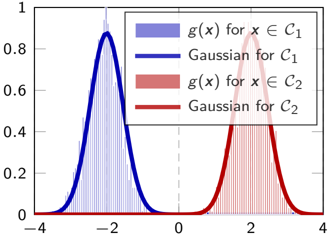
$$\nu^{-\frac{1}{2}} \left(g(\boldsymbol{x}) - m_a \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

where

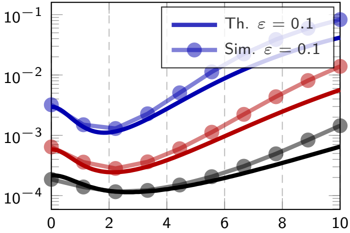
$$m_a \equiv (-1)^a \sqrt{\frac{\varepsilon}{1+\alpha^2(1-\varepsilon)}} \frac{\boldsymbol{\mu}^\intercal \bar{\boldsymbol{Q}}(\gamma)\boldsymbol{\mu}}{1+\delta(\gamma)}$$

$$\nu \equiv \frac{1}{(1+\delta(\gamma))^2} \left(\eta(\boldsymbol{C}_1) + \frac{\varepsilon}{1+\alpha^2(1-\varepsilon)} \times \left[\boldsymbol{\mu}^\intercal \left(\boldsymbol{\Delta}(\boldsymbol{C}_1) - \bar{\boldsymbol{Q}}(\gamma) \right) \boldsymbol{\mu} - \frac{2\,\eta(\boldsymbol{C}_1)\boldsymbol{\mu}^\intercal \bar{\boldsymbol{Q}}(\gamma)\boldsymbol{\mu}}{1+\delta(\gamma)} \right] \right)$$

Take Away Messages



Test (generalization) scores $g(\boldsymbol{x})$



Test Error in terms of α

Highlights:

- Existence of $\alpha \neq 0$ which minimizes **the test error**.
- In our setting, such α satisfies $\frac{1}{m_a}\frac{\partial m_a}{\partial \alpha} = \frac{1}{\sqrt{\nu}}\frac{\partial \sqrt{\nu}}{\partial \alpha}$.

Perspectives:

- Extend the analysis to a k-**class model** with α_ℓ's for each class.
- Validation of the α-Dropout approach with **real data**.
- Extend to **multi-layers** networks.