A two-scale Complexity Measure for Deep Learning Models

Massimiliano Datres[1,2], Gian Paolo Leonardi[1], Alessio Figalli[3], David Sutter[4]

[1] University of Trento, [2] Bruno Kessler Foundation, [3] IBM Research Zurich, [4] ETH Zurich

Introduction & Contribution

- Neural Networks (NN) achieve outstanding performances in solving complex tasks such as image classification problems, object detection;

- Quantify expressivity pre-training ➔ **complexity measures**

- A good complexity measure for NN should:

  1. Give **useful pre-training** information;
  2. Be more **efficient/scalable** than full training;
  3. Be applicable **to over-parametrized regimes**;
  4. Provide **insights about generalization**.

Introduction & Contribution

- Neural Networks (NN) achieve outsta
  classification problems, object detec

- Quantify expressivity pre-traini

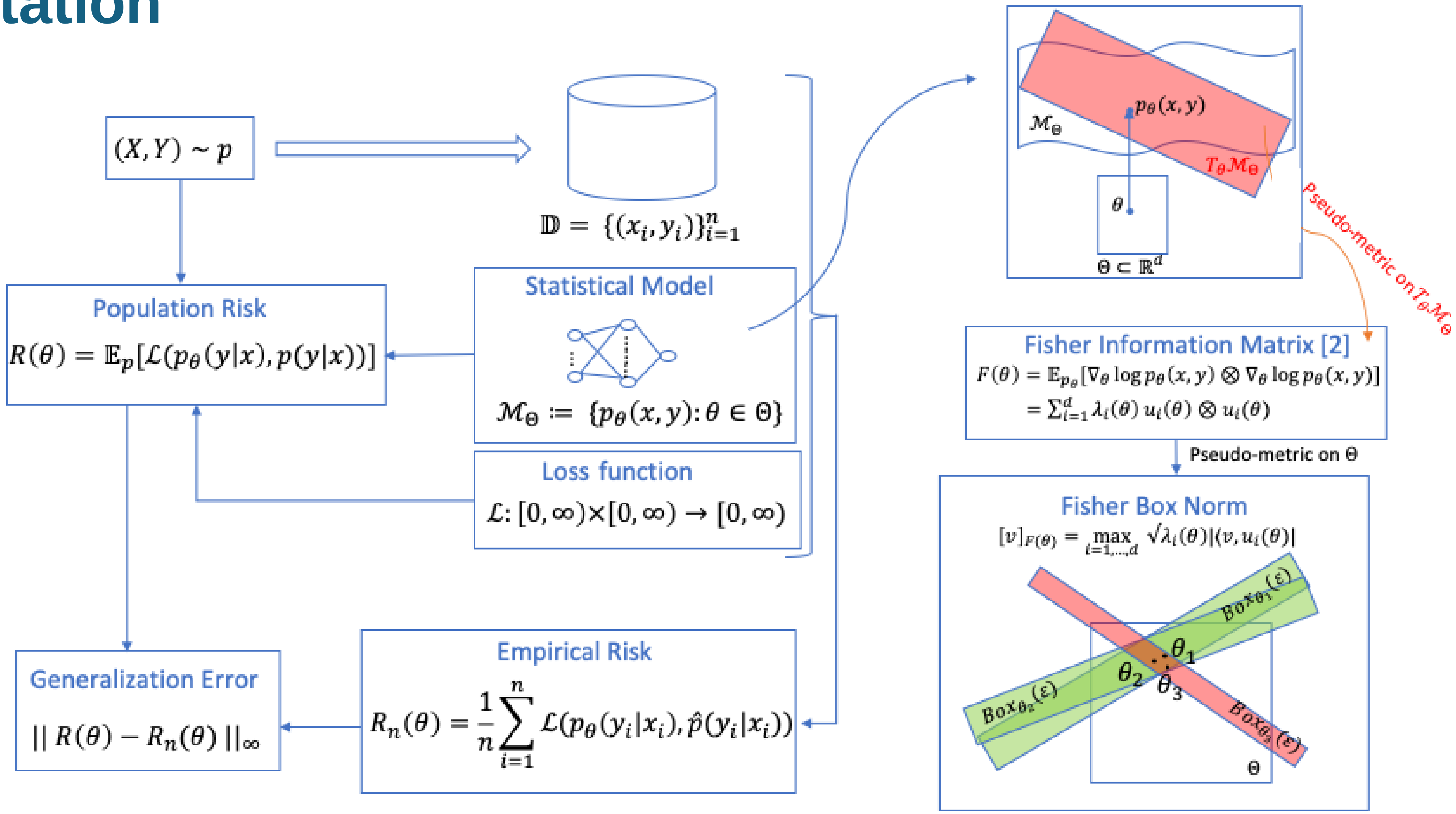
- A good complexity measure for NN

Our contribution

- A new complexity measure, the two-scale effective dimension (2sED) satisfying (1), (3), (4);
- Approximation for Markovian models satisfying (2);
- Empirical validation of (1), (2), (3).

1. Give **useful pre-training** information;
2. Be more **efficient/scalable** than full training;
3. Be applicable **to over-parametrized regimes**;
4. Provide **insights about generalization**.

Notation



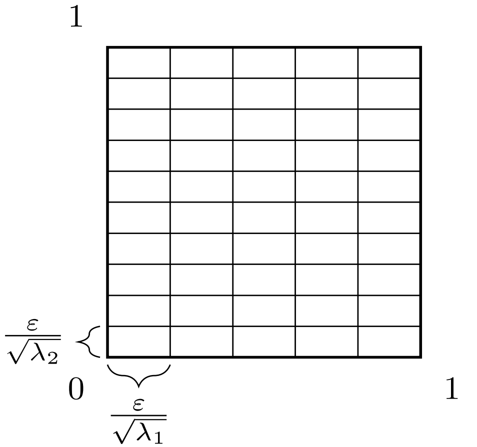
$(X, Y) \sim p$
$\mathbb{D} = \{(x_i, y_i)\}_{i=1}^n$
Population Risk
$R(\theta) = \mathbb{E}_p[\mathcal{L}(p_\theta(y|x), p(y|x))]$
Statistical Model
$\mathcal{M}_\Theta := \{p_\theta(x, y): \theta \in \Theta\}$
Loss function
$\mathcal{L}: [0, \infty) \times [0, \infty) \to [0, \infty)$
Generalization Error
$\| R(\theta) - R_n(\theta) \|_\infty$
Empirical Risk
$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(p_\theta(y_i|x_i), \hat{p}(y_i|x_i))$
$p_\theta(x, y)$
$\mathcal{M}_\Theta$
$T_\theta \mathcal{M}_\Theta$
θ
$\Theta \subset \mathbb{R}^d$
Pseudo-metric on $T_\theta \mathcal{M}_\Theta$
Fisher Information Matrix [2]
$F(\theta) = \mathbb{E}_{p_\theta}[\nabla_\theta \log p_\theta(x, y) \otimes \nabla_\theta \log p_\theta(x, y)]$
$= \sum_{i=1}^{d} \lambda_i(\theta)\, u_i(\theta) \otimes u_i(\theta)$
Pseudo-metric on Θ
Fisher Box Norm
$[v]_{F(\theta)} = \max_{i=1,\dots,d} \sqrt{\lambda_i(\theta)}\, |\langle v, u_i(\theta)\rangle|$
$Box_{\theta_1}(\varepsilon)$
$Box_{\theta_2}(\varepsilon)$
$Box_{\theta_3}(\varepsilon)$
θ_1
θ_2
θ_3
Θ

The Effective Dimension

$$\mathrm{effdim}_{eff,\varepsilon}(\mathcal{M}_\Theta) \coloneqq \frac{\log \mathcal{N}_\theta(\varepsilon)}{|\log \varepsilon|}$$

where $\mathcal{N}_\theta(\varepsilon)$ is the **minimum number of Fisher boxes of size $\boldsymbol{\varepsilon}$** needed to cover Θ

Easiest Case

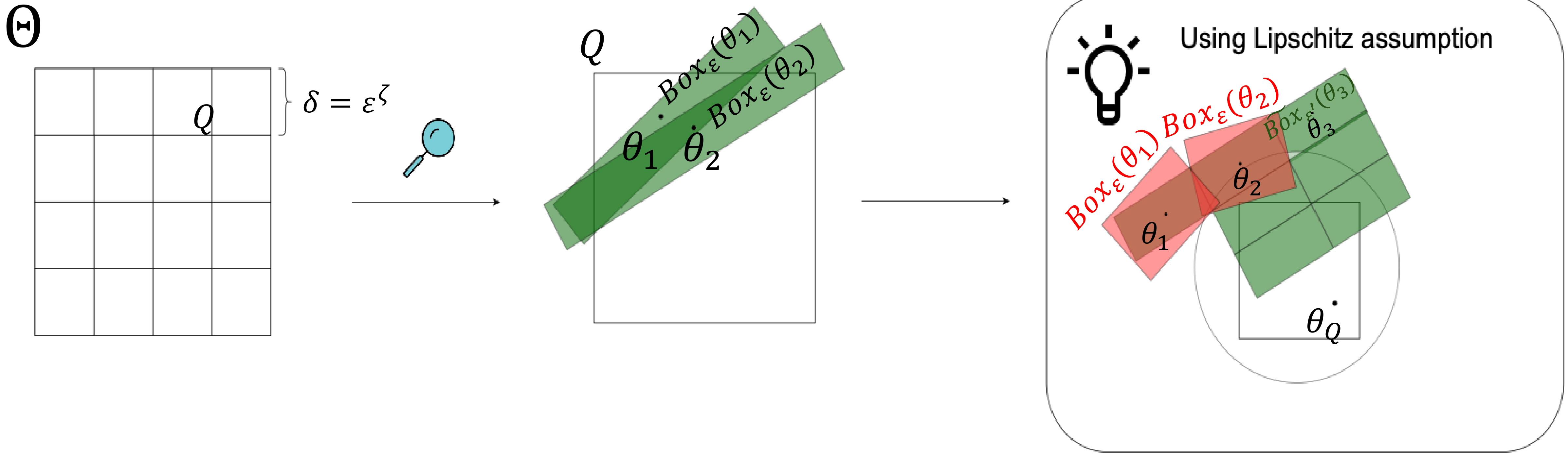


$\Theta = [0,1]^{\mathrm{d}}$ and $F = diag(\lambda_1, \ldots, \lambda_d)$:

$$\mathcal{N}_\theta(\varepsilon) \leq \prod_{i=1}^{d} \left\lceil \frac{\sqrt{\lambda_i}}{\varepsilon} \right\rceil \leq \det(Id + \varepsilon^{-1}\sqrt{F})$$

$where\ \lceil t \rceil = \min\{k \in \mathbb{Z}: \max(t, 1)\}$

Harder Case



The 2-scale Effective Dimension

Definition

Given $0 < \varepsilon < 1$ and $0 \leq \zeta < 1$, we define the ***two-scale effective dimension*** (or simply 2sED) as:

$$d_\zeta(\varepsilon) = \zeta d + (1-\zeta)\frac{\log \mathbb{E}_\theta[\det(I_d + \varepsilon^{\zeta-1}\, \hat{F}(\theta)^{1/2})]}{|\log \varepsilon^{\zeta-1}|}$$

where:

$$\hat{F}(\theta) = \begin{cases} \frac{d}{\mathbb{E}_\theta[Tr\, F(\theta)]} F(\theta) & if\ \mathbb{E}_\theta[Tr\, F(\theta)] > 0 \\ 0 & otherwise \end{cases}$$

Generalization Buond

Under suitable assumptions:

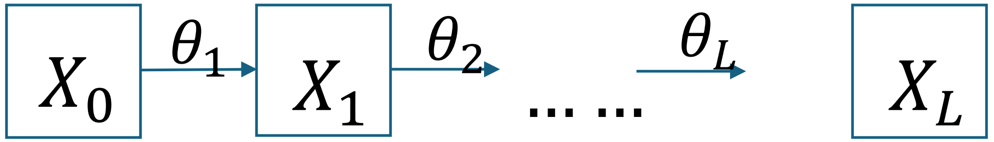
1. The model $\theta \to p_\theta$ is $C^{1,1}$ and $\exists\ 0 < \alpha_1 \leq \alpha_2$ such that $\alpha_1 \leq p, p_\theta \leq \alpha_2$;
2. The FIM $F(\theta)$ and the loss function $\mathcal{L}$ are bounded and Lipschitz;
3. The meso-scale exponent $\zeta \in \left[\frac{2}{3}, 1\right]$;

Theorem

Under assumptions (1), (2), (3), there exists $C, H, K, n_0 > 0$ such that $\forall \gamma \in (0,1], n \geq n_0$ and $\varepsilon_n = \left({}^{\log n}/_{\gamma n}\right)^{3/8}$:

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |R(\theta) - R_n(\theta)| \geq C\varepsilon_n\right) \leq H \varepsilon_n^{-d_\zeta(\varepsilon)} n^{-\frac{K}{\gamma}}$$

Markovian Models



$$p_\theta(x_0, x_1, \dots, x_L) = p_{\theta_L}(x_L | x_{L-1}) \cdots p_{\theta_1}(x_1 | x_0) p(x_0)$$

FIM Diagonal Block

$$F(\theta) = \begin{bmatrix} F_1(\theta_1) & 0 & \cdots & 0 \\ 0 & F_2(\theta_1, \theta_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & F_L(\theta_1, \dots, \theta_L) \end{bmatrix}$$

where:

$$F_j = F_j(\theta_1, \dots, \theta_j)$$

$$= \mathbb{E}_{x, p_{\theta_1}(x_1|x_0), \dots, p_{\theta_j}(x_j|x_{j-1})}\left[\int_{x_j} \underbrace{\left[\nabla_{\theta_j} l_{\theta_j}(x_j | x_{j-1})\right]^{\otimes 2}}_{\nabla_{\theta_j} \log p_{\theta_j}(x_j|x_{j-1}) \otimes \nabla_{\theta_j} \log p_{\theta_j}(x_j|x_{j-1})} p_{\theta_j}(dx_j | x_{j-1})\right]$$

Jensen

Lower 2sED - $d_\zeta(\varepsilon)$

$$d_\zeta^1(\varepsilon) = \zeta d + \log \int_{\Theta_1} \det 1 \, d\theta_1$$

$$\vdots$$

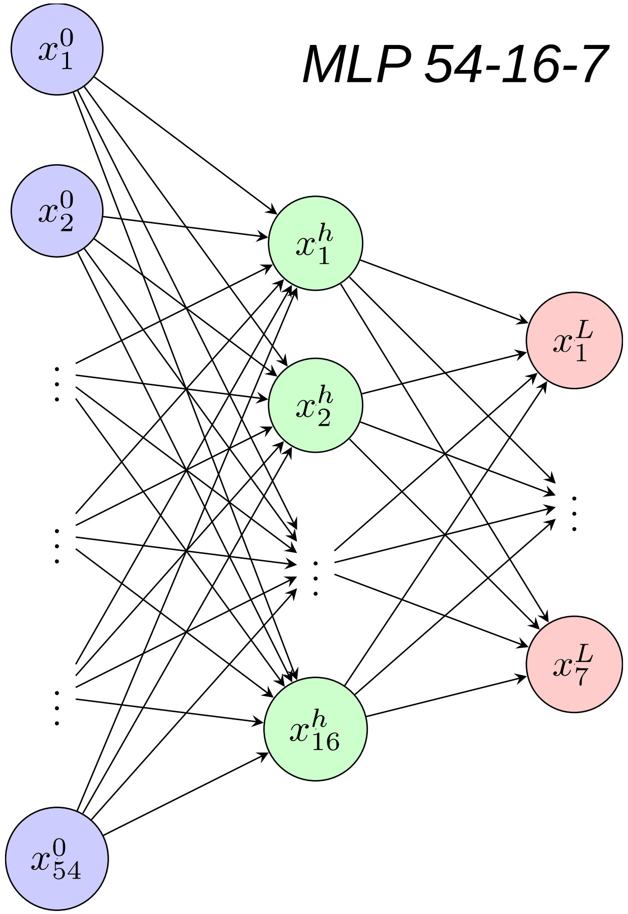
$$d_\zeta^m(\varepsilon) = d_\zeta^{m-1}(\varepsilon) + \int_{\widehat{\Theta}_m} \int_{\Theta_m} detm \, d\theta_m d\Phi_m$$

$$detm = \det(Id + \varepsilon^{-1} F_m)$$

$$\widehat{\Theta}_m = \Theta_1 \times \cdots \times \Theta_m$$

$$d\Phi_m = \frac{1}{\prod_{j=1}^{m-1} |\Theta_j|} \prod_{j=1}^{m-1} detj \, d\theta_1 \cdots d\theta_{m-1}$$

Selected Experiment



- $O_i(\cdot) \coloneqq act(W^i \cdot)$ where *act* is the activation function;
- *Stochasticity:*

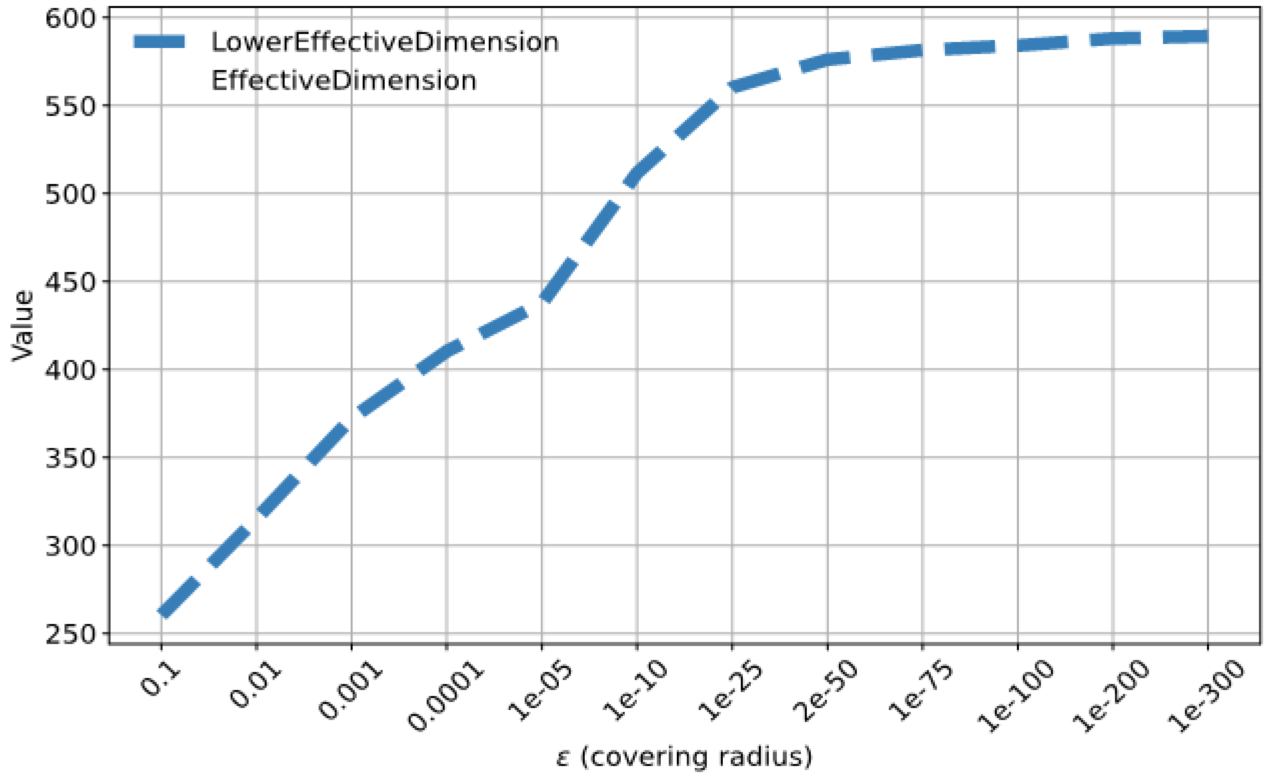
$$O_i^\sigma \coloneqq O_i + \nu_i \sim \mathcal{N}(O_i, \sigma^2 Id)$$

  where $\nu_i \sim \mathcal{N}(0, \sigma^2)$;
- Compare three models with similar amount of parameters and ReLU activation functions;

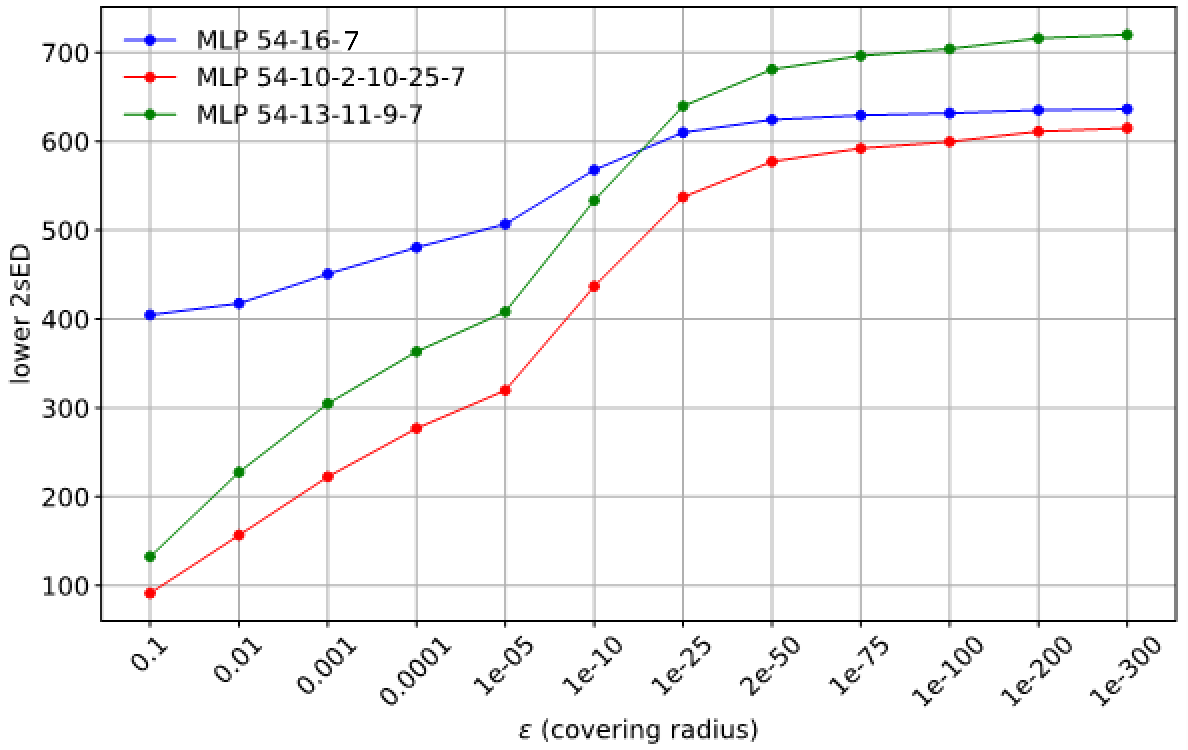
Model	Number of Parameters
MLP 54-16-7	976
MLP 54-13-11-9-7	1007
MLP 54-10-2-10-25-7	1005

CoverType Dataset [4]: Classification of pixels into 7 forest cover types based on attributes such as elevation, aspect, slope, hillshade, soil-type, and more.

Selected Experiment

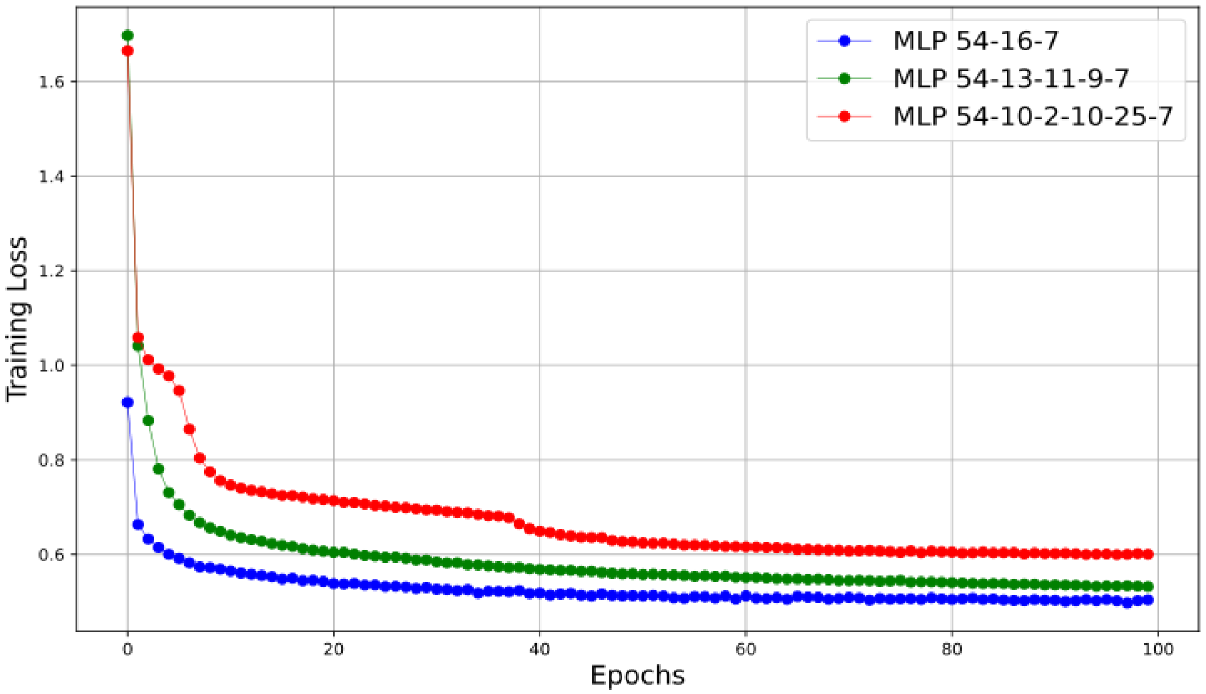


Comparison $d_\zeta(\varepsilon)$ and $d_\zeta(\varepsilon)$ for MLP 54-16-7

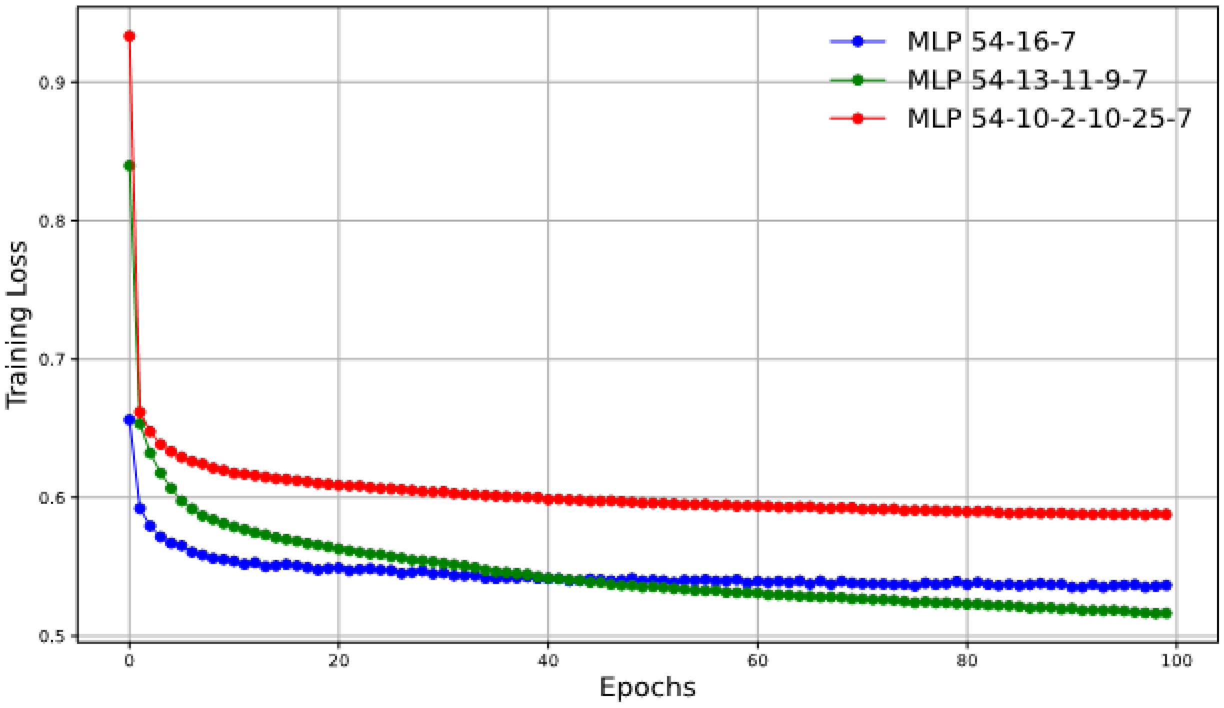


Estimated $\underline{d_\zeta(\varepsilon)}$ of three different MLP architectures using 100 Covertype samples and 100 different vectors of parameters for the Monte Carlo estimation of $\hat{F}_N$;

Selected Experiment



Training loss plots of MLPs on 10000 random CoverType samples using Adam with learning rate $1e^{-3}$ and a batch size 64;



Training loss plots of MLPs on 100000 random CoverType samples using Adam with learning rate $1e^{-3}$ and a batch size 64;

References

[1]Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*, *1*(6), 403-409.

[2] Liang, T., Poggio, T., Rakhlin, A., & Stokes, J. (2019, April). Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics* (pp. 888-896). PMLR.Marge

[3] Berezniuk, O., Figalli, A., Ghigliazza, R., & Musaelian, K. (2020). A scale-dependent notion of effective dimension. arXiv preprint arXiv:2001.10872.

[4] http://archive.ics.uci.edu/dataset/31/covertype

Thanks for the attention!