A TWO-SCALE COMPLEXITY MEASURE FOR DEEP LEARNING MODELS

MASSIMILIANO DATRES, GIAN PAOLO LEONARDI, ALESSIO FIGALLI, AND DAVID SUTTER

Abstract. We introduce a novel capacity measure 2sED for statistical models based on the effective dimension. The new quantity provably bounds the generalization error under mild assumptions on the model. Furthermore, simulations on standard data sets and popular model architectures show that 2sED correlates well with the training error. For Markovian models, we show how to efficiently approximate 2sED from below through a layerwise iterative approach, which allows us to tackle deep learning models with a large number of parameters. Simulation results suggest that the approximation is good for different prominent models and data sets.

1. Introduction

Deep learning models are achieving outstanding performances in solving several complex tasks such as image classification problems, object detection [KSH12, LBH15] and natural language processing [BMR+20]. The reason for the vast success of deep neural networks (DNNs) is mainly due to the technological achievements that have made possible to set up and train parametric models defined by a huge number of parameters (typically, much larger than the cardinality of the training datasets) [ZBH+21]. Over-parametrized regimes make DNNs able to extract valuable information from data. Quite surprisingly, and despite the fact that an abundance of parameters could in principle give rise to over-fitting problems, DNNs typically exhibit impressive generalization capabilities after training [ZBH+21, NTS14]. As the demand of computational resources and training time increases with the number of parameters, the trial-and-error procedures that are typically adopted for selecting the most appropriate models for a given task become extremely expensive, or even impractical, when the number of parameters is huge.

Finding appropriate complexity measures for deep learning models can help in understanding and quantifying their generalization capabilities. Hereafter we propose some essential features that, ideally, should characterize a complexity measure for parametric models:

(P1) it should provide pre-training information consistent with post-training performances;
(P2) its computation should be more efficient and scalable in comparison with a full training & validation process;
(P3) in the case of a feedforward-type model, it should be "modular", i.e., computable in some iterative fashion[1];

Notions of complexity measures have appeared in the context of machine learning, with early studies focusing, e.g., on the complexity of decision tree models [BFSO84] and logistic regression models [CSMV99, KST08, BMR19].

From the perspective of statistical learning theory, the Vapnik-Chervonenkis dimension, commonly called VC dimension [Vap99], is an established complexity measure defined in term of the largest number of points that can be shattered by a class of functions [HFT09]. This complexity dimension has been used to establish data-independent generalization bounds for statistical models [SSBD14, PL20].

2020 *Mathematics Subject Classification.* Primary: ADD Secondary: ADD.

Key words and phrases. ADD.

ADD.

[1]Here we are referring to the typical structure of a feedforward-type model, which is a composition of parametric layer maps.

Other notions of model complexity, specifically designed for deep learning models, have been more recently proposed, with the aim of quantifying the expressivity of a DNN [MPCB14, RPK$^+$17, LPRS19, HCP$^+$21].

There is, however, a supported evidence that data-independent generalization bound are not universally effective [BM02, KMNR95]. For this reason, data-dependent notions of complexity have also been introduced, like the Rademacher complexity and the Gaussian complexity. These notions of complexity evaluate the expected noise-fitting-ability of a function class over all data sets drawn according to an unknown data distribution. By means of such data-dependent complexity measures, one obtains generalization bounds that are considerably better than those involving the VC dimension [BM02, PL20, SSBD14]. In any case, computing or estimating the VC dimension, or the Rademacher complexity, is generally a challenging task, feasible only under strong model constraints. Some tight approximations and bounds have been obtained in some specific cases, however they are not general enough to be applicable to complex models like modern deep neural networks [BHLM19, BM02, VLLC94]. With the aim to provide more easily computable notions of complexity, other definitions have been considered based on the minimum description length (MDL) and on the notion of stochastic complexity [Grü07, Ris97, BMR19].

Other complexity measures of more geometric flavour have been defined in terms of the Fisher information associated with the statistical model [Ris96]. More recently, [BFGM20, ASZ$^+$21] propose a notion of *effective dimension*, that is, a box-covering dimension related to the number of "Fisher boxes" of a given size that are needed to cover the parameter space. The size of such boxes represents a "scale" at which the model is analysed. Under suitable regularity assumptions on the statistical model and on the loss functional, the generalization error (i.e., the gap between the population error and the empirical error) can be controlled by an expression involving the effective dimension computed with respect to an explicit scale parameter, that depends on the number of samples defining the empirical error. One of the main challenges in the computation of the effective dimension is to determine the eigenvalues of the Fisher information matrix. Note that, for high-dimensional models, even the storage of the Fisher information becomes impractical, despite sophisticated approximation methods such as K-FAC [MG15].

This work proposes and studies variants of the original effective dimension. Our aim is to address some issues that affect the previous definitions. On the one hand, the generalization bounds proved in [ASZ$^+$21] (see also [ASFW21]) require strong regularity assumptions on the statistical model. Specifically, the logarithm of the Fisher information matrix is required to be Lipschitz, and in particular its eigenvalues cannot become too small, hence this assumption excludes the case of over-parametrized models [KAA19]. On the other hand, the original definition requires a global computation for the statistical model as a whole, hence it does not satisfy properties (P2) and (P3).

Our starting point is the definition of the following, *two-scale effective dimension* (2sED)

$$d_\zeta(\varepsilon) = \zeta d + (1-\zeta)\frac{\log \fint_\Theta \det\left(I_d + \varepsilon^{\zeta-1}\sqrt{\hat{F}(\vartheta)}\right) d\vartheta}{|\log \varepsilon^{\zeta-1}|}, \tag{1}$$

where $\hat{F}(\vartheta)$ denotes the normalized Fisher information matrix (see Section 2) and $\fint_\Theta f(\vartheta)d\vartheta = \frac{1}{V_\Theta}\int_\Theta f(\vartheta)d\vartheta$, with $V_\Theta = \int_\Theta d\vartheta$. Two parameters show up in the above definition: a micro-scale $\varepsilon > 0$ and an exponent $\zeta \in [0,1)$ defining a meso-scale $\delta = \varepsilon^\zeta$. The micro-scale is related to the size of *Fisher boxes* (defined in (19)) that are used to cover a component of the parameter space, while the meso-scale ε^ζ represents the size of the components of a partition of the parameter space, that needs to be fixed in order to localise and adapt the micro-scale covering. Note that when $\zeta = 0$ we essentially obtain the effective dimension of [ASZ$^+$21], up to a slight technical difference due to the presence of the square root of the renormalized Fisher matrix $\hat{F}$. More generally, the 2sED is a convex combination of the dimension of the parameter space and the effective dimension. Our main theoretical result is Theorem 4.1, which establishes a generalization bound explicitly dependent upon the cardinality of the dataset and the 2sED.

We stress that the proof of this result requires considerably weaker regularity assumptions than those of [ASZ$^+$21, Theorem 1], which is relevant for various models that are considered in practice.

The 2sED defined in (1) depends on the eigenvalues of the Fisher information matrix and hence is not straightforward to evaluate efficiently for large models with many parameters. Furthermore we need to average the determinant of these eigenvalues over the full parameter space which requires computing a high-dimensional integral. To overcome this problem, we introduce a modular version of $d_\zeta(\varepsilon)$, that is specifically tailored for Markovian models, which can be though of as stochastic generalizations of feedforward DNNs. This new quantity, called *lower 2sED* and denoted by $\underline{d}_\zeta(\varepsilon)$, provides a lower bound for $d_\zeta(\varepsilon)$. It is obtained by exploiting the concavity property of the logarithm, and has the advantage of being computed sequentially layer-by-layer, thanks to the block structure of the Fisher information matrix of a Markovian model. As a result, the computational cost required to evaluate $\underline{d}_\zeta(\varepsilon)$ is drastically reduced compared to that of $d_\zeta(\varepsilon)$. Moreover, the need to store the full Fisher information matrix is eliminated, since only the i-th layer block of the matrix needs an iterative evaluation. Consequently, the lower 2sED $\underline{d}_\zeta(\varepsilon)$ satisfies (P3) and therefore can be computed for models that are more complex (deeper) than those considered in [ASZ$^+$21].

We finally present numerical simulations based on Monte Carlo approximations of $\underline{d}_0$ for various models and datasets. The experiments remarkably confirm properties (P1), (P2), and (P3). Concerning (P2) and (P3), these are satisfied by the approximation of $\underline{d}_0$, thanks to the analytical properties of $\underline{d}_0$. Moreover, we have observed a systematic correlation between the approximations of d_0 and $\underline{d}_0$, thus they provide the same comparative information about models. Finally, we have an experimental evidence that the post-training performances of a given parametric model are strongly correlated to higher values of $\underline{d}_0$, which is an experimental confirmation of (P1).

To summarize, we present a new capacity measure called 2sED that

(i) provably bounds the generalization error under mild assumptions on the model (see Theorem 4.1), providing non-vacuous estimates on the generalization error in the under-parametrized regime;
(ii) correlates well with the training error for popular models and data sets (see Section 6);
(iii) can be approximated efficiently for Markovian models, which allows us to compute it for deep neural networks that have a large number of parameters (see Sections 5 and 6);
(iv) satisfies properties (P1)–(P3).

2. Preliminaries

Take $\mathcal{X} \subset \mathbb{R}^{d_{in}}$ and $\mathcal{Y} \subset \mathbb{R}^{d_{out}}$ nonempty Borel sets, and denote by $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ a pair of random vectors with (unknown) joint probability distribution $p = p(x, y)$. Let $(X_1, Y_1), \dots, (X_N, Y_N)$ be i.i.d. copies of (X, Y). A dataset $\mathbb{D} := \{(x_i, y_i) : i = 1, \dots, N\}$ is understood as a realization of the N random pairs considered before. A *statistical model* on the sample space $\mathcal{X} \times \mathcal{Y}$ is a collection

$$\mathcal{M}_\Theta(\mathcal{X}, \mathcal{Y}) := \{p_\vartheta : \vartheta \in \Theta\},$$

where $p_\vartheta = p_\vartheta(x, y)$ is a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ for each $\vartheta \in \Theta$, and $\Theta \subseteq \mathbb{R}^d$ is a bounded domain called *parameter space*. In order to stress the functional relation between the input x and the output y, it is customary to assume p_ϑ of the form:

$$p_\vartheta(x, y) = p_\vartheta(y|x)\, p(x),$$

where $p_\vartheta(y|x)$ is a parametric conditional probability, and $p(x)$ denotes the marginal of the unknown distribution $p(x, y)$ on $\mathcal{X}$ (with a slight abuse of notation). We will also assume that the parameter space Θ is equipped with a probability measure, that we do not formally specify (just to fix the ideas, one could take the uniform probability, i.e., the normalized Lebesgue measure restricted to Θ). Consequently, we will denote as $\mathbb{E}_\vartheta$ the expectation with respect to this measure.

Under the hypothesis that the log-likelihood of the model is differentiable with respect to ϑ for p-a.e. x (in principle we should also require summability with respect to p, and could consider a weak L^2-derivative with respect to ϑ) we define the *Fisher information matrix* as:

$$F(\vartheta) := \mathbb{E}_{(x,y)\sim p_\vartheta}[(\nabla_\vartheta \log\, p_\vartheta(x,y)) \otimes (\nabla_\vartheta \log\, p_\vartheta(x,y))]\,, \tag{2}$$

where by $a \otimes b$ we mean $a \cdot b^T$ (with the convention that a and b are column vectors). In other words, the Fisher information matrix is the expectation of the orthogonal projector onto the direction of the gradient of the log-likelihood, scaled by the squared norm of that gradient. It is a symmetric and positive semidefinite $d \times d$ matrix. Its empirical version is:

$$F_N(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} (\nabla_\vartheta \log\, p_\vartheta(X_i,Y_i))) \otimes (\nabla_\vartheta \log\, p_\vartheta(X_i,Y_i))$$

for some $(X_1,Y_1),\dots,(X_N,Y_N) \overset{i.i.d.}{\sim} p_\vartheta$.
For each $\vartheta \in \Theta$, we define the *pointed Fisher norm* of a (tangent) vector $v \in \mathbb{R}^d$ as

$$\|v\|_\vartheta := \sqrt{\langle F(\vartheta)v, v\rangle}\,.$$

If $F(\vartheta)$ is smooth and positive-definite, then $\langle F(\vartheta)u, v\rangle$ defines a Riemannian metric on the parameter space Θ, that from now on will be called *Fisher metric* (we shall adopt the same terminology also when the metric is degenerate). In general, the Fisher metric can be considered as the pull-back of a (possibly degenerate) Riemannian metric on $\mathcal{M}_\Theta(\mathcal{X},\mathcal{Y})$ [LPRS19].

We conveniently introduce some terminology and a few definitions concerning $d\times d$ symmetric matrices and matrix fields. We denote by $S_+^d(\mathbb{R})$ the set of real $d \times d$ symmetric and positive semidefinite matrices. Given $\beta > 0$ and $A \in S_+^d(\mathbb{R})$, we define A_β as the matrix obtained from A by replacing all the eigenvalues of A smaller than β with β, that is, if we write

$$A = \sum_{i=1}^{d} \lambda_i u_i \otimes u_i\,,$$

where $\{u_i\}_{i=1}^d$ is the spectral basis of A, then

$$A_\beta := \sum_{i=1}^{d} \max\big(\lambda_i, \beta\big) u_i \otimes u_i\,. \tag{3}$$

Let $A \in S_+^d(\mathbb{R})$, then for any $v \in \mathbb{R}^d$ we set

$$[v]_A := \max_{i=1,\dots,d} \sqrt{\lambda_i}|\langle v, u_i\rangle|\,, \tag{4}$$

where λ_i and u_i are, respectively, the i-th eigenvalue and the corresponding eigenvector of A. We say that a matrix field $A : \Theta \to S_+^d(\mathbb{R})$ is L-Lipschitz if it can be written as

$$A(\vartheta) = \sum_{i=1}^{d} \lambda_i(\vartheta)\, u_i(\vartheta) \otimes u_i(\vartheta)\,,$$

where $\{u_i\}_{i=1}^d$ is an orthonormal frame, and u_i, λ_i are L-Lipschitz for all i.

Some further terminology must be recalled before discussing the generalization bounds. Given a *loss function* $\mathfrak{L}$, i.e., a continuous function $\mathfrak{L} : [0,+\infty) \times [0,+\infty) \to [0,+\infty)$ such that $\mathfrak{L}(a,b) = 0$ if and only if $a = b$, we define the *population risk*

$$R(\vartheta) := \mathbb{E}_{(x,y)\sim p}[\mathfrak{L}(p_\vartheta(y|x), p(y|x))],$$

and the *empirical risk*

$$R_n(\vartheta) := \frac{1}{n}\sum_{i=1}^{n} \mathfrak{L}(p_\vartheta(Y_i|X_i), p(Y_i|X_i))\,,$$

where $(X_i, Y_i) \overset{i.i.d.}{\sim} p$, $i = 1, \dots, n$. Then, the *generalization error* is defined as

$$\|R - R_n\|_\infty = \sup_{\vartheta \in \Theta} |R(\vartheta) - R_n(\vartheta)| \,. \tag{5}$$

3. The two-scale effective dimension

Here we consider a notion of complexity for a statistical model $\mathcal{M}_\Theta$, that depends on the properties of the Fisher metric on the parameter space. Given $0 < \varepsilon < 1$ and $0 \leq \zeta < 1$, we define the *two-scale effective dimension* (or simply 2sED) as

$$d_\zeta(\varepsilon) = \zeta d + (1-\zeta) \frac{\log \mathbb{E}_\vartheta \left[\det \left(I_d + \varepsilon^{\zeta - 1} \hat{F}(\vartheta)^{\frac{1}{2}} \right) \right]}{|\log(\varepsilon^{\zeta-1})|} \,, \tag{6}$$

where

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

is the normalized Fisher information matrix, so that whenever the statistical model is not trivial (i.e. not constant with respect to ϑ) the expectation of the trace of $\hat{F}$ satisfies

$$\mathbb{E}_\vartheta[\operatorname{Tr} \hat{F}(\vartheta)] = d \,.$$

Note that $d_\zeta(\varepsilon)$ is the convex combination of the dimension d of the parameter space with a slight variant of the effective dimension studied in [ASZ+21], which is obtained in the special case $\zeta = 0$.

Remark 3.1. *The effective dimension $d_\zeta(\varepsilon)$ can be shown to converge to $\zeta d + (1-\zeta)\hat{r}$ as $\varepsilon \to 0$, where $\hat{r} := \max_{\vartheta \in \Theta} \operatorname{rank}(\hat{F}(\vartheta))$, see Proposition A.1. The proof follows the strategy of [ASZ+21, Remark 1], and is presented in Appendix A for completeness.*

Example 3.2 (1D Gaussians with fixed variance σ^2)**.** Let us consider the statistical model

$$\mathcal{M}_\Theta := \left\{ p_\vartheta(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - f_\vartheta(x))^2}{2\sigma^2}} \,:\, \vartheta = [0,1] \right\} \,,$$

where $f_\vartheta : \mathbb{R} \to \mathbb{R}$ is a given parametric function. Let $p(x,y) = p(y|x)p(x)$ be an unknown joint probability distribution. Since

$$\partial_\vartheta \log\, p_\vartheta(y|x) = \frac{(y - f_\vartheta(x))}{\sigma^2} \partial_\vartheta f_\vartheta(x) \,,$$

one obtains

$$\begin{aligned} F(\vartheta) &= \mathbb{E}_{(x,y)\sim p_\vartheta} \left[\left(\partial_\vartheta \log\, p_\vartheta(y|x) \right)^2 \right] \\ &= \mathbb{E}_{x\sim p} \left[(\partial_\vartheta f_\vartheta(x))^2 \, \mathbb{E}_{y \sim p_\vartheta(\cdot|x)} \left[\frac{(y - f_\vartheta(x))^2}{\sigma^4} \right] \right] \\ &= \frac{1}{\sigma^2} \mathbb{E}_{x\sim p} \left[(\partial_\vartheta f_\vartheta(x))^2 \right] \end{aligned}$$

for all $\vartheta \in \Theta$. The normalized Fisher information matrix is then:

$$\hat{F}(\vartheta) = \frac{1}{\mathbb{E}_\vartheta \left[\mathbb{E}_{x\sim p} \left[(\partial_\vartheta f_\vartheta(x))^2 \right] \right]} \mathbb{E}_{x\sim p} \left[(\partial_\vartheta f_\vartheta(x))^2 \right]$$

Hence we get

$$d_\zeta(\varepsilon) = \zeta + (1-\zeta)\frac{\log \mathbb{E}_\vartheta\left(1+\varepsilon^{-1}\frac{\sqrt{\mathbb{E}_{x\sim p}\left[(\partial_\vartheta f_\vartheta(x))^2\right]}}{\sqrt{\mathbb{E}_\vartheta\left[\mathbb{E}_{x\sim p}\left[(\partial_\vartheta f_\vartheta(x))^2\right]\right]}}\right)}{|\log\varepsilon|}$$
$$= \zeta + (1-\zeta)\frac{\log\left(1+\varepsilon^{-1}\frac{\mathbb{E}_\vartheta\left[\sqrt{\mathbb{E}_{x\sim p}\left[(\partial_\vartheta f_\vartheta(x))^2\right]}\right]}{\sqrt{\mathbb{E}_\vartheta\left[\mathbb{E}_{x\sim p}\left[(\partial_\vartheta f_\vartheta(x))^2\right]\right]}}\right)}{|\log\varepsilon|}\,.$$

Note that by Jensen's inequality we obtain

$$d_\zeta(\varepsilon) \le \zeta + (1-\zeta)\frac{\log\left(1+\varepsilon^{-1}\right)}{|\log\varepsilon|} \to 1 \qquad \text{as } \varepsilon\to 0\,. \tag{7}$$

Let us consider now the simpler case of a linear model $f_\vartheta(x) = \vartheta x$. In this case we have

$$\hat{F}(\vartheta) = \frac{1}{\mathbb{E}_\vartheta\left[\mathbb{E}_{x\sim p(x)}\left[x^2\right]\right]}\mathbb{E}_{x\sim p(x)}\left[x^2\right] = 1$$

for all $\vartheta\in\Theta$. Therefore, the 2sED of the one-dimensional linear model is

$$d_\zeta(\varepsilon) = \zeta + (1-\zeta)\frac{\log(1+\varepsilon^{-1})}{|\log\,\varepsilon|} = 1 + (1-\zeta)\frac{\log(1+\varepsilon)}{|\log\varepsilon|}\,,$$

hence it saturates the upper bound (7) and converges to 1 as $\varepsilon\to 0$, in accordance with Remark 3.1. Therefore, when $d=1$, linear models reach the highest possible effective dimension.

4. Generalization bounds

It is known that the Fisher information of a statistical model degenerates asymptotically with the number of parameters [KAA19]. This suggests that, in the case of a large (over-parametrized) model, like a deep neural network with high-dimensional layers, the corresponding Fisher information matrix $F(\vartheta)$ should have a lot of small (or possibly zero) eigenvalues. For this reason, in Theorem 4.1 below we will not require the Lipschitz regularity of $\log(F(\vartheta))$, as done in [ASZ+21, Theorem 1], because this assumption would imply uniform positive lower bounds on the eigenvalue of $F(\vartheta)$. Without loss of generality, we directly assume $F=\hat{F}$ and $\Theta=[0,1]^d$, as this can be enforced by a suitable scaling of the model.

We list below a set of hypotheses, that will be required in the generalization bounds:

(i) the map $\vartheta\mapsto p_\vartheta(y|x)$ is of class $C^{1,1}$ uniformly in (x,y);
(ii) there exist two constants $0<\alpha_1\le\alpha_2$ such that

$$\alpha_1 \le p(x,y),\ p_\vartheta(x,y)\le\alpha_2$$

for all $x\in\mathcal{X}, y\in\mathcal{Y}, \vartheta\in\Theta$;
(iii) the Fisher matrix $F(\vartheta)$ is L-Lipschitz and its eigenvalues are smaller than μ (for some fixed $\mu>0$);
(iv) the loss function $\mathcal{L}$ is bounded by $2b$ and is Λ-Lipschitz, for some $b,\Lambda>0$;
(v) the meso-scale parameter ζ satisfies $\zeta\in[\frac{2}{3},1)$.

Some comments about the previous properties are in order. First, property (i) is a mild regularity assumption on the model. Property (ii) prevents degeneration of both probability densities $p(x,y)$ and $p_\vartheta(x,y)$. The L-Lipschitz property (iii) corresponds to the existence of a L-Lipschitz spectral frame, such that the corresponding eigenvalues are also L-Lipschitz as functions of ϑ (this assumption is crucial to compare the pointed Fisher norm computed in different $\vartheta\in\Theta$). Lipschitz regularity and boundedness of the loss function $\mathcal{L}$ (property (iv)) are standard assumptions (see, e.g., [LG23]). Finally, property (v) is structurally needed in the proof of the generalization bound (Theorem 4.1).

Theorem 4.1. *Let us assume (i)–(v). Then, there exist explicit constants* $C, H, K, n_0 > 0$[2] *such that for any* $\gamma \in (0,1]$, $n \geq n_0$, *and* $\varepsilon_n = \left(\frac{\log n}{\gamma n}\right)^{3/8}$, *we obtain*

$$\mathbb{P}\left(\sup_{\vartheta\in\Theta} |R(\vartheta) - R_n(\vartheta)| \geq C\varepsilon_n\right) \leq H\varepsilon_n^{-d_\zeta(\varepsilon_n)} n^{-\frac{K}{\gamma}} . \tag{8}$$

The proof of Theorem 4.1 is given in Appendix B.

Remark 4.2. *The above result implies the existence of* $\gamma_0 > 0$ *such that, for* $0 < \gamma < \gamma_0$, *the right hand side of* (8) *vanishes as* $n \to \infty$. *The upper bound* γ_0 *is explicit and depends only on the dimension* d *and on the properties of the model, see* (31). *By choosing* γ *as above, the right-hand side of* (8) *gives an explicit upper bound of the generalization error, that is non-vacuous also for finite* n *(even though this can be granted only in the under-parametrized regime, i.e. for* n *large enough).*

5. The Effective Dimension of a Markovian Model

Markovian models are a family of probabilistic models characterized by a sequential, feed-forward-type structure, see the Markovian property stated below.

Let us consider an integer $L \geq 2$, a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a random vector $X_j : \Omega \to \mathcal{X}_j$ for $j = 0, \dots, L$. Given a parameter space $\Theta = \Theta_1 \times \cdots \times \Theta_L$, a parametric statistical model $\mathcal{M}_\Theta(\mathcal{X}_0, \dots, \mathcal{X}_L)$ satisfies the Markovian property if and only if for each $p_\vartheta(x_0, \dots, x_L) \in \mathcal{M}_\Theta(\mathcal{X}_0, \dots, \mathcal{X}_L)$ and for each $\vartheta = (\vartheta_1, \dots, \vartheta_L) \in \Theta = \Theta_1 \times \cdots \times \Theta_L$ we have:

$$p_\vartheta(x_0, \dots, x_L) = p(x_0) p_{\vartheta_1}(x_1|x_0) \cdots p_{\vartheta_L}(x_L|x_{L-1}) \tag{9}$$

where $\vartheta_1, \dots, \vartheta_L$ are the parameters associated to the model's distribution of $X_1, \dots, X_L$ respectively. Many well-known and commonly used neural network architectures, such as feed-forward neural networks, can be interpreted as Markovian models with concentrated, Dirac-type probability distributions. A specific evaluation of the effective dimension of these models seems therefore particularly interesting. Exploiting the Markovian property, for $j = 1, \dots, L$ we define

$$F(\vartheta_j|x_{j-1}) := \int_{\mathcal{X}_j} \nabla_{\vartheta_j} \log\, p_{\vartheta_j}(x_j|x_{j-1}) \otimes \nabla_{\vartheta_j} \log\, p_{\vartheta_j}(x_j|x_{j-1})\; p_{\vartheta_j}(dx_j|x_{j-1})$$

and

$$F_j := F_j(\vartheta_1, \dots, \vartheta_j) := \mathbb{E}_{x_0}\mathbb{E}_{x_1|x_0} \cdots \mathbb{E}_{x_{j-1}|x_{j-2}}[F(\vartheta_j|x_{j-1})]\,,$$

where by $\mathbb{E}_{x_0}$ and $\mathbb{E}_{x_j|x_{j-1}}$ we denote the (conditional) expectations with respect to $p(x_0)$ and $p_{\vartheta_j}(x_j|x_{j-1})$, respectively. Clearly F_j is a symmetric and positive semidefinite $d_j \times d_j$ matrix (where d_j is the dimension of Θ_j) and represents the j-th block of the Fisher information matrix

$$F(\vartheta) = \begin{pmatrix} F_1(\vartheta_1) & 0 & \cdots & 0 \\ 0 & F_2(\vartheta_1, \vartheta_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & F_L(\vartheta_1, \dots, \vartheta_L) \end{pmatrix} . \tag{10}$$

We recall that the two-scale effective dimension (2sED) is

$$d_\zeta(\varepsilon) = \zeta d + \frac{1-\zeta}{|\log\varepsilon|} \log \fint_{\Theta_1} \cdots \fint_{\Theta_L} \prod_{j=1}^{L} \det\left(I_j + \varepsilon^{-1} F_j(\vartheta_1, \dots, \vartheta_j)^{\frac{1}{2}}\right) d\vartheta_1 \cdots d\vartheta_m\,, \tag{11}$$

where I_j denotes the $d_j \times d_j$ identity matrix. Since F_j depends on all the parameters of the previous blocks, it is not possible to directly factorize the multiple integral in (11). Nevertheless, one obtains a more easily computable lower bound of $d_\zeta(\varepsilon)$, called *lower 2sED*, by a single

[2] the constants can be computed/estimated in terms of the assumptions.

application of Jensen's inequality as hereafter described. Let $d_\zeta^m(\varepsilon)$ be the 2sED associated with the composition of the first m layers, $m \geq 2$. Then:

$$
\begin{aligned}
&d_\zeta^m(\varepsilon) - d_\zeta^{m-1}(\varepsilon) \\
&\quad = \frac{1-\zeta}{|\log\varepsilon|} \log\left(\frac{1}{\Phi_m(\hat{\Theta}_m)} \int_{\hat{\Theta}_m} \fint_{\Theta_m} \det(I_m + \varepsilon^{-1} F_m(\vartheta_1,\dots,\vartheta_m)^{\frac{1}{2}}) d\vartheta_m \; \Phi_m(d\vartheta_1,\dots,d\vartheta_{m-1}) \right) \\
&\quad \geq \frac{1-\zeta}{|\log\varepsilon|} \fint_{\hat{\Theta}_m} \fint_{\Theta_m} \log\det(I_m + \varepsilon^{-1} F_m(\vartheta_1,\dots,\vartheta_m)^{\frac{1}{2}}) d\vartheta_m \Phi_m(d\vartheta_1,\dots,d\vartheta_{m-1}) \,,
\end{aligned}
$$

where we have set

$$
\begin{aligned}
&\hat{\Theta}_m := \Theta_1 \times \cdots \times \Theta_{m-1} \\
&\Phi_m(d\vartheta_1,\dots,d\vartheta_{m-1}) := \frac{1}{\prod_{j=1}^{m-1} |\Theta_j|} \prod_{j=1}^{m-1} \det\left(I_j + \varepsilon^{-1} F_j(\vartheta_1,\dots,\vartheta_j)^{\frac{1}{2}} \right) d\vartheta_1 \cdots d\vartheta_{m-1} \,.
\end{aligned}
$$

Now, a lower bound of $d_\zeta^m(\varepsilon)$ can be iteratively defined for $m = 1,\dots,L$ as follows:

(12)

$$
\begin{aligned}
&\underline{d}_\zeta^1(\varepsilon) = \zeta d + \frac{1-\zeta}{|\log\varepsilon|} \log \fint_{\Theta_1} \det(I_1 + \varepsilon^{-1} F_1(\vartheta_1)^{\frac{1}{2}}) \, d\vartheta_1 \\
&\underline{d}_\zeta^2(\varepsilon) = \underline{d}_\zeta^1(\varepsilon) + \frac{1-\zeta}{|\log\varepsilon|} \fint_{\hat{\Theta}_2} \fint_{\Theta_2} \log\det(I_2 + \varepsilon^{-1} F_2(\vartheta_1,\vartheta_2)^{\frac{1}{2}}) \, d\vartheta_2 \, \Phi_2(d\vartheta_1) \\
&\qquad \vdots \\
&\underline{d}_\zeta^m(\varepsilon) = \underline{d}_\zeta^{m-1}(\varepsilon) + \frac{1-\zeta}{|\log\varepsilon|} \fint_{\hat{\Theta}_m} \fint_{\Theta_m} \log\det(I_m + \varepsilon^{-1} F_m(\vartheta_1,\dots,\vartheta_m)^{\frac{1}{2}}) \, d\vartheta_m \, \Phi_m(d\vartheta_1,\dots,d\vartheta_{m-1}) \,.
\end{aligned}
$$

From now on we set $\underline{d}_\zeta = \underline{d}_\zeta^L$ and call it the *lower effective dimension* of the Markovian model $\mathcal{M}_\Theta$.

6. Experiments

In this section, we present experimental evidence that the post-training performance of given parametric models is related both with 2sED (6) and the lower 2sED (12). We compute $\underline{d}_\zeta$ and d_ζ of different feed-forward neural networks (FNN) such as convolutional neural networks (CNN) and multi-layer perceptron (MLP). The feed-forward neural network choice is justified by their architecture characterised by a Markovian dependency structure. Indeed, the flow of information in FNN is unidirectional from input to output, making them representable with a finite acyclic graph. We evaluate $\underline{d}_\zeta$ and d_ζ on real-world datasets, including Covertype dataset [Bla98], MNIST dataset [Den12], and CIFAR10 [KNH].

To simplify notation and enhance readability, we denote with "MLP N_0-N_1-...-N_n" a MLP with n linear layers , each with a width of N_i for $i = 0,\dots,N$, followed by ReLU activation functions on all layers except the final layer n. If we denote with $W^i \in R^{N_i \times N_{i-1}}$ the parameters of the i-th layer, we can describe "MLP N_0-N_1-...-N_n" through n blocks of operations defined as $O_i(\cdot) = ReLU(W^i \cdot)$ for $i = 1,\dots,n$. Similarly, "CNN N_0-N_1-...-$N_{n_1}|L_1$-...-L_{n_2}" refers to a convolutional neural network with n_1 convolutional blocks each one of kernel size N_i for $i = 1,\dots,N_1$ followed by a flattening layer and n_2 MLP blocks of width L_i for $i = 1,\dots,n_2$. Within each convolutional block, the operations of convolution, batch normalization, ReLU activation, and max pooling are performed sequentially. Moreover, the flattening operation is executed by applying a common convolutional kernel to all the channels of the last convolutional layer. Hence, given the input $A \in \mathbb{R}^{N_c \times k \times k}$ the flattening operation $Flat_K : \mathbb{R}^{N_c \times k \times k} \to \mathbb{R}^{N_c}$ is

defined as follows:

$$[Flat_K(A)]_l := A_{l::} \star K = \sum_{i=1}^{k}\sum_{j=1}^{k} A_{l,i,j}K_{i,j} \qquad\qquad l = 1, \dots, N_c\,,$$

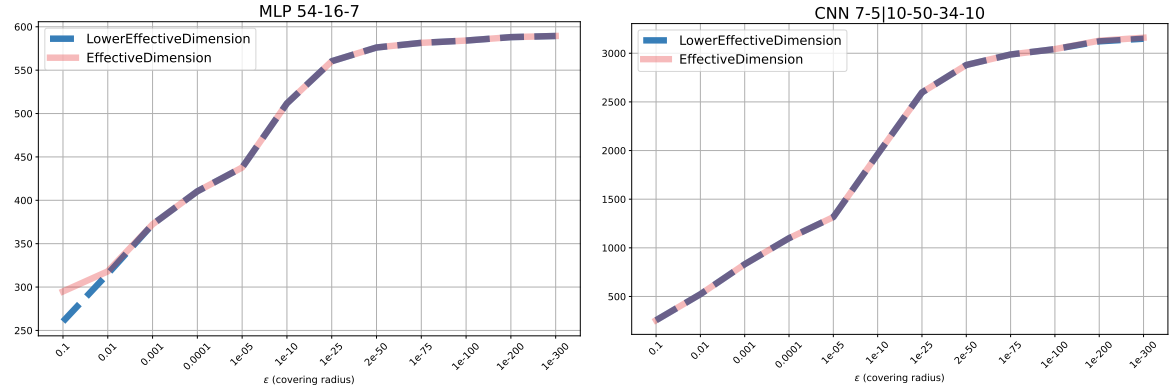
where $A_{l::}$ denotes the $\mathbb{R}^{k\times k}$ obtained by fixing the first dimension at index l and K is a $k \times k$ convolutional kernel which is a parametric matrix in applications. This approach effectively reduce the number of parameters allowing us to compute the effective dimension in reasonable time.

In applications, the core architectures of many deep learning models is deterministic, and the stochasticity is usually introduced in the training pipeline rather than in the model itself. This makes deep learning models, like MLPs and CNNs, incompatible with our setting. Therefore, we approximate deterministic feed-forward neural networks with stochastic variants, where the output of each block is Gaussian with mean the current block deterministic output and a small fixed variance σ^2. In other words, if N is the number of blocks, the output of the i-th block O_i^σ is given by

$$O_i^\sigma = O_i + \nu \sim \mathcal{N}(O_i, \sigma^2 I)\,,$$

where O_i is the deterministic output of the i-th block, $\nu \sim \mathcal{N}(0, \sigma^2)$.

For all the subsequent experiments, we will specifically focus on the computation of 2sED and lower-2sED for $\zeta = 0$ and considering the empirical Fisher information matrix $\hat{F}_N$. To empirically validate the lower 2sED, we compute $\underline{d}_0$ and d_0 for different stochastic perturbations of feed-forward neural networks, also varying the covering radius ε. We keep constant both the 100 samples used to estimate $\hat{F}_N$ and the 100 parametrizations employed for estimating the integrals appearing in (6) and (12). The results are visualized in Figure 1.



(a) **(b)**

Figure 1. In this figures the difference between 2sED and lower 2sED cannot be appreciated, which means that the second is a tight lower bound of the first. In plot **(a)**, the lower 2sED and 2sED of MLP 54-16-7 using 100 Covertype samples and 100 parametrizations. Plot **(b)** shows the lower 2sED and the 2sED of CNN 7-5|10-50-34-10 using 100 MNIST samples and 100 parametrizations.

Notably, the lower bound is sharp in both the MLP and the CNN case suggesting the conclusions regarding model complexity obtained using $\underline{d}_\zeta(\varepsilon)$ are equivalent to those obtained when considering $d_\zeta(\varepsilon)$ for all covering radius ε. It is also worth to notice that lower 2sED exhibits a sequential form, reducing the computational demands when investigating how the model's complexity changes by modifying only its final components.

We study now the impact of variance σ^2 on 2sED and lower 2sED. We vary the values of σ^2 while computing the $\underline{d}_\zeta$ and d_ζ for different models on Covertype and MNIST dataset. As before, we fix the data and the parametrizations used to estimate the integrals to avoid discrepancies stemming from these estimations. The results in Figure 3 show that the impact of σ^2 on $\underline{d}_\zeta$ and

d_ζ is negligible. This observation ensures the meaningfulness of 2sED, and lower 2sED, when considering real deep learning models as the limit of our stochastic version with $\sigma^2 \to 0$.

Monte Carlo integration is crucial in the estimation of both the Fisher information matrix and the integral within Θ appearing in (6). To ensure the reliability of our results, we conduct a robustness analysis of the lower 2sED with respect to variations in the number of samples and parameterizations employed for integrals estimation. In particular, the 2sED is computed for three different models on Covertype, MNIST and Cifar10 dataset. Figure 4 confirms the stability of the lower 2sED plots with respect to the number of points used in the Monte Carlo approximation. Indeed, even if the value of the lower 2sED vary together with the number of estimation points, the conclusions remain consistent.

Finally, we test the relationship between the lower 2sED and the loss minimization. We expect that models with higher values of the lower 2sED can achieve higher accuracy after training. Furthermore, it is crucial to gain a deeper understanding of the role played by the covering radius, denoted as ε, in the 2sED definition. We compute the lower 2sED for three different models with similar dimension on CIFAR10 and Covertype dataset. The dimension of these models is reported in Table 1.

Model	**Number of Parameters**
MLP 54-16-7	976
MLP 54-13-11-9-7	1007
MLP 54-10-2-10-25-7	1005
CNN 7-5|10-50-34-10	4493
CNN 3-5-3-6|10-50-34-10	10034
CNN 3-6-5-3|10-50-34-10	10041

Table 1. Number of model's parameters

MLP 54-10-2-10-25-7 is characterized by a bottleneck structure in the middle of its architecture. A loss of information due to this bottleneck is therefore expected as data are mapped into a significantly lower dimensional space. Consequently, the expressiveness of this model is expected to be lower compared to the other two models, even though it is bigger than MLP 54-16-7 in terms of number of parameters. In Figure 2, this expected behaviour is effectively captured by the lower 2sE, as indicated by the lower position of the red curve in comparison to the other two curves. Furthermore the position of the curves change varying the covering radius ε. Indeed, the blue curve remains above the other two curves up to a certain scale, suggesting that the MLP 54-16-7 model exhibits greater expressiveness within this range of ϵ. Conversely, for smaller values of ϵ, MLP 54-13-11-9-7 appears to be more expressive. This behaviour is empirically validated by the experiments. In plot (c), we observe the training loss curve for the three models when trained with only 10000 data. In this scenario, the blue model achieves a lower training loss minimum compared to the other two models, and the shapes of these curves mirror the lower 2dED plots for a small ε.
Increasing the number of training data to 500000, MLP 54-13-11-9-7 is the one achieving the lower training loss. The empirical correlation between training losses and lower 2sED underscores the capacity of 2sED as a reliable measure for describing the training capabilities of neural networks. We conducted additional experiments, manipulating the number of training data points. The outcomes align consistently with the previously described results. This further confirms its effectiveness as a capacity metric. Other experiments in this direction are performed on the CIFAR10 dataset and the results are reported in Figure 2. We also conducted experiments on MNIST dataset with varying batch sizes, as illustrated in Figure 5 providing additional empirical evidence that supports our findings.

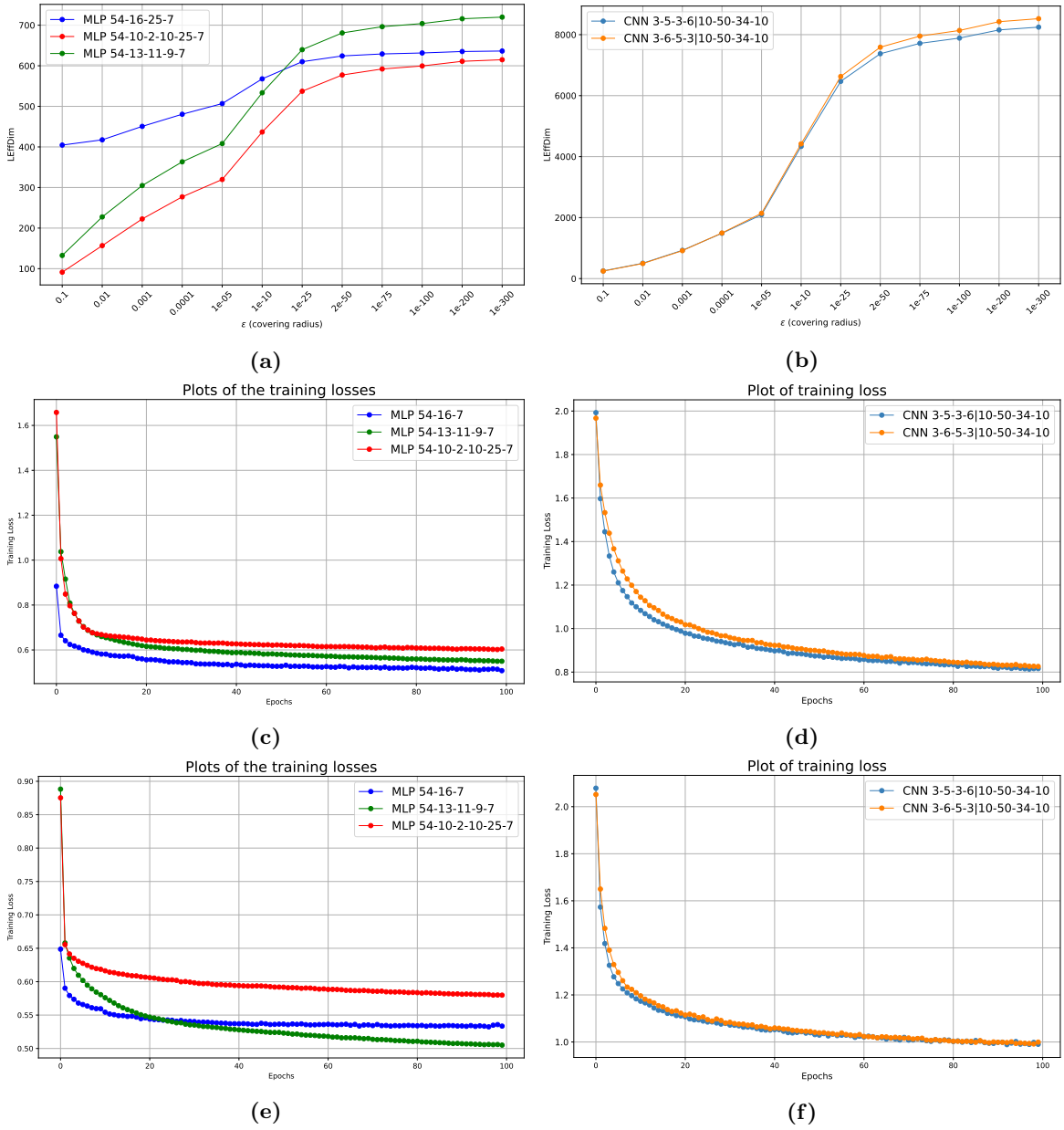


Figure 2. This pictures highlight the relation between the lower 2sED (Figure **(a)**) and the training loss curves of three MLPS (Figure **(c)** and Figure **(d)**). In particular, it also shows that we should consider covering radius ε based on the number of training data as suggested by Theorem 4.1. The same experiment is performed for two different CNNs (Figure **(b)**, Figure **(e)**, Figure **(f)**).
(a) Estimated lower 2sED of MLPs using 100 Covertype samples and 100 parametrizations; **(b)** Estimated lower 2sED of CNNs using 100 CIFAR10 samples and 100 parametrizations; **(c)** Training loss plots of MLPs on 10000 Covertype samples using Adam with learning rate $1e^{-3}$ and a batch size 64 **(d)** Training loss plots of CNNs on CIFAR10 with Adam optimizer with learning rate $1e^{-3}$ and a batch size 512; **e** Training loss plot of MLPs on 100000 Covertype samples using same optimization algorithm as (c);**(f)** Training loss plots of CNNs on augmented CIFAR10 (double the original size) with optimization strategy as in (e);

References

[ASFW21] Amira Abbas, David Sutter, Alessio Figalli, and Stefan Woerner. Effective dimension of machine learning models. *arXiv:2112.04807*, 2021.

[ASZ+21] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

[BFGM20] Oksana Berezniuk, Alessio Figalli, Raffaele Ghigliazza, and Kharen Musaelian. A scale-dependent notion of effective dimension. *arXiv:2001.10872*, 2020.

[BFSO84] L Breiman, J Friedman, C Stone, and R Olshen. Classification and regression trees (crc, boca raton, fl). 1984.

[BHLM19] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

[Bla98] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[BM02] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[BMR19] Nicola Bulso, Matteo Marsili, and Yasser Roudi. On the complexity of logistic regression models. *Neural computation*, 31(8):1592–1623, 2019.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[CSMV99] Vladimir Cherkassky, Xuhui Shao, Filip M Mulier, and Vladimir N Vapnik. Model complexity control for regression using vc generalization bounds. *IEEE transactions on Neural Networks*, 10(5):1075–1089, 1999.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Grü07] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[HCP+21] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63:2585–2619, 2021.

[HFT09] Trevor Hastie, Jerome H Friedman, and Robert Tibshirani. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[KAA19] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019.

[KMNR95] Michael Kearns, Yishay Mansour, Andrew Y Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 21–30, 1995.

[KNH] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[KST08] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LG23] Mohammad Lashkari and Amin Gheibi. Lipschitzness effect of a loss function on generalization performance of deep neural networks trained by adam and adamw optimizers. *arXiv preprint arXiv:2303.16464*, 2023.

[LPRS19] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pages 888–896. PMLR, 2019.

[MG15] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2408–2417. JMLR.org, 2015.

[MPCB14] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

[NTS14] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[PL20] Guillermo Valle Pérez and Ard A Louis. Generalization bounds for deep learning. *CoRR, vol. abs/2012.04115*, 2020.

[Ris96] Jorma J Rissanen. Fisher information and stochastic complexity. *IEEE transactions on information theory*, 42(1):40–47, 1996.

[Ris97] Jorma Rissanen. Stochastic complexity in learning. *journal of computer and system sciences*, 55(1):89–95, 1997.

[RPK+17] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pages 2847–2854. PMLR, 2017.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Vap99] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[VLLC94] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the vc-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix A. Asymptotic property of $d_\zeta(\varepsilon)$

In this section, we prove the following result.

Proposition A.1. *Let $\hat{r}$ denote the maximum rank of the Fisher matrix $\hat{F}(\vartheta)$ and let $\mu > 0$ be an upper bound for all the eigenvalues of $\hat{F}(\vartheta)$, for all $\vartheta \in \Theta$. Then, for all $\zeta \in [0,1)$ and $0 < \varepsilon < 1$ we have*

$$d_\zeta(\varepsilon) \le \zeta d + \hat{r}\left(1 - \zeta + \frac{\log(1+\mu^{1/2})}{|\log\varepsilon|}\right) \tag{13}$$

and, moreover,

$$\lim_{\varepsilon\to 0} d_\zeta(\varepsilon) = \zeta d + (1-\zeta)\hat{r}\,.$$

Proof. Let us fix ζ, ε as above. Denoting by r_ϑ the rank of $\hat{F}(\vartheta)$, we have:

$$\begin{aligned}
d_\zeta(\varepsilon) &= \zeta d + \frac{\log \fint_\Theta \det(Id_d + \varepsilon^{\zeta-1}\hat{F}^{1/2}(\vartheta))\,d\vartheta}{|\log\varepsilon|} \\
&= \zeta d + \frac{\log \fint_\Theta \prod_{i=1}^{r_\vartheta}(1+\varepsilon^{\zeta-1}\lambda_i^{1/2}(\vartheta))\,d\vartheta}{|\log\varepsilon|} \\
&\le \zeta d + \log\,\frac{\fint_\Theta \varepsilon^{(\zeta-1)r_\vartheta}\prod_{i=1}^{r_\vartheta}(1+\lambda_i^{1/2}(\vartheta))\,d\vartheta}{|\log\varepsilon|} \\
&\le \zeta d + \frac{\log \varepsilon^{(\zeta-1)\hat{r}}}{|\log\varepsilon|} + \frac{\log \fint_\Theta \prod_{i=1}^{r_\vartheta}(1+\lambda_i^{1/2}(\vartheta))\,d\vartheta}{|\log\varepsilon|}\,,
\end{aligned}$$

where $\lambda_i(\vartheta)$ are the nonzero eigenvalues of $\hat{F}(\vartheta)$. Notice that $\log \fint_\Theta \prod_{i=1}^{r_\vartheta}(1+\lambda_i^{1/2}(\vartheta))\,d\vartheta$ is finite. Indeed, $0 \le \lambda_i(\vartheta) \le \mu$ by assumption. Then it holds

$$1 \le \prod_{i=1}^{r_\vartheta}(1+\lambda_i^{1/2}(\vartheta)) \le (1+\mu^{1/2})^{\hat{r}}\,.$$

This implies that

$$1 \le \fint_\Theta \prod_{i=1}^{r_\vartheta}(1+\lambda_i^{1/2}(\vartheta)) \le (1+\mu^{1/2})^{\hat{r}}$$

and therefore (13). We thus conclude that

$$\lim_{\varepsilon\to 0} d_\zeta(\varepsilon) \le \zeta d + (1-\zeta)\hat{r}\,. \tag{14}$$

To see the other inequality, let us consider $\mathcal{A} := \{\vartheta \in \Theta : r_\vartheta = \hat{r}\}$. Notice that $\mathcal{A} \subset \Theta$ and hence $|\mathcal{A}| < \infty$. Also, by continuity of the Fisher matrix, the set $\mathcal{A}$ has positive measure. Then,

we have

$$\begin{aligned} d_\zeta(\varepsilon) &= \zeta d + \frac{\log \fint_\Theta \det(Id_d + \varepsilon^{\zeta-1}\hat{F}^{\frac{1}{2}}(\vartheta))\, d\vartheta}{|\log\varepsilon|} \\ &\geq \zeta d + \frac{\log \fint_{\mathcal{A}} \det(Id_d + \varepsilon^{\zeta-1}\hat{F}^{\frac{1}{2}}(\vartheta))\, d\vartheta}{|\log\varepsilon|} \\ &= \zeta d + \frac{\log \fint_{\mathcal{A}} \det(Id_{d_\vartheta} + \varepsilon^{\zeta-1}\hat{F}_0^{\frac{1}{2}}(\vartheta))\, d\vartheta}{|\log\varepsilon|}\,, \end{aligned}$$

where d_ϑ is the number of non-zero eigenvalues of $\hat{F}(\vartheta)$ and $\hat{F}_0(\vartheta)$ is the diagonal $d_\vartheta \times d_\vartheta$ containing only the d_ϑ non-zero eigenvalues of $\hat{F}(\vartheta)$ for all $\vartheta \in \Theta$. This yields

$$\begin{aligned} d_\zeta(\varepsilon) &\geq \zeta d + \frac{\log \fint_{\mathcal{A}} \det(Id_d) + \det(\varepsilon^{\zeta-1}\hat{F}_0^{\frac{1}{2}}(\vartheta))\, d\vartheta}{|\log\varepsilon|} \\ &= \zeta d + \frac{\log |\mathcal{A}|}{|\log\varepsilon|} + \frac{\log \fint_{\mathcal{A}} \prod_{i=1}^{\hat{r}} \varepsilon^{\zeta-1}\lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta}{|\log\varepsilon|} \\ &= \zeta d + \frac{\log |\mathcal{A}|}{|\log\varepsilon|} + \frac{\log \fint_{\mathcal{A}} \prod_{i=1}^{\hat{r}} \varepsilon^{\zeta-1}\lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta}{|\log\varepsilon|} \\ &= \zeta d + \hat{r}(\zeta-1)\frac{\log\varepsilon}{|\log\varepsilon|} + \frac{\log \fint_{\mathcal{A}} \lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta}{|\log\varepsilon|} \\ &= \zeta d + \hat{r}(1-\zeta) + \frac{\log \fint_{\mathcal{A}} \lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta}{|\log\varepsilon|}\,. \end{aligned}$$

Notice now that since $\lambda_i(\vartheta) \neq 0$ for all $\vartheta \in \Theta$, it holds that $log \fint_{\mathcal{A}} \lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta < \infty$ and so:

$$\lim_{\varepsilon\to 0} \frac{\log \fint_{\mathcal{A}} \lambda_i^{\frac{1}{2}}(\vartheta)\, d\vartheta}{|\log\varepsilon|} = 0\,.$$

Therefore

$$\lim_{\varepsilon\to 0} d_\zeta(\varepsilon) \geq \zeta d + \hat{r}(1-\zeta)\,. \tag{15}$$

Combining (14) and (15), we conclude

$$\lim_{\varepsilon\to 0} d_\zeta(\varepsilon) = \zeta d + \hat{r}(1-\zeta)\,.$$

□

Appendix B. Proof of the generalization bound

In this section we prove Theorem 4.1. We start with a preliminary lemma (see Section 2 for the notation).

Lemma B.1. *Let $A : \Theta \to S^d_+(\mathbb{R})$ be a Lipschitz tensor field. Then for all $\beta > 0$, $v \in \mathbb{R}^d$, and $\vartheta_1, \vartheta_2 \in \Theta$, one has*

$$\left| [v]^2_{A_\beta(\vartheta_1)} - [v]^2_{A_\beta(\vartheta_2)} \right| \leq \omega_\beta(|\vartheta_1 - \vartheta_2|)[v]^2_{A_\beta(\vartheta_1)}\,, \tag{16}$$

where

$$\omega_\beta(t) = \frac{(2\mu+1)L}{\beta}\, t\,. \tag{17}$$

Proof. Set

$$G(\vartheta) := [v]^2_{A_\beta(\vartheta)} \quad \text{and} \quad \lambda_{i,\beta}(\vartheta) := \max(\lambda_i(\vartheta), \beta)\,.$$

For all $\vartheta_1, \vartheta_2 \in \Theta$, we have

$$|\lambda_{i,\beta}(\vartheta_1) - \lambda_{i,\beta}(\vartheta_2)| \le |\lambda_i(\vartheta_1) - \lambda_i(\vartheta_2)|\,, \tag{18}$$

hence

$$\begin{aligned}
|G(\vartheta_1) - G(\vartheta_2)| &= \left| \max_{i=1\dots d} \lambda_{i,\beta}(\vartheta_1)\langle v, u_i(\vartheta_1)\rangle^2 - \max_{i=1\dots d} \lambda_{i,\beta}(\vartheta_2)\langle v, u_i(\vartheta_2)\rangle^2 \right| \\
&\le \max_{i=1\dots d} \left| \lambda_{i,\beta}(\vartheta_1)\langle v, u_i(\vartheta_1)\rangle^2 - \lambda_{i,\beta}(\vartheta_2)\langle v, u_i(\vartheta_2)\rangle^2 \right| \\
&= \max_{i=1\dots d} \big| \big(\lambda_{i,\beta}(\vartheta_1) - \lambda_{i,\beta}(\vartheta_2)\big)\, \langle v, u_i(\vartheta_1)\rangle^2 \\
&\qquad + \lambda_{i,\beta}(\vartheta_2)\langle v, u_i(\vartheta_1) + u_i(\vartheta_2)\rangle \langle v, u_i(\vartheta_1) - u_i(\vartheta_2)\rangle \big| \\
&\le |v|^2 \max_{i=1\dots d} \big(|\lambda_{i,\beta}(\vartheta_1) - \lambda_{i,\beta}(\vartheta_2)|\, |u_i(\vartheta_1)|^2 \\
&\qquad + \lambda_{i,\beta}(\vartheta_2) |u_i(\vartheta_1) + u_i(\vartheta_2)|\, |u_i(\vartheta_1) - u_i(\vartheta_2)| \big) \\
&\le |v|^2 \max_{i=1\dots d} \big(|\lambda_i(\vartheta_1) - \lambda_i(\vartheta_2)|\, |u_i(\vartheta_1)|^2 + 2\lambda_{i,\beta}(\vartheta_2) |u_i(\vartheta_1) - u_i(\vartheta_2)| \big) \\
&= \frac{(2\mu+1)L|\vartheta_1 - \vartheta_2|}{\beta}\, \beta |v|^2 \\
&\le \omega_\beta(|\vartheta_1 - \vartheta_2|) [v]^2_{A_\beta(\vartheta_1)}\,,
\end{aligned}$$

and the proof is concluded. □

The Fisher box centered in $\vartheta_0 \in \Theta$ of radius $\varepsilon > 0$ is defined as:

$$\mathrm{Box}_\varepsilon(\vartheta_0) := \left\{ \vartheta \in \Theta :\ [\vartheta - \vartheta_0]_{F(\vartheta_0)} < \varepsilon \right\}\,. \tag{19}$$

Lemma B.2. *Let $0 < \varepsilon < 1$, $\zeta \in [\frac{2}{3}, 1)$, $\Theta = [0,1]^d$, and assume that the Fisher matrix $F(\vartheta)$ is L-Lipschitz and that $\lambda_i(\vartheta) \le \mu$ for all $i = 1, \dots, d$ and for all $\vartheta \in \Theta$. Then, F admits an L-Lipschitz extension to the whole $\mathbb{R}^d$, and Θ can be covered by $C_d\, \varepsilon^{-d_\zeta(\varepsilon)}$ Fisher boxes of radius ε, where $d_\zeta(\varepsilon)$ is as in (6), and C_d is a dimensional constant.*

Proof. The fact that any L-Lipschitz mapping from a subset of $\mathbb{R}^d$ into $\mathbb{R}^m$ admits an L-Lipschitz extension to the whole $\mathbb{R}^d$ is classically known as Kirszbraun's Theorem. Consider now a partition $\mathcal{Q}$ of Θ made by closed cubes with mutually disjoint interior and side $\delta = \delta(Q) = \varepsilon^\zeta$, and let Q be one of these cubes. Set

$$\beta = \varepsilon^2/\delta^2 = \varepsilon^{2-2\zeta}\,, \tag{20}$$

and fix a generic $\vartheta_Q \in Q$, then define the β-Fisher box of radius ε and center ϑ_Q as

$$\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q) = \left\{ \vartheta \in \Theta :\ [\vartheta - \vartheta_Q]_{F_\beta(\vartheta_Q)} < \varepsilon \right\}\,.$$

Let S_Q be the Euclidean ball circumscribed to Q. Consider a partition of $\mathbb{R}^d$ by means of translated copies of $\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q)$, then the minimum number of such boxes that have a nonempty intersection with Q is bounded from above by the number $\tilde{k} = \tilde{k}(Q, \beta, \varepsilon)$ of boxes that have a nonempty intersection with S_Q. The volume of each copy of $\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q)$ is given by

$$|\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q)| = \prod_{i=1}^d \frac{2\varepsilon}{\sqrt{\lambda_{i,\beta}(\vartheta_Q)}}\,.$$

At the same time, the union of the covering boxes is contained in $S_Q + B_{2\varepsilon\sqrt{d}/\sqrt{\beta}}$, i.e., in a Euclidean ball B_R with

$$R = \sqrt{d}\left(\frac{\delta}{2} + 2\frac{\varepsilon}{\sqrt{\beta}}\right) = \frac{5}{2}\sqrt{d}\delta\,, \tag{21}$$

hence its volume is bounded from above by $|B_R| = \alpha_d R^d$, where $\alpha_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of an Euclidean ball of radius 1 in $\mathbb{R}^d$, and $\Gamma(\cdot)$ is Euler's Gamma function. Therefore we can estimate $\tilde{k}$ from above by the ratio between the upper bound on the volume of the union of the boxes and the volume of a single box. We obtain

$$\begin{aligned}\tilde{k} &\leq \frac{|B_R|}{|\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q)|} = \frac{\alpha_d \left(\sqrt{d}(\delta/2 + 2\varepsilon/\sqrt{\beta})\right)^d}{\prod_{i=1}^d \frac{2\varepsilon}{\sqrt{\lambda_{i,\beta}(\vartheta_Q)}}} = \frac{\alpha_d \left(5/2\sqrt{d}\delta\right)^d}{\prod_{i=1}^d \frac{2\varepsilon}{\sqrt{\lambda_{i,\beta}(\vartheta_Q)}}} \\ &\leq c_d \prod_{i=1}^d \left\lceil \sqrt{\frac{\lambda_{i,\beta}(\vartheta_Q)}{\beta}} \right\rceil = c_d \prod_{i=1}^d \left\lceil \sqrt{\frac{\lambda_i(\vartheta_Q)}{\beta}} \right\rceil ,\end{aligned}$$

where we have used the special rounding function

$$\lceil x \rceil = \min\{k \in \mathbb{N} : k \geq \max(x, 1)\}$$

(note that $\lceil x \rceil \geq 1$ for all x), and where $c_d = \alpha_d (5/4)^d d^{d/2}$. Note that, by Stirling's formula $\Gamma(x+1) \sim \sqrt{2\pi x}(x/e)^x$ valid as $x \to +\infty$, we deduce that $c_d \leq 4(25/8\pi e)^{d/2}$ for d large enough.

Now we notice that for all $\vartheta \in S_\varepsilon$ the translated copy of $B_{\beta,\varepsilon}(\vartheta_Q)$ centered in ϑ is contained in $B_{\beta,\varepsilon'}(\vartheta)$, with

$$\varepsilon' = \varepsilon \sqrt{1 + \omega_\beta(5\sqrt{d}\delta)}\,.$$

Indeed, let ϑ be the center of the translated copy $\widetilde{\mathrm{Box}}$ of $\mathrm{Box}_{\beta,\varepsilon}(\vartheta_Q)$, then for each $\xi \in S_Q \cap \widetilde{\mathrm{Box}}$ one has by definition $[\xi - \vartheta]_{F_\beta(\vartheta_Q)} < \varepsilon$. Consequently, by Lemma B.1 and by the fact that both ϑ and ϑ_Q are contained in B_R, one gets

$$\begin{aligned}[\xi - \vartheta]_{F_\beta(\vartheta)} &\leq [\xi - \vartheta]_{F_\beta(\vartheta_Q)} \sqrt{1 + \omega_\beta(|\vartheta - \vartheta_Q|)} \\ &\leq \varepsilon \sqrt{1 + \omega_\beta(|\vartheta - \vartheta_Q|)} \\ &\leq \varepsilon \sqrt{1 + \omega_\beta(5\sqrt{d}\delta)}\,.\end{aligned}$$

The previous estimate shows that there exists a covering of Q by means of at most k_Q boxes of the form $\mathrm{Box}_j = \mathrm{Box}_{\beta,\varepsilon'}(\vartheta_j)$, with $j = 1, \ldots, k_Q$, with

$$\begin{aligned}k_Q &\leq \tilde{k} \\ &\leq c_d \prod_{i=1}^d \left\lceil \sqrt{\delta^2(1 + \omega_\beta(5\sqrt{d}\delta))\lambda_i(\vartheta_Q)/(\varepsilon')^2} \right\rceil \\ &\leq\ c_d (1 + \omega_\beta(5\sqrt{d}\delta))^{\frac{d}{2}} |Q| \prod_{i=1}^d \left(\varepsilon^{-1}\sqrt{\lambda_i(\vartheta_Q)} + \delta^{-1}\right) ,\end{aligned}$$

and where we have used that $\lceil xy \rceil \leq xy + 1$ for all $x, y \geq 0$. If now we assume $\zeta \geq 2/3$, we have

$$\omega_\beta(5\sqrt{d}\delta) = 5\sqrt{d}(2\mu+1)L\varepsilon^{3\zeta-2} \leq 5\sqrt{d}(2\mu+1)L\,,$$

therefore we conclude

$$k_Q \leq C_d\, \varepsilon^{-\zeta d} |Q| \prod_{i=1}^d \left(1 + \varepsilon^{\zeta-1}\sqrt{\lambda_i(\vartheta_Q)}\right) = C_d\, \varepsilon^{-\zeta d} |Q| \det\left(I + \varepsilon^{\zeta-1} F(\vartheta_Q)^{\frac{1}{2}}\right) ,$$

where $C_d = 5\sqrt{d} c_d (2\mu+1)L$. Finally, if we denote by $k(\varepsilon)$ the cardinality of the least number of Fisher boxes of size ε that are needed to cover Θ, by summing over Q and choosing ϑ_Q as a minimum point for $\det(I + \varepsilon^{\zeta-1} F(\vartheta)^{\frac{1}{2}})$ when $\vartheta \in Q$, we obtain

$$k(\varepsilon) \leq C_d\, \varepsilon^{-\zeta d} \int_\Theta \det\left(I + \varepsilon^{\zeta-1} F(\vartheta)^{\frac{1}{2}}\right) d\vartheta = C_d\ \varepsilon^{-d_\zeta(\varepsilon)}\,,$$

as wanted. □

The following result exploits the link between the generalization bound and the covering bound proved in Lemma B.2.

Lemma B.3. *Under the assumption of Theorem 4.1, there exist $\varepsilon_0, C, K > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ we have*

$$\mathbb{P}\left\{\sup_{\vartheta\in\Theta} |R(\vartheta) - R_n(\vartheta)| \geq C\varepsilon\right\} \leq 4\, k(\varepsilon)\, \exp\left(-Kn\varepsilon^{8/3}\right), \tag{22}$$

where $k(\varepsilon)$ is a bound on the cardinality of a covering by β_ε-Fisher boxes of radius ε, with $\beta_\varepsilon = \varepsilon^{2-2\zeta}$ (see Lemma B.2).

Proof. As a first step, we need to "discretize" the estimate of the left-hand side of (22) at the micro-scale ε, using the β-Fisher box covering from Lemma B.2, with $\beta = \varepsilon^{2-2\zeta}$. Recalling that $S_n(\vartheta) = R(\vartheta) - R_n(\vartheta)$, for all $\vartheta_1, \vartheta_2 \in \Theta$ such that $|\vartheta_1 - \vartheta_2| < 2R$ (where R is defined in (21)) we have

$$|S_n(\vartheta_1) - S_n(\vartheta_2)| \leq |R(\vartheta_1) - R(\vartheta_2)| + |R_n(\vartheta_1) - R_n(\vartheta_2)|\,. \tag{23}$$

Now we estimate each term in the right-hand side of (23). We set $\vartheta(t) = t\vartheta_1 + (1-t)\vartheta_2$ for $t \in [0,1]$, and we estimate the first term:

$$\begin{aligned}
|R(\vartheta_1) - R(\vartheta_2)| &\leq \int_{\mathcal{X}\times\mathcal{Y}} |\mathfrak{L}(p_{\vartheta_1}(y|x)) - \mathfrak{L}(p_{\vartheta_2}(y|x))|\, p(dx,dy) \\
&\leq \int_{\mathcal{X}\times\mathcal{Y}} \left|\mathfrak{L}(p_{\vartheta(0)}(y|x)) - \mathfrak{L}(p_{\vartheta(1)}(y|x))\right| p(dx,dy) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \left| \int_0^1 \partial_1\mathcal{L}(p_{\vartheta(t)}(y|x), p(y|x)) \left\langle \nabla_\vartheta p_{\vartheta(t)}(y|x), \vartheta_2 - \vartheta_1 \right\rangle dt \right| p(dx,dy) \\
&\leq \int_0^1 \int_{\mathcal{X}\times\mathcal{Y}} \left|\partial_1\mathcal{L}(p_{\vartheta(t)}(y|x), p(y|x))\right| \left| \left\langle \nabla_\vartheta p_{\vartheta(t)}(y|x), \vartheta_2 - \vartheta_1 \right\rangle \right| p(dx,dy)dt \\
&\leq \Lambda \int_0^1 \int_{\mathcal{X}\times\mathcal{Y}} \left| \left\langle \nabla_\vartheta p_{\vartheta(t)}(x,y), \vartheta_2 - \vartheta_1 \right\rangle \right| p(dx,dy)\, dt \\
&= \Lambda \int_0^1 \int_{\mathcal{X}\times\mathcal{Y}} \left| \left\langle \nabla_\vartheta \log p_{\vartheta(t)}(x,y), \vartheta_2 - \vartheta_1 \right\rangle \right| p_{\vartheta(t)}(x,y)\, p(dx,dy)\, dt \\
&= \Lambda \int_0^1 \int_{\mathcal{X}\times\mathcal{Y}} \left| \left\langle \nabla_\vartheta \log p_{\vartheta(t)}(x,y), \vartheta_2 - \vartheta_1 \right\rangle \right| p(x,y)\, p_{\vartheta(t)}(dx,dy)\, dt\,,
\end{aligned}$$

where we have used the fundamental theorem of calculus, Fubini's theorem, the Λ-Lipschitzianity of $\mathcal{L}$, and the fact that $\nabla_\vartheta \log p_\vartheta(y|x) = \nabla_\vartheta \log p_\vartheta(x,y)$. Then, by Cauchy-Schwarz inequality, we obtain for all $\beta > 0$:

$$\begin{aligned}
|R(\vartheta_1) - R(\vartheta_2)| &\leq \Lambda \int_0^1 \mathbb{E}_{p_{\vartheta(t)}}[p^2(x,y)]^{1/2} \cdot \left\langle F(\vartheta(t))(\vartheta_2 - \vartheta_1), \vartheta_2 - \vartheta_1 \right\rangle^{1/2} dt \\
&\leq \Lambda \int_0^1 \mathbb{E}_{p_{\vartheta(t)}}[p^2(x,y)]^{1/2} \cdot \left\langle F_\beta(\vartheta(t))(\vartheta_2 - \vartheta_1), \vartheta_2 - \vartheta_1 \right\rangle^{1/2} dt \\
&\leq \Lambda C_1 \int_0^1 \left\langle F_\beta(\vartheta(t))(\vartheta_2 - \vartheta_1), \vartheta_2 - \vartheta_1 \right\rangle^{1/2} dt\,,
\end{aligned}$$

for some constant $C_1 > 0$ depending on α_1, α_2, thanks to hypothesis (ii). Now, Lemma B.1 implies that

$$\left\langle F_\beta(\vartheta(t))(\vartheta_2 - \vartheta_1), \vartheta_2 - \vartheta_1 \right\rangle \leq d\big(1 + \omega_\beta(t|\vartheta_2 - \vartheta_1|)\big)[\vartheta_2 - \vartheta_1]_{F_\beta(\vartheta_1)}\,. \tag{24}$$

By (24) we conclude that

$$|R(\vartheta_1) - R(\vartheta_2)| \le \Lambda C_1 d \int_0^1 \left(1 + \omega_\beta(t|\vartheta_2 - \vartheta_1|)\right)^{1/2} dt \; [\vartheta_2 - \vartheta_1]_{F_\beta(\vartheta_1)}$$
$$\le C_2 [\vartheta_2 - \vartheta_1]_{F_\beta(\vartheta_1)}\,, \tag{25}$$

where C_2 is a constant depending only on Λ, C_1 and the dimension d. Indeed, we observe that the assumptions $\beta = \varepsilon^{2-2\zeta}$ and $\zeta \ge 2/3$, coupled with $|\vartheta_1 - \vartheta_2| \le 2R$, imply that

$$\omega_\beta(t|\vartheta_2 - \vartheta_1|) \le 5(2\mu+1)L\sqrt{d}\varepsilon^{3\zeta-2} \le 5(2\mu+1)L\sqrt{d}\,. \tag{26}$$

Now we estimate the second term in the r.h.s. of (23). By a similar computation we obtain

$$|R_n(\vartheta_1) - R_n(\vartheta_2)| \le \Lambda \int_0^1 \left(\frac{1}{n}\sum_{i=1}^n \left\langle \nabla \log p_{\vartheta(t)}(X_i, Y_i), \vartheta_2 - \vartheta_1 \right\rangle^2 \frac{p_{\vartheta(t)}(X_i, Y_i)}{p(X_i, Y_i)} \right)^{1/2}$$
$$\cdot \left(\frac{1}{n}\sum_{i=1}^n p_{\vartheta(t)}(X_i, Y_i) p(X_i, Y_i) \right)^{1/2} dt$$
$$\le \alpha_2 \int_0^1 \left(\frac{1}{n}\sum_{i=1}^n \left\langle \nabla \log p_{\vartheta(t)}(X_i, Y_i), \vartheta_2 - \vartheta_1 \right\rangle^2 \frac{p_{\vartheta(t)}(X_i, Y_i)}{p(X_i, Y_i)} \right)^{1/2} dt\,. \tag{27}$$

Let us set

$$Z_i(t) := \left\langle \nabla \log p_{\vartheta(t)}(X_i, Y_i), \vartheta_2 - \vartheta_1 \right\rangle^2 \frac{p_{\vartheta(t)}(X_i, Y_i)}{p(X_i, Y_i)}$$

and

$$T := \sup \frac{p_\vartheta(x,y)}{p(x,y)} |\nabla_\vartheta \log p_\vartheta(x,y)|^2\,,$$

where the supremum is computed w.r.t $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $\vartheta \in \Theta$. By (i) and (ii) we obtain

$$T \le B \sup |\nabla_\vartheta \log p_\vartheta(x,y)|^2 < \infty\,,$$

where $B = \alpha_2/\alpha_1$. Thus we also get

$$0 \le Z_i(t) \le T|\vartheta_2 - \vartheta_1|^2\,.$$

The expectation of $Z_i(t)$ is

$$\overline{Z}_i(t) = \mathbb{E}_{(x,y)\sim p}[Z_i(t)] = \int \left\langle \nabla \log p_{\vartheta(t)}(x,y), \vartheta_2 - \vartheta_1 \right\rangle^2 \frac{p_{\vartheta(t)}(x,y)}{p(x,y)}\, p(dx,dy)$$
$$= \int \left\langle \nabla \log p_{\vartheta(t)}(x,y), \vartheta_2 - \vartheta_1 \right\rangle^2 p_{\vartheta(t)}(dx,dy)$$
$$= \left\langle F(\vartheta(t))(\vartheta_2 - \vartheta_1), \vartheta_2 - \vartheta_1 \right\rangle$$

hence also $\frac{1}{n}\sum_{i=1}^n Z_i(t)$ has the same expectation, by independence of the $Z_i(t)$.

Now, from Lemma B.2 we know that Θ can be covered with $k = k(\varepsilon) \le C_d \varepsilon^{-d_\zeta(\varepsilon)}$ Fisher boxes $\mathrm{Box}_1, \dots, \mathrm{Box}_k$ of size ε. Let now $\eta = C\varepsilon$ for some $C > 0$ to be chosen later, and evaluate

$$\mathbb{P}\left\{ \sup_{\vartheta\in\Theta} |S_n(\vartheta)| \ge \eta \right\} \le \mathbb{P}\left\{ \bigcup_{j=1}^k \sup_{\vartheta\in \mathrm{Box}_j} |S_n(\vartheta)| \ge \eta \right\} \le \sum_{j=1}^k \mathbb{P}\left\{ \sup_{\vartheta\in\mathrm{Box}_j} |S_n(\vartheta)| \ge \eta \right\}.$$

Now for all $j = 1, \dots, k$ we bound the probability of an event involving the computation of the supremum of $|S_n(\vartheta)|$ over Box_j with another one involving only the pointwise evaluation of S_n

at the center ϑ_j of Box_j. Indeed by (25) and (27), and with ϑ, ϑ_j respectively replacing ϑ_2, ϑ_1, we deduce

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\vartheta\in\mathrm{Box}_j}|S_n(\vartheta)|\geq\eta\right\}\\
&\quad\leq\mathbb{P}\left\{|S_n(\vartheta_j)|+\sup_{\vartheta\in\mathrm{Box}_j}\big(|S_n(\vartheta)-S_n(\vartheta_j)|\big)\geq\eta\right\}\\
&\quad\leq\mathbb{P}\left\{|S_n(\vartheta_j)|+\sup_{\vartheta\in\mathrm{Box}_j}\left(C_2[\vartheta-\vartheta_j]_{F_\beta(\vartheta_j)}+\alpha_2\int_0^1\left(\frac{1}{n}\sum_{i=1}^n Z_i(t)\right)^{1/2}dt\right)\geq\eta\right\}\\
&\quad\leq\mathbb{P}\left\{|S_n(\vartheta_j)|\geq\frac{\eta}{2}\right\}+\mathbb{P}\left\{\exists\, t\in[0,1]:\ \frac{1}{n}\sum_{i=1}^n Z_i(t)\geq\frac{\eta^2}{16\,\alpha_2^2}\right\},
\end{aligned}$$

where in the last inequality we have used $[\vartheta-\vartheta_j]_{F_\beta(\vartheta_j)}<\varepsilon=\frac{\eta}{C}$ and required $C\geq 4C_2$. Owing to Lemma B.4 and (v), we get

$$\mathbb{P}\left(|S_n(\vartheta_j)|\geq\frac{\eta}{2}\right)=\mathbb{P}\left(|R_n(\vartheta_j)-R(\vartheta_j)|\geq\frac{\eta}{2}\right)\leq 2\exp\left(-\frac{n\eta^2}{2b^2}\right). \tag{28}$$

and

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i(t)-\Big\langle F(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle\right|\geq\xi\right\}\leq 2\exp\left(-\frac{2n\xi^2}{T^2|\vartheta-\vartheta_j|^2}\right). \tag{29}$$

By (29) we find

$$\begin{aligned}
&\mathbb{P}\left\{\exists\, t\in[0,1]:\ \frac{1}{n}\sum_{i=1}^n Z_i(t)\geq\frac{\eta^2}{16\,\alpha_2^2}\right\}\\
&\qquad\leq\ \mathbb{P}\left\{\Big\langle F(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle\geq\frac{\eta^2}{32\,\alpha_2^2}\right\}\\
&\qquad\qquad+\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n Z_i(t)-\Big\langle F(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle\geq\frac{\eta^2}{32\,\alpha_2^2}\right\}\\
&\qquad\leq\mathbb{P}\left\{\Big\langle F_\beta(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle\geq\frac{\eta^2}{32\,\alpha_2^2}\right\}+2\exp\left(-\frac{n\eta^4}{2^9\,\alpha_2^4T^2|\vartheta-\vartheta_j|^2}\right)\\
&\qquad\leq 2\exp\left(-C_4n\eta^{4-2\zeta}\right),
\end{aligned} \tag{30}$$

where

$$C_4=\frac{C_2^{2\zeta}}{3^2 2^{9-2\zeta}\alpha_2^4T^2d}$$

and where the last inequality follows from the impossibility of the first event, up to a further update of the constant C, that we shall explain hereafter. Indeed, using Lemma B.1:

$$\begin{aligned}
\Big\langle F_\beta(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle &\le \Big\langle F_\beta(\vartheta_j)(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle + \Big\langle (F_\beta(\vartheta(t))-F_\beta(\vartheta_j))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle \\
&\le \Big\langle F_\beta(\vartheta_j)(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle + \Big|\Big\langle (F_\beta(\vartheta(t))-F_\beta(\vartheta_j))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle\Big| \\
&\le d\left([\vartheta-\vartheta]^2_{F_\beta(\vartheta_j)} + \Big|[\vartheta-\vartheta_j]^2_{F_\beta(\vartheta_j)} - [\vartheta-\vartheta_j]^2_{F_\beta(\vartheta)}\Big|\right) \\
&\le d([\vartheta-\vartheta]^2_{F_\beta(\vartheta_j)} + \omega_\beta(t|\vartheta-\vartheta_j|)[\vartheta-\vartheta]^2_{F_\beta(\vartheta_j)}) \\
&= d(1+\omega_\beta(t|\vartheta-\vartheta_j|))[\vartheta-\vartheta]^2_{F_\beta(\vartheta_j)} \\
&\le d\big(1+\omega_\beta(t|\vartheta-\vartheta_j|)\big)\,\varepsilon^2 \le d(5(2\mu+1)L\sqrt{d}\varepsilon^{3\zeta}+\varepsilon^2) \\
&\le d\frac{2+5(2\mu+1)L\sqrt{d}}{2C^2}\eta^2\,,
\end{aligned}$$

where the last two inequalities follow from (26) and (v). Therefore, if we choose C such that

$$d(2+5(2\mu+1)L\sqrt{d}) < \frac{C^2}{16\alpha_2^2}\,,$$

we obtain

$$\Big\langle F_\beta(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle < \frac{\eta^2}{32\alpha_2^2}$$

and thus we enforce, as claimed, the impossibility of the event

$$\Big\langle F_\beta(\vartheta(t))(\vartheta-\vartheta_j),\vartheta-\vartheta_j\Big\rangle \ge \frac{\eta^2}{32\,\alpha_2^2}\,,$$

which completes the proof of (30).

Finally by (28) and (30), and observing that the second exponential represents the leading term, we get

$$\begin{aligned}
\mathbb{P}\left(\sup_{\vartheta\in\Theta}|S_n(\vartheta)|\ge\eta\right) &\le \sum_{i=1}^{k}\mathbb{P}\left(\sup_{\vartheta\in\mathrm{Box}_i}|S_n(\vartheta)|\ge\eta\right) \\
&\le 2\,k(\varepsilon)\left[\exp\left(-\frac{n\eta^2}{2b^2}\right)+\exp\left(-C_4 n\eta^{4-2\zeta}\right)\right] \\
&\le 4\,k(\varepsilon)\ \exp\left(-C_5 n\eta^{8/3}\right)\,,
\end{aligned}$$

where $C_5=\min(C_4,(2b^2)^{-1})$ (and since $4-2\zeta<3$ by our assumption on ζ). In conclusion we obtain (22) with $K=C_5C^{8/3}$. □

Proof of Theorem 4.1. We choose $\gamma>0$, and let $\varepsilon_n=\left(\frac{\log n}{\gamma n}\right)^{3/8}$ and K be as in Lemma B.3. By combining Lemma B.2 with Lemma B.3 we obtain

$$\begin{aligned}
\mathbb{P}\left(\sup_{\vartheta\in\Theta}|R(\vartheta)-R_n(\vartheta)|\ge C\varepsilon_n\right) &\le 4\,k(\varepsilon_n)\exp\left(-K\frac{\log n}{\gamma}\right) \\
&\le H\varepsilon_n^{-d_\zeta(\varepsilon_n)}n^{-\frac{K}{\gamma}}\,,
\end{aligned}$$

where C is as in Lemma B.3 and $H=4C_d$. □

We can now explain Remark 4.2 by noting that, if we choose

$$0<\gamma<\gamma_0 := \frac{8K}{3d(1+\log(1+\mu^{1/2}))}\,, \tag{31}$$

with K as in Lemma B.3, then the generalization bound becomes infinitesimal as $n \to +\infty$. Indeed, by the upper estimate (13) we have

$$d_\zeta(\varepsilon) \leq \zeta d + \hat{r}\left(1 - \zeta + \frac{\log(1+\mu^{1/2})}{|\log(\varepsilon)|}\right) \leq d(1+\log(1+\mu^{1/2})) =: \bar{d}\,,$$

whenever $\varepsilon < \exp(-1)$, so that

$$\varepsilon_n^{-d_\zeta(\varepsilon_n)} n^{-\frac{K}{\gamma}} = \left(\frac{\gamma n}{\log n}\right)^{3d_\zeta(\varepsilon_n)/8} n^{-\frac{K}{\gamma}} \leq \gamma^{3\bar{d}/8} n^{3\bar{d}/8 - K/\gamma}\,. \tag{32}$$

Hence, the infinitesimality of the generalization bound as $n \to \infty$ follows from $3\bar{d}/8 - K/\gamma < 0$, as wanted.

We recall Hoeffding's estimate, which is used in the proof of Lemma B.3.

Lemma B.4 (Hoeffding's estimate)**.** *Let Z_i, $i = 1, \ldots, n$, be independent random variables, such that $Z_i \in [a, b]$ almost surely. Define $V_n = \frac{1}{n}\sum_{i=1}^n Z_i$ and take $\varepsilon > 0$, then*

$$\mathbb{P}\left(\left|V_n - \mathbb{E}[V_n]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)\,.$$

Appendix C. Figures

The appendix contains a comprehensive collection of figures and tables that complement and enhance the understanding of the main content presented in this document. These figures provide visual representations the results related to the experiments section discussed in the main part of the paper.

Model	**Number of Parameters**
CNN 7-5|10-50-34-10	4493
CNN 5-7|10-50-34-10	4753
CNN 5-4-3|10-50-34-10	4985
CNN 3-2-3-4|10-50-34-10	5749
CNN 3-4-3-2|10-50-34-10	5749

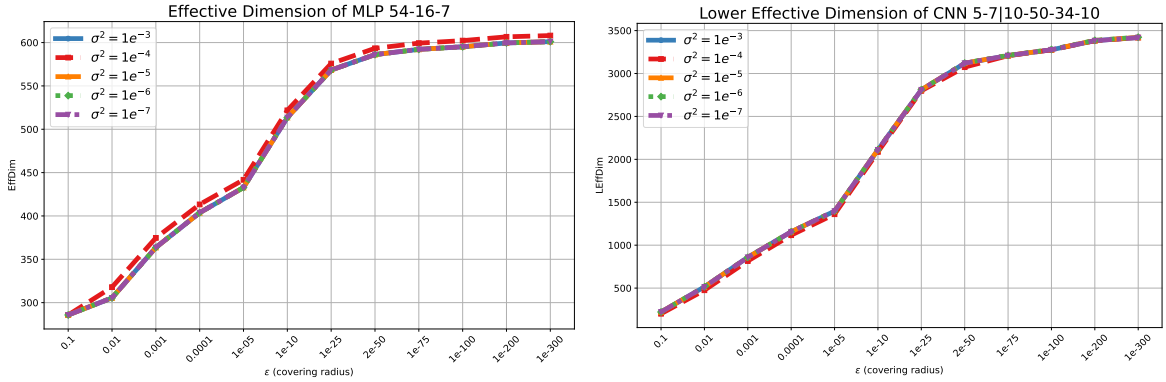
Table 2. Number of parameters of CNNs

(Massimiliano Datres) Dipartimento di Matematica, Università di Trento, via Sommarive 14, IT-38123 Povo - Trento (Italy)
Email address: massimiliano.datres@unitn.it

(Gian Paolo Leonardi) Dipartimento di Matematica, Università di Trento, via Sommarive 14, IT-38123 Povo - Trento (Italy)
Email address: gianpaolo.leonardi@unitn.it

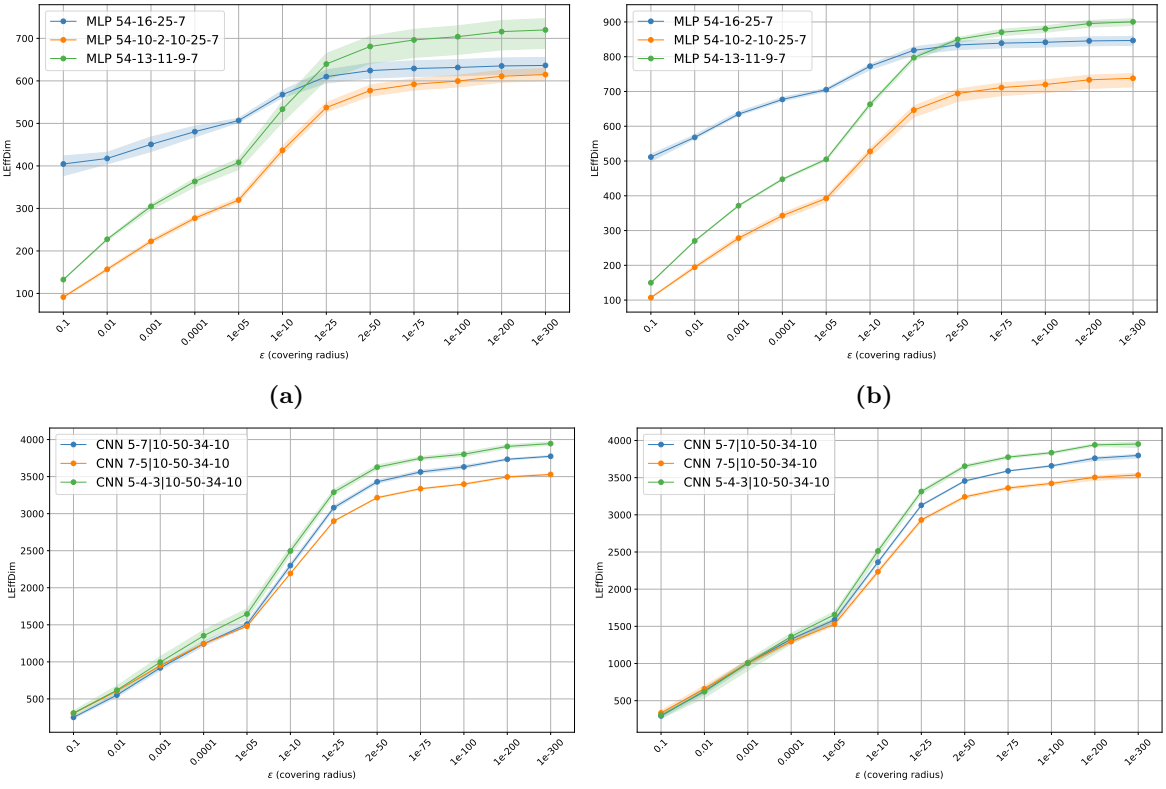
(Alessio Figalli) Department of Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zürich, Switzerland
Email address: alessio.figalli@math.ethz.ch

(David Sutter) IBM Quantum, IBM Research Europe – Zurich, Switzerland
Email address: dsu@zurich.ibm.com



(a) **(b)**

Figure 3. Figure **(a)** and Figure **(b)** show that the impact of σ^2 on $\underline{d}_\zeta$ and d_ζ is negligible both for MLPs and CNNs.
(a) 2sED of MLP 54-16-7 with fixed seed varying σ^2; **(b)** Lower 2sED of CNN 5-7—10-50-34-10 with fixed seed varying σ^2



(a) **(b)**

(c) **(d)**

Figure 4. These experiments confirms the stability of the lower 2sED with respect to the number of points (and the points themself) used in the Monte Carlo approximation. **(a)** Estimated lower 2sED of MLPs using 100 Covertype samples and 100 parametrization with error margin; **(b)** Estimated lower 2sED of MLPs using 1000 Covertype samples and 1000 parametrization with error margin; **(c)** Estimated lower 2sED of CNNs using 100 MNIST samples and 100 parametrization with error margin; **(d)** Estimated lower 2sED of CNNs using 500 MNIST samples and 100 parametrization with error margin;

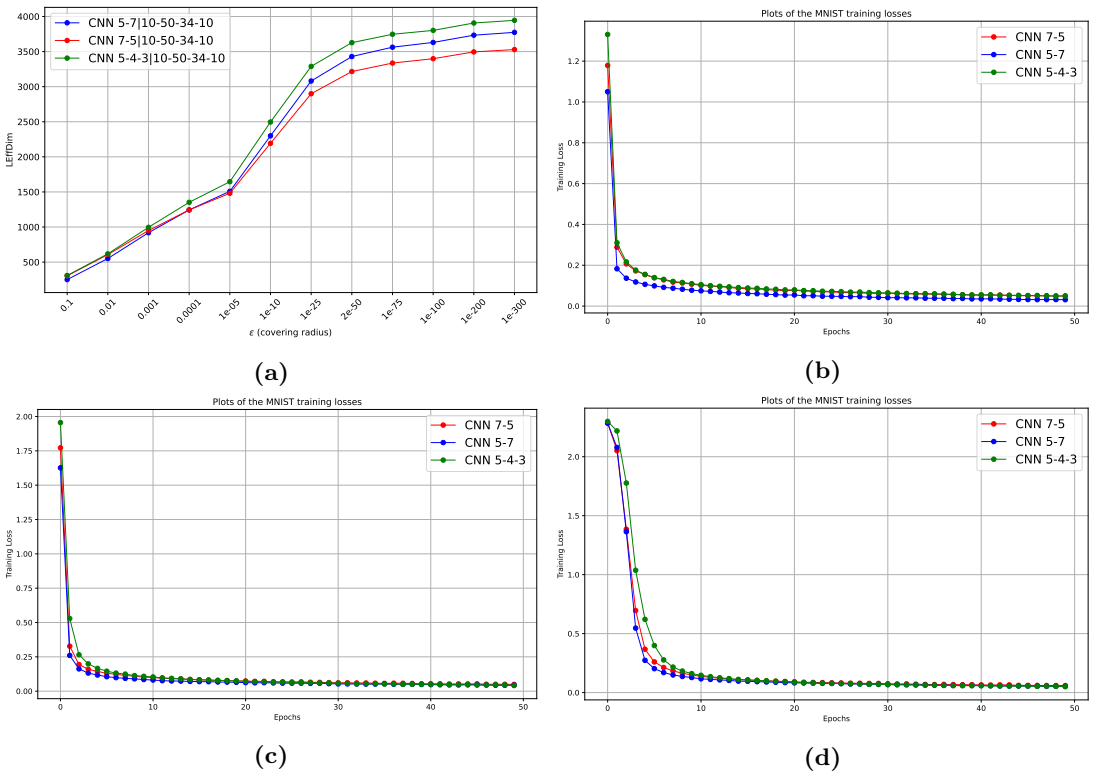


Figure 5. This figures show that the relation between the lower 2sED and the training curves does not depend on the choices of batch size and learning rate.
(a) Lower 2sED of CNNs using 100 MNIST samples and 100 parametrization; (**b**) Training loss plots of CNNs on MNIST using Adam with learning rate $1e^{-3}$ and a batch size 256; **(c)** Training loss plots of CNNs on MNIST using Adam with learning rate $1e^{-3}$ and a batch size 512;**(d)** Training loss plots of CNNs on MNIST using Adam with learning rate $1e^{-3}$ and a batch size 2048;