GENERALIZATION THROUGH THE LENS OF LEAVE-ONE-OUT ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the tremendous empirical success of deep learning models to solve various learning tasks, our theoretical understanding of their generalization ability is very limited. Classical generalization bounds based on tools such as the VC dimension or Rademacher complexity, are so far unsuitable for deep models and it is doubtful that these techniques can yield tight bounds even in the most idealistic settings (Nagarajan & Kolter, 2019). In this work, we instead revisit the concept of leave-one-out error to measure the generalization ability of deep and wide models. While popular in statistics, it has been largely overlooked in the context of deep learning. By building upon the recently established connection between neural networks and kernel learning, we leverage the closed-form expression for the leave-one-out error, giving us access to an efficient proxy for the test error. We show both theoretically and empirically that the leave-one-out error is capable of capturing various phenomena in generalization theory, such as double descent, random labels or transfer learning. Our work therefore demonstrates that the leave-one-out error provides a tractable way to estimate the generalization ability of wide, deep neural networks, opening the door to potential, new research directions in the field of generalization.

1 INTRODUCTION

Neural networks have achieved astonishing performance across many learning tasks such as in computer vision (He et al., 2016), natural language processing (Devlin et al., 2019) and graph learning (Kipf & Welling, 2017) among many others. Despite the large overparametrized nature of these models, they have shown a surprising ability to generalize well to unseen data. The theoretical understanding of this generalization ability has been an active area of research where contributions have been made using diverse analytical tools (Arora et al., 2018; Bartlett et al., 2019; 2017; Neyshabur et al., 2015; 2018). Yet, the theoretical bounds derived by these approaches typically suffer from one or several of the following limitations: i) they are known to be loose or even vacuous (Jiang et al., 2019), ii) they only apply to randomized networks (Dziugaite & Roy, 2018; 2017), and iii) they rely on the concept of uniform convergence which was shown to be non-robust against adversarial perturbations (Nagarajan & Kolter, 2019).

In this work, we revisit the concept of leave-one-out error (LOO) and demonstrate its surprising ability to predict generalization for very wide, deep neural networks. This object is an important statistical estimator of the generalization performance of an algorithm that is theoretically motivated by a connection to the concept of uniform stability (Pontil, 2002). It has also recently been advocated by Nagarajan & Kolter (2019) as a potential substitute for uniform convergence bounds. Despite being popular in classical statistics (Cawley & Talbot, 2003; Vehtari et al., 2016; Fukunaga & Hummels, 1989; Zhang, 2003), LOO has largely been overlooked in the context of deep learning, likely due to its high computational cost. However, recent advances from the neural tangent kernel perspective (Jacot et al., 2018) render the LOO error all of a sudden tractable thanks to the availability of a closed-form expression due to Stone (1974, Eq. 3.13). While the role of the LOO error as an estimator of the generalization performance is debated in the statistics community (Zhang & Yang, 2015; Bengio & Grandvalet, 2004; Breiman, 1996; Kearns & Ron, 1999), the recent work by Patil et al. (2021) established a consistency result in the case of ridge regression, ensuring that LOO converges to the generalization error in the large sample limit, under mild assumptions on the data distribution.

Inspired by these recent advances, in this work, we investigate the use of LOO as a generalization measure for wide, deep neural networks. We find that LOO is a surprisingly rich descriptor of the generalization error, capturing a wide range of phenomena such as random label fitting, double descent and transfer learning. Specifically, we make the following contributions:

- We extend the LOO expression to the multi-class setting and derive a new closed-form formula for the resulting LOO accuracy.
- We investigate both the LOO loss and accuracy in the context of deep learning through the lens of the NTK, and demonstrate empirically that they both capture the generalization ability in a variety of settings.
- We showcase the utility of LOO for practical networks by accurately predicting their transfer learning performance.
- We build on the mathematically convenient form of the LOO loss to derive some novel insights into double descent and the role of regularization.

Our work is structured as follows. We give an overview of related works in Section 2. In Section 3 we introduce the setting along with the LOO error and showcase our extensions to the multi-class case and LOO accuracy. Then, in Section 4 we analyze the predictive power of LOO in various settings and present our theoretical results on double descent. We discuss our findings and future work in Section 5.

2 Related Work

Originally introduced by Lachenbruch (1967); Mosteller & Tukey (1968), the leave-one-out error is a standard tool in the field of statistics that has been used to study and improve a variety of models, ranging from support vector machines (Weston, 1999), Fisher discriminant analysis (Cawley & Talbot, 2003), linear regression (Hastie et al., 2020; Patil et al., 2021) to non-parametric density estimation (Fukunaga & Hummels, 1989). Various theoretical perspectives have justified the use of the LOO error, including for instance the framework of uniform stability (Bousquet & Elisseeff, 2002) or generalization bounds for KNN models (Devroye & Wagner, 1979). Elisseeff et al. (2003) focused on the stability of a learning algorithm to demonstrate a formal link between LOO error and generalization. The vanilla definition of the LOO requires access to a large number of trained models which is typically computationally expensive. Crucially, Stone (1974) derived an elegant closed-form formula for kernels, completely removing the need for training multiple models. In the context of deep learning, the LOO error remains largely unexplored. Shekkizhar & Ortega (2020) propose to fit polytope interpolators to a given neural network and estimate its resulting LOO error. Alaa & van der Schaar (2020) rely on a Jackknife estimator, applied in a LOO fashion to obtain better confidence intervals for Bayesian neural networks.

Recent works have established a connection between kernel regression and deep learning. Surprisingly, it is shown that an infinitely-wide, fully-connected network behaves like a kernel both at initialization (Lee et al., 2018) as well as during gradient flow training (Jacot et al., 2018). Various follow-up works have refined these results, extending the analysis to various architectures (Arora et al., 2019a; Huang et al., 2020; Du et al., 2019) and discrete gradient descent (Lee et al., 2019). The direct connection to the field of kernel regression enables the usage of the closed-form expression for LOO error in the context of deep learning, which to the best of our knowledge has not been explored yet.

3 Setting and Background

In the following, we establish the notations required to describe the problem of interest and the setting we consider. We study the standard learning setting, where we are given a dataset of input-target pairs $\mathcal{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ where each $(\boldsymbol{x}_i, \boldsymbol{y}_i) \overset{i.i.d.}{\sim} \mathcal{D}$ for $i = 1, \ldots, n$ is distributed according to some probability measure $\mathcal{D}$. We refer to $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ as the input and to $\boldsymbol{y} \in \mathcal{Y} \subset \mathbb{R}^K$ as the target. We will often use matrix notations to obtain more compact formulations, thus summarizing all inputs and targets as matrices $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times K}$. We will mainly be

concerned with the task of multiclass classification, where targets $\boldsymbol{y}_i$ are encoded as one-hot vectors. We consider a **function space** $\mathcal{F}$ and an associated **loss function** $L_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that measures how well a given function $f \in \mathcal{F}$ performs at predicting the targets. More concretely, we define the regularized empirical error as

$$\hat{L}_{\mathcal{S}}^{\lambda} : \mathcal{F} \to \mathbb{R}, \quad f \mapsto \frac{1}{n} \sum_{i=1}^{n} L_f(\boldsymbol{x}_i, \boldsymbol{y}_i) + \lambda \Omega(f),$$

where $\lambda > 0$ is a regularization parameter and $\Omega : \mathcal{F} \to \mathbb{R}$ is a functional. In the context of classification, we also define the **empirical accuracy** $\hat{A}_{\mathcal{S}} : \mathcal{F} \to \mathbb{R}$ as the number of instances for which $f \in \mathcal{F}$ provides the correct prediction. For $\boldsymbol{v} \in \mathbb{R}^K$, let $v^* = \operatorname{argmax}_{k \leq K} v_k$ and $a_f(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{1}_{\{f^*(\boldsymbol{x}) = y^*\}}$. Then we can write the empirical accuracy $\hat{A}_{\mathcal{S}}$ as

$$\hat{A}_{\mathcal{S}} : \mathcal{F} \to [0, 1], \quad f \mapsto \frac{1}{n} \sum_{i=1}^{n} a_f(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

Finally, we introduce a **learning algorithm** $\mathcal{Q} : \mathcal{F} \times 2^{\mathcal{X}^n} \to \mathcal{F}$, that given a function class $\mathcal{F}$ and a training set $\mathcal{S} \subset \mathcal{X}^n$, chooses a function $f \in \mathcal{F}$. We will exclusively be concerned with learning through empirical risk minimization, i.e.

$$\mathcal{Q}_{\lambda}^{\text{ERM}}(\mathcal{F}, \mathcal{S}) := \hat{f}_{\mathcal{S}}^{\lambda} := \operatorname{argmin}_{f \in \mathcal{F}} \hat{L}_{\mathcal{S}}^{\lambda}(f),$$

and use the short-cut $\hat{f}_{\mathcal{S}} := \hat{f}_{\mathcal{S}}^{\lambda=0}$. In practice however, the most important measures are given by the **generalization** loss and accuracy of the model, defined as

$$L : \mathcal{F} \to \mathbb{R}, \quad f \mapsto \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \left[L_f(\boldsymbol{x}, \boldsymbol{y}) \right], \qquad A : \mathcal{F} \to \mathbb{R}, \quad f \mapsto \mathbb{P}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \left(f^*(\boldsymbol{x}) = y^* \right).$$

A central goal in machine learning is to control the difference between the empirical error $\hat{L}_{\mathcal{S}}(f)$ and the generalization error $L(f)$. In this work, we mainly focus on the mean squared loss,

$$L_f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \sum_{k=1}^{K} (f_k(\boldsymbol{x}) - y_k)^2,$$

and also treat classification as a regression task, as commonly done in the literature (Chen et al., 2020; Arora et al., 2019a; Bachmann et al., 2021). Finally, we mainly consider the function spaces induced by kernels and neural networks, which we present in the following.

3.1 Kernel Learning

A kernel $\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric, positive semi-definite function, i.e. $\mathrm{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{K}(\boldsymbol{x}', \boldsymbol{x}) \ \forall \, \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and for any set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$, the matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi-definite. It can be shown that a kernel induces a *reproducing kernel Hilbert space* (RKHS), a function space equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, containing elements

$$\mathcal{H} = \mathrm{cl} \left(\left\{ f : f_k(\cdot) = \sum_{i=1}^{n} \alpha_{ik} \mathrm{K}(\cdot, \boldsymbol{x}_i) \text{ for } \boldsymbol{\alpha} \in \mathbb{R}^{n \times K} \right\} \right)$$

where cl is the closure operator. Although $\mathcal{H}$ is an infinite dimensional space, it turns out that the minimizer $\hat{f}_{\mathcal{S}}^{\lambda}$ can be obtained in closed form for mean squared loss, given by

$$\hat{f}_{\mathcal{S}}^{\lambda}(\boldsymbol{x}) = \boldsymbol{K}_{\boldsymbol{x}}^T \left(\boldsymbol{K} + \lambda \mathbf{1}_n \right)^{-1} \boldsymbol{Y} \xrightarrow{\lambda \to 0} \hat{f}_{\mathcal{S}}(\boldsymbol{x}) := \boldsymbol{K}_{\boldsymbol{x}}^T \boldsymbol{K}^{\dagger} \boldsymbol{Y}$$

where the regularization functional is induced by the inner product of the RKHS, $\Omega(f) = ||f||_{\mathcal{H}}^2$, $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ has entries $K_{ij} = \mathrm{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{K}_{\boldsymbol{x}} \in \mathbb{R}^n$ has entries $(K_{\boldsymbol{x}})_i = \mathrm{K}(\boldsymbol{x}, \boldsymbol{x}_i)$. $\boldsymbol{A}^{\dagger}$ denotes the pseudo-inverse of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. We will analyze both the regularized ($\hat{f}_{\mathcal{S}}^{\lambda}$) and unregularized model ($\hat{f}_{\mathcal{S}}$) and their associated LOO errors in the following. For a more in-depth discussion of kernels and a theoretical account on RKHS, we refer the reader to Hofmann et al. (2008); Paulsen & Raghupathi (2016).

3.2 Neural Networks and Associated Kernels

Another function class that has seen a rise in popularity is given by the family of neural networks of depth $L \in \mathbb{N}$,

$$\mathcal{F}_{NN} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}^{(L)} \sigma \left(\boldsymbol{W}^{(L-1)} \dots \sigma \left(\boldsymbol{W}^{(1)} \boldsymbol{x} \right) \dots \right) \right\},$$

where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ for $d_l \in \mathbb{N}$, $d_0 = d$, $d_L = K$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is an element-wise non-linearity. We denote by $\boldsymbol{\theta} \in \mathbb{R}^m$ the concatenation of all parameters $\left(\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}\right)$, where $m \in \mathbb{N}$ is the number of parameters of the model. The parameters are initialized randomly according to $W_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_w^2)$ for $\sigma_w > 0$ and then adjusted to optimize the empirical error $\hat{L}_{\mathcal{S}}$, usually through an iterative procedure using a variant of gradient descent. Usually, the entire vector $\boldsymbol{\theta}$ is optimized via gradient descent, leading to a highly non-convex problem which does not admit a closed-form solution $\hat{f}$, in contrast to kernel learning. Next, we will review settings under which the output of a deep neural network can be approximated using a kernel formulation.

Random Features. Instead of training all the parameters $\boldsymbol{\theta}$, we restrict the function space by only optimizing over the top layer $\boldsymbol{W}^{(L)}$ and freezing the other variables $\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L-1)}$ to their value at initialization. This restriction makes the model linear in its parameters and induces a finite-dimensional feature map

$$\phi_{\text{RF}} : \mathbb{R}^d \to \mathbb{R}^{d_{L-1}}, \quad \boldsymbol{x} \mapsto \sigma \left(\boldsymbol{W}^{(L-1)} \dots \sigma \left(\boldsymbol{W}^{(1)} \boldsymbol{x} \right) \dots \right).$$

Originally introduced by Rahimi & Recht (2008) to enable more efficient kernel computations, the random feature model has recently seen a strong spike in popularity and has been the topic of many theoretical works (Jacot et al., 2020; Mei & Montanari, 2021).

Infinite Width. Another, only recently established correspondence emerges in the infinite-width scenario. By choosing $\sigma_w = \frac{1}{\sqrt{d_l}}$ for each layer, Lee et al. (2018) showed that at initialization, the network exhibits a Gaussian process behaviour in the infinite-width limit. More precisely, it holds that $f_k \overset{i.i.d.}{\sim} \mathcal{GP}\left(0, \Sigma^{(L)}\right)$, for $k = 1, \dots, K$, where $\Sigma^{(L)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel, coined NNGP, obeying the recursive relation

$$\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(l-1)}|_{\boldsymbol{x}, \boldsymbol{x'}}\right)} \left[\sigma(z_1) \sigma(z_2) \right]$$

with base $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \frac{1}{\sqrt{d}} \boldsymbol{x}^T \boldsymbol{x'}$ and $\boldsymbol{\Sigma}^{(l-1)}|_{\boldsymbol{x}, \boldsymbol{x'}} \in \mathbb{R}^{2 \times 2}$ is $\Sigma^{(l-1)}$ evaluated on the set $\{\boldsymbol{x}, \boldsymbol{x'}\}$. In order to reason about networks trained with gradient descent, Jacot et al. (2018) extended this framework and proved that the empirical neural tangent kernel converges to a constant, deterministic kernel Θ. The so-called neural tangent kernel $\Theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is also available through a recursive relation involving the NNGP:

$$\Theta^{(L+1)}(\boldsymbol{x}, \boldsymbol{x'}) = \Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x'}) \dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x'}) + \Sigma^{(L+1)}(\boldsymbol{x}, \boldsymbol{x'}),$$

with base $\Theta^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'})$ and $\dot{\Sigma}^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}^{(l-1)}|_{\boldsymbol{x}, \boldsymbol{x'}}\right)} \left[\dot{\sigma}(z_1) \dot{\sigma}(z_2) \right]$. Training infinitely-wide networks with gradient descent thus reduces to kernel learning with the NTK $\Theta^{(L)}$.

3.3 Leave-one-out Error

The fundamental goal of machine learning is to choose a model $f \in \mathcal{F}$ with minimal generalization error $L(f)$. Since the data distribution $\mathcal{D}$ is usually not known, practitioners obtain an estimate for $L(f)$ by using an additional test set $\mathcal{S}_{\text{test}} = \{(\boldsymbol{x}_1^t, \boldsymbol{y}_1^t), \dots, (\boldsymbol{x}_b^t, \boldsymbol{y}_b^t)\}$ not used as input to the learning algorithm $\mathcal{Q}$, and calculate

$$L_{\text{test}} = \frac{1}{b} \sum_{i=1}^{b} L_f(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t), \qquad A_{\text{test}} = \frac{1}{b} \sum_{i=1}^{b} a_f(\boldsymbol{x}_i^t, \boldsymbol{y}_i^t) \tag{1}$$

While this is a simple approach, practitioners cannot always afford to put a reasonably sized test set aside. This observation led the field of statistics to develop the idea of a leave-one-out error (LOO), allowing one to obtain an estimate of $L(f)$ without having to rely on additional data (Lachenbruch, 1967; Mosteller & Tukey, 1968). We give a formal definition below.

Definition 3.1. *Consider a learning algorithm $\mathcal{Q}$ and a training set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. Let $\mathcal{S}_{-i} = \mathcal{S} \setminus \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ be the training set without the i-th data point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and define $\hat{f}^{-i} := \mathcal{Q}(\mathcal{F}, \mathcal{S}_{-i})$, the corresponding model obtained by training on $\mathcal{S}_{-i}$. The leave-one-out loss/accuracy of $\mathcal{Q}$ on $\mathcal{S}$ is then defined as the average loss/accuracy of $\hat{f}^{-i}$ incurred on the left out point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, given by*

$$L_{LOO}(\mathcal{Q}, \mathcal{F}; \mathcal{S}) = \frac{1}{n}\sum_{i=1}^n L_{\hat{f}^{-i}}(\boldsymbol{x}_i, \boldsymbol{y}_i), \qquad A_{LOO}(\mathcal{Q}, \mathcal{F}; \mathcal{S}) = \frac{1}{n}\sum_{i=1}^n a_{\hat{f}^{-i}}(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

Notice that calculating L_{LOO} can indeed be done solely based on the training set. On the other hand, using the training set to estimate $L(f)$ comes at a cost. Computing $\hat{f}^{-i}$ for $i = 1, \ldots, n$ requires fitting the given model class n times. While this might be feasible for smaller models such as random forests, the method clearly reaches its limits for larger models and datasets used in deep learning. For instance, evaluating L_{LOO} for ResNet-152 on *ImageNet* requires fitting a model with 6×10^7 parameters 10^7 times, which clearly becomes extremely inefficient. At first glance, the concept of leave-one-out error therefore seems to be useless in the context of deep learning. However, we will now show that the recently established connection between deep networks and kernel learning gives us an efficient way to compute L_{LOO}.

In the following, we will use the short-cut $\hat{f}_{\mathcal{S}}^{\lambda} := \hat{f}^{\lambda}$. Surprisingly, it turns out that for kernels under mean-squared loss, we can write the corresponding leave-one-out error as a closed-form expression that only relies on a *single model* obtained on the full training set $\boldsymbol{X}$:

Theorem 3.2. *Consider a kernel $\mathrm{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and the associated objective with regularization parameter $\lambda > 0$ under mean squared loss. Define $\boldsymbol{A} = \boldsymbol{K}(\boldsymbol{K} + \lambda \mathbf{1}_n)^{-1}$. Then it holds that*

$$L_{LOO}\left(\mathcal{Q}_{\lambda}^{ERM}\right) = \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \left(\Delta_{ik}^{\lambda}\right)^2, \qquad A_{LOO}\left(\mathcal{Q}_{\lambda}^{ERM}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\left\{(\boldsymbol{y}_i - \Delta_{i\bullet})^* = \boldsymbol{y}_i^*\right\}} \tag{2}$$

where the residual $\Delta_{ik}^{\lambda} \in \mathbb{R}$ for $i = 1, \ldots, n$, $k = 1, \ldots, K$ is given by $\Delta_{ik}^{\lambda} = \frac{Y_{ik} - \hat{f}_k^{\lambda}(\boldsymbol{x}_i)}{1 - A_{ii}}$.

For the remainder of this text, we will use the short-cut $L_{\text{LOO}}^{\lambda} := L_{\text{LOO}}\left(Q_{\lambda}^{\text{ERM}}\right)$. While the expression for L_{LOO}^{λ} for $K = 1$ has been known in the statistics community for decades (Stone, 1974), its multiclass extension and the formula for A_{LOO}^{λ} are to the best of our knowledge novel results. We provide a proof of Theorem 3.2 in the Appendix A.1. Notice that Theorem 3.2 not only allows for an efficient computation of L_{LOO}^{λ} and A_{LOO}^{λ} but also provides us with a simple formula to assess the generalization behaviour of the model. It turns out that L_{LOO}^{λ} captures a great variety of generalization phenomena, typically encountered in deep learning. Leveraging the neural tangent kernel together with Theorem 3.2 allows us to investigate their origin.

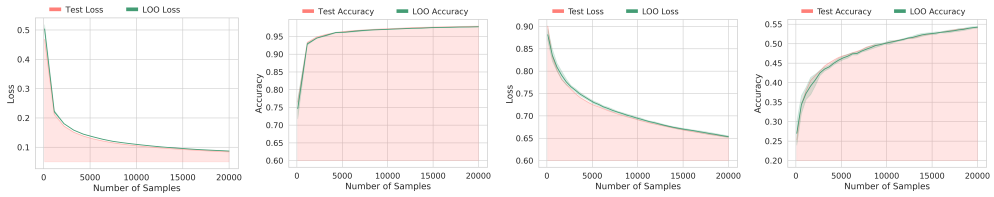
As a first step, we consider the zero regularization limit $\lambda \to 0$, i.e. L_{LOO}^{0} since this is the standard setting considered in deep learning. A quick look at the formula reveals however that care has to be taken; as soon as the kernel is invertible, we have a $\frac{0}{0}$ situation.

Corollary 3.3. *Consider the eigendecomposition $\boldsymbol{K} = \boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\omega})\boldsymbol{V}^T$ for $\boldsymbol{V} \in O(n)$ and $\boldsymbol{\omega} \in \mathbb{R}^n$. Denote its rank by $r = \mathrm{rank}(\boldsymbol{K})$. Then it holds that the residuals $\Delta_{ik}^{\lambda} \in \mathbb{R}$ can be expressed as*

$$\Delta_{ik}^{\lambda}(r) = \sum_{l=1}^n Y_{lk} \frac{\sum_{j=r+1}^n V_{ij}V_{lj} + \sum_{j=1}^r \frac{\lambda}{\lambda+\omega_j} V_{ij}V_{lj}}{\sum_{j=r+1}^n V_{ij}^2 + \sum_{j=1}^r \frac{\lambda}{\lambda+\omega_j} V_{ij}^2}$$

Moreover in the zero regularization limit, i.e. $\lambda \to 0$, it holds that

$$\Delta_{ik}^{\lambda}(r) \to \Delta_{ik}(r) = \begin{cases} \sum_{l=1}^n Y_{lk} \frac{\sum_{j=r+1}^n V_{ij}V_{lj}}{\sum_{j=r+1}^n V_{ij}^2} & \text{if } r < n \\ \sum_{l=1}^n Y_{lk} \frac{\sum_{j=1}^n \frac{1}{\omega_j} V_{ij}V_{lj}}{\sum_{j=1}^n \frac{1}{\omega_j} V_{ij}^2} & \text{if } r = n \end{cases}$$



(a) L_{LOO}, *MNIST* (b) A_{LOO}, *MNIST* (c) L_{LOO}, *CIFAR10* (d) A_{LOO}, *CIFAR10*

Figure 1: Test and LOO losses (a,c) and accuracies (b, d) as a function of sample size n. We use a 5-layer fully-connected NTK model on *MNIST* and *CIFAR10*.

Remark. At first glance for $\lambda = 0$, a distinct phase transition appears to take place when moving from $r = n - 1$ to $r = n$. Notice however that a small eigenvalue ω_n will allow for a smooth interpolation, i.e. $\Delta_{ik}(n) \xrightarrow{\omega_n \to 0} \Delta_{ik}(n-1)$ for $i = 1, \dots, n$, $k = 1, \dots, K$.

A similar result for the special case of linear regression and $r = n$ has been derived in Hastie et al. (2020). We now explore how well these formulas describe the complex behaviour encountered in deep models.

4 Leave-One-Out as a Generalization Proxy

In this section, we study the derived formulas in the context of deep learning, through the lens of the neural tangent kernel and random feature models. Through empirical studies, we investigate LOO for varying sample sizes, different amount of noise in the targets as well as different regimes of overparametrization. We evaluate the models on the benchmark vision datasets *MNIST* (LeCun & Cortes, 2010) and *CIFAR10* (Krizhevsky & Hinton, 2009). To estimate the true generalization error, we employ the test sets as detailed in equation 1. We provide theoretical insights into double descent, establishing the explosion in loss around the interpolation threshold. Finally, we demonstrate the utility of LOO for state-of-the-art networks by predicting their transferability to new tasks. We provide more experiments for varying depth and double descent with random labels in the Appendix C.

4.1 Function of Sample Size

While known to exhibit only a small amount of bias, the LOO error may have a high variance for small sample sizes (Varoquaux, 2018). Motivated by the consistency results obtained for ridge regression in Patil et al. (2021), we study how the LOO error (and its variance) of an infinitely-wide network behaves as a function of the sample size n. In the following we thus interpret LOO as a function of n. In order to quantify how well LOO captures the generalization error for different sample regimes, we evaluate a 5-layer fully-connected NTK model for varying sample sizes on *MNIST* and *CIFAR10*. We display the resulting test and LOO losses in Figure 1. We plot the average result of 5 runs along with confidence intervals. We observe that both L_{LOO} and A_{LOO} closely follow their corresponding test quantity, even for very small sample sizes of around $n = 500$. As we increase the sample size, the variance of the estimator decreases and the quality of both LOO loss and accuracy improves even further, offering a very precise estimate at $n = 20000$.

4.2 Random Labels

Next we study the behaviour of the generalization error when a portion of the labels is randomized. In Zhang et al. (2017), the authors investigated how the performance of deep neural networks is affected when the functional relationship between inputs $\boldsymbol{x}$ and targets $\boldsymbol{y}$ is destroyed by replacing $\boldsymbol{y}$ by a random label. More specifically, given a dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and one-hot targets $\boldsymbol{y} = \boldsymbol{e}_k$ for $k \in \{1, \dots, K\}$, we fix a noise level $p \in [0, 1]$ and replace a subset of size pn of the targets with random ones, i.e. $\boldsymbol{y} = \boldsymbol{e}_{\tilde{k}}$ for $\tilde{k} \sim \mathcal{U}(\{1, \dots, K\})$. We then train the model using the noisy targets $\tilde{\boldsymbol{y}}$ while measuring the generalization error with respect to the clean target distribution. As observed in Zhang et al. (2017); Rolnick et al. (2018), neural networks manage to achieve very small empirical

(a) L_{LOO}, *MNIST* (b) A_{LOO}, *MNIST* (c) L_{LOO}, *CIFAR10* (d) A_{LOO}, *CIFAR10*

Figure 2: Test and LOO losses (a, c) and accuracies (b, d) as a function of noise. We use a 5-layer fully-connected NTK model on *MNIST* and *CIFAR10* with $n = 20000$.

error even for $p = 1$ whereas their generalization consistently worsens with increasing p. Random labels has become a very popular experiment to assess the quality of generalization bounds (Arora et al., 2019b; 2018; Bartlett et al., 2017). We now show how the leave-one-out error exhibits the same behaviour as the test error under label noise, thus serving as an ideal test bed to analyze this phenomenon. In the following, we consider the leave-one-out error as a function of the labels, i.e. $L_{\text{LOO}} = L_{\text{LOO}}(\boldsymbol{y})$. The phenomenon of noisy labels is of particular interest when the model has enough capacity to achieve zero empirical error. For kernels, this is equivalent to $\operatorname{rank}(\boldsymbol{K}) = n$ and we will thus assume in this section that the kernel matrix has full rank. However, a direct application of Theorem 3.2 is not appropriate, since randomizing some of the training labels also randomizes the "implicit" test set. In other words, our aim is not to evaluate $\hat{f}_{\tilde{\mathcal{S}}}^{-i}$ against $\tilde{\boldsymbol{y}}_i$ since $\tilde{\boldsymbol{y}}_i$ might be one of the randomized labels. Instead, we want to compare $\hat{f}_{\tilde{\mathcal{S}}}^{-i}$ with the true label $\boldsymbol{y}_i$, in order to preserve a clean target distribution. To fix this issue, we derive a formula for the leave-one-out error for a model trained on labels $\tilde{\boldsymbol{Y}}$ but evaluated on $\boldsymbol{Y}$:

Proposition 4.1. *Consider a kernel with spectral decomposition* $\boldsymbol{K} = \boldsymbol{V}\operatorname{diag}(\boldsymbol{\omega})\boldsymbol{V}^T$ *for* $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ *orthogonal and* $\boldsymbol{\omega} \in \mathbb{R}^n$. *Assume that* $\operatorname{rank}(\boldsymbol{K}) = n$. *Then it holds that the leave-one-out error* $L_{LOO}(\tilde{\boldsymbol{Y}}; \boldsymbol{Y})$ *for a model trained on* $\tilde{\mathcal{S}} = \{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^n$ *but evaluated on* $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ *is given by*

$$L_{LOO}(\tilde{\boldsymbol{Y}}; \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\tilde{\Delta}_{ik} + Y_{ik} - \tilde{Y}_{ik}\right)^2, \qquad A_{LOO}(\tilde{\boldsymbol{Y}}; \boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\left\{(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{\Delta}}_{i\bullet})^* = \boldsymbol{y}_i^*\right\}}$$

where $\tilde{\Delta}_{ik} = \Delta_{ik}(\tilde{\boldsymbol{Y}}) \in \mathbb{R}$ *is defined as in Corollary 3.3.*

We defer the proof to the Appendix A.4. Attentive readers will notice that we indeed recover Theorem 3.2 when $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$. Using this generalization of LOO, we can study the setup in Zhang et al. (2017) to test whether LOO indeed correctly reflects the behaviour of the generalization error. To this end, we perform a random label experiment on *MNIST* and *CIFAR10* with $n = 20000$ for a NTK model of depth 5, where we track test and LOO losses while increasingly perturbing the targets with noise. We display the results in Figure 2. We see an almost perfect match between test and LOO loss as well test and LOO accuracy, highlighting how precisely LOO reflects the true error across all settings.

4.3 Double Descent

Originally introduced by Belkin et al. (2019), the double descent curve describes the peculiar shape of the generalization error as a function of the model complexity. The error first follows the classical U-shape, where an increase in complexity is beneficial until a specific threshold. Then, increasing the model complexity induces overfitting, which leads to worse performance and eventually a strong spike around the interpolation threshold. However, as found in Belkin et al. (2019), further increasing the complexity again reduces the error, often leading to the overall best performance. A large body of works provide insights into the double descent curve but most of them rely on very restrictive assumptions such as Gaussian input distributions (Harzli et al., 2021; Adlam & Pennington, 2020; d'Ascoli et al., 2020) or teacher models (Bartlett et al., 2020; Harzli

et al., 2021; Adlam & Pennington, 2020; Mei & Montanari, 2021; Hastie et al., 2020) and require very elaborate proof techniques. Here we show under arguably weaker assumptions that the LOO loss also exhibits a spike around the interpolation threshold. Our proof is very intuitive, relying largely on Equation 2 and standard probability tools.

In order to reason about the effect of complexity, we have to consider models with a finite dimensional feature map $\phi_m : \mathbb{R}^d \to \mathbb{R}^m$. Increasing the dimensionality of the feature space $\mathbb{R}^m$ naturally overparametrizes the model and induces a dynamic in the rank $r(m) := \text{rank}(\boldsymbol{K}) = \text{rank}\left(\phi_m(\boldsymbol{X})\right)$. The exact rank dynamics $r(m)$ are obviously specific to the employed feature map ϕ_m, but by construction always monotonic. We restrict our attention to models that admit an interpolation point, i.e. $\exists m^* \in \mathbb{N}$ such that $r(m^*) = n$. While the developed theory holds for any kernel, we focus the empirical evaluations largely to the random feature model due to its aforementioned connection to neural networks. In the following, we will interpret the leave-one-out-error as a function of m and n, i.e. $L_{\text{LOO}} = L^n_{\text{LOO}}(m)$. Notice that $L^n_{\text{LOO}}(m)$ measures the generalization error of a model trained on $n-1$ points, instead of n. The underlying interpolation point thus shifts, i.e. we denote $m^* \in \mathbb{N}$ such that $r(m^*) = n-1$. In the following we will show that $L^n_{\text{LOO}}(m^*(n)) \xrightarrow{n\to\infty} \infty$.

Intuition. We give a high level explanation for our theoretical insights in this paragraph. For $m < m^*$ the dominating term in $L^n_{\text{LOO}}(m)$ is

$$g(r,\lambda) = \sum_{i=1}^{n} \frac{1}{\sum_{l=r+1}^{n} V_{il}^2 + \sum_{l=1}^{r} \frac{\lambda}{\lambda+\omega_l} V_{il}^2}$$

We observe a blow-up for small λ due to the unit vector nature of both $\boldsymbol{V}_{i\bullet}$ and $\boldsymbol{V}_{\bullet j}$. For instance, considering the interpolation point where $r = n-1$ and $\lambda = 0$, we get $g(n-1,0) = \sum_{i=1}^{n} \frac{1}{V_{in}^2}$, the sum over the elements of $\boldsymbol{V}_{\bullet n}$, which likely has a small entry due to its unit norm nature. On the other hand, reducing r or increasing λ has a dampening effect as we combine multiple positive terms together, increasing the likelihood to move away from the singularity. A similar dampening behaviour in $L^n_{\text{LOO}}(m)$ can also be observed for $m > m^*$ when $\omega_n >> 0$.

In order to mathematically prove the spike at the interpolation threshold, we need an assumption on the spectral decomposition of $\boldsymbol{K}$:

Assumption 4.2. *Consider again the eigendecomposition $\boldsymbol{K} = \boldsymbol{V}\,\text{diag}\left(\boldsymbol{\omega}\right)\boldsymbol{V}^T$. We impose the following distributional assumption on the rotation matrix $\boldsymbol{V}$:*

$$\boldsymbol{V} \sim \mathcal{U}\left(O_n\right)$$

where $\mathcal{U}\left(O_n\right)$ is the Haar measure on the orthogonal group O_n.

Under Assumption 4.2, we can prove that at the interpolation point, the leave-one-out error exhibits a spike, which worsens as we increase the amount of data:

Theorem 4.3. *For large enough $n \in \mathbb{N}$ and $\lambda \to 0$, it holds that*

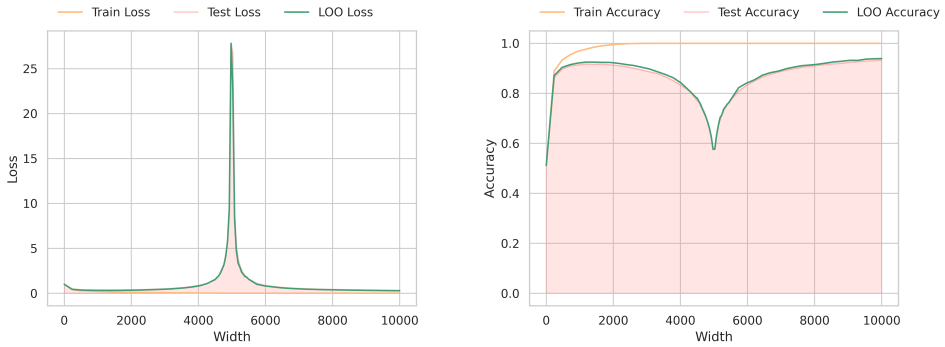
$$L^n_{LOO}(m^*) \gtrapprox 2nA$$

where $A \sim \Gamma(\frac{1}{2}, 1)$ is independent of n. $L^n_{LOO}(m^*)$ hence diverges a.s. with $n \to \infty$.

It is surprising that Assumption 4.2 suffices to prove Theorem 4.3, highlighting the universal nature of double descent. We provide empirical evidence for the double descent behaviour of LOO in Figure 3. We fit a random feature model of varying width to the binarized *MNIST* dataset with $n = 5000$. We observe that the LOO error indeed closely follows the test error, exhibiting the spike as predicted by Theorem 4.3.

4.4 Transfer Learning

Finally, we study more realistic networks often employed for large-scale computer vision tasks. We study the task of transfer learning (Yosinski et al., 2014) by considering classifiers trained on



(a) Loss (b) Accuracy

Figure 3: (a) Train, Test and LOO losses (a) and accuracies (b) as a function of width. We use a random feature model on binary *MNIST* with $n = 5000$.

ImageNet (Krizhevsky et al., 2012) and fine-tuning their top-layer to the *CIFAR10* dataset. Notice that this corresponds to training a kernel with a data-dependent feature map ϕ_{data} induced by the non-output layers. For standard kernels however, the feature map is not dependent on any training data and consistency results such as Patil et al. (2021) indeed rely on this fact. One could thus expect that the data-dependent feature map might lead to a too optimistic LOO error. Our experiments on *ResNet18* (He et al., 2016), *AlexNet* (Krizhevsky et al., 2012), *VGG* (Simonyan & Zisserman, 2015) and *DenseNet* (Huang et al., 2017) however demonstrate that transferring between tasks indeed preserves the stability of the networks and leads to a very precise estimation of the test accuracy. We display the results in Table 1. We consider a small data regime where $n = 10000$, as often encountered in practice for transfer learning tasks. In order to highlight the role of pre-training, we also evaluate random feature maps ϕ_{rand} where we use standard initialization for all the parameters in the network. Indeed we clearly observe the benefits of pre-training, leading to a high increase in performance for all considered models.

MODEL	$A_{\text{TEST}}(\phi_{\text{DATA}})$	$A_{\text{LOO}}(\phi_{\text{DATA}})$	$A_{\text{TEST}}(\phi_{\text{RAND}})$	$A_{\text{LOO}}(\phi_{\text{RAND}})$
RESNET18	67.2 ± 0.1	67.5 ± 0.4	37.7 ± 0.2	37.1 ± 0.1
ALEXNET	57.6 ± 0.2	58.2 ± 0.2	24.2 ± 0.2	23.7 ± 0.2
VGG16	56.6 ± 0.2	56.8 ± 0.5	29.3 ± 0.5	29.2 ± 0.8
DENSENET161	72.5 ± 0.2	71.3 ± 0.3	51.7 ± 0.3	49.9 ± 0.3

Table 1: Test and LOO accuracies in percentage for models pre-trained on *ImageNet* and transferred to *CIFAR10* by re-training the top layer. We use 5 runs, each with a different training set of size $n = 10000$. We compare the pre-trained networks with random networks to illustrate the benefits of transfer learning.

5 DISCUSSION

In this work we investigated the concept of leave-one-out error as a measure for generalization in the context of deep learning. Through the use of NTK, we derive new expressions for the LOO loss and accuracy of deep models in the kernel regime and observe that they capture the generalization behaviour for a variety of settings. Moreover, we mathematically prove that LOO exhibits a spike at the interpolation threshold, needing only a minimal set of theoretical tools. Finally, we demonstrate that LOO accurately predicts the predictive power of deep models when applied to the problem of transfer learning.

Notably, the simple form of LOO might open the door to new types of theoretical analyses to better understand generalization, allowing for a disentanglement of the roles of various factors at play such as overparametrization, noise and architectural design.