Deep Double Descent via Smooth Interpolation

Anonymous authors
Paper under double-blind review

Abstract

Overparameterized deep networks can interpolate noisy data while at the same time showing good generalization performance. Common intuition from polynomial regression suggests that large networks are able to sharply interpolate noisy data without considerably deviating from the ground-truth signal. At present, a precise characterization of this phenomenon for deep networks is missing. In this work, we present an empirical study of input-space smoothness of the loss landscape of deep networks over volumes around cleanly- and noisily-labelled training samples, as we systematically increase the number of model parameters and training epochs. Our findings show that loss sharpness in the input space follows both model- and epoch-wise double descent, with worse peaks observed around noisy labels. While small interpolating models sharply fit both clean and noisy data, large interpolating models express a smooth loss landscape, where noisy targets are predicted over large volumes around training data points, in contrast to existing intuition.

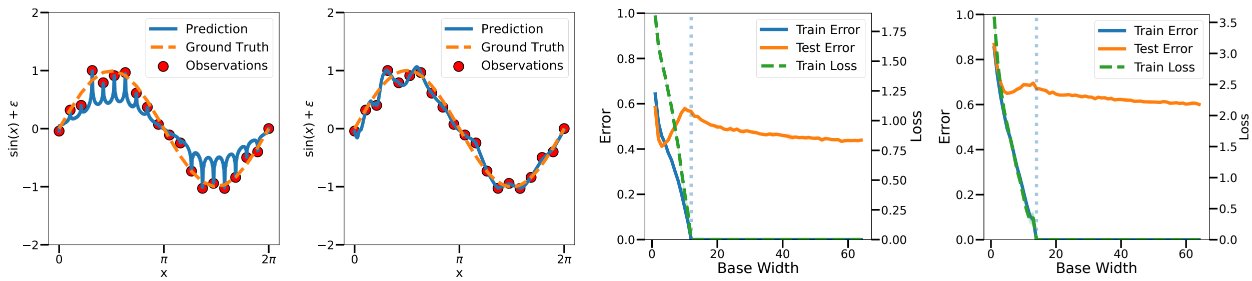
1 Introduction

Recent years have seen increased interest in the study of smoothness of deep networks in relationship to their generalization ability. For networks interpreted as functions of their parameters, smoothness of the loss landscape has been related to improved generalization (Ma & Ying, 2021; Foret et al., 2020; Rosca et al., 2020), increased stability to perturbations (Keskar et al., 2017), reduced minimum description length (Hochreiter & Schmidhuber, 1997), as well as better compression (Chang et al., 2021). Additionally, input-space sensitivity of the networks' learned function has been connected to generalization performance (LeJeune et al., 2019; Novak et al., 2018). Indeed mounting evidence, both empirical (Gamba et al., 2022; Novak et al., 2018) as well as theoretical (Bubeck & Sellke, 2021; Neyshabur et al., 2018), suggests that large state-of-the-art models achieve robust generalization (Ma & Ying, 2021) via smoothness of the learned function. While overparameterization alone is not enough to guarantee strong robustness (Chen et al., 2021; Rice et al., 2020), the large number of parameters of modern networks is thought to promote some form of implicit regularization (Gamba et al., 2022; Bubeck & Sellke, 2021; Neyshabur et al., 2018; 2015).

The ability of overparameterized networks to interpolate training data (Zhang et al., 2018) is related to the double descent phenomenon (Belkin et al., 2019; Nakkiran et al., 2019b; Belkin et al., 2018) – most markedly in the presence of noisy labels – with large networks showing good generalization performance well past the *interpolation threshold* (Belkin et al., 2019), namely the smallest model size achieving zero training error.

In this work, we study deep double descent (Nakkiran et al., 2019b) through the lens of smooth interpolation of the training data, as the model size as well as the number of training epochs vary.

Current intuition from linear and polynomial regression suggests that, under some hypothesis on the training sample, large overparameterized models are able to perfectly interpolate both cleanly- and noisily-labeled samples, without considerably deviating from the ground-truth signal (Muthukumar et al., 2020; Bartlett et al., 2020; Nakkiran et al., 2019a). Figure 1a illustrates this phenomenon, showing a polynomial of large degree that perfectly interpolates the training data, with predictive function sharply interpolating noisy samples (intuitively corresponding to a spike at each training point), while overall remaining close to the data-generating function. In this work, we show that such intuition does not hold for deep networks, which instead smoothly interpolate both clean and noisy data (Figure 1b).



(a) (b) (c) (d)

Figure 1: **Intuition from overparameterized regression.** a) Polynomial of large degree, trained with gradient descent to fit noisy scalar data, reproducing the polynomial regression experiment of Nakkiran et al. (2019a), and reflecting common intuition on double descent, suggesting that the generalization ability of large interpolating models is tied to sharply fitting of noisy data, thus resulting in models that do not deviate considerably from the ground truth signal. b) In this work we show that, contrary to intuition, deep networks *smoothly* interpolate both clean and noisy data, and that improved generalization in the interpolating regime is tied to smoothness of the loss w.r.t. the input variable. c) Double descent curve for the test error on CIFAR-10 with 20% noisy labels for the family of ConvNets considered in this study. d) Double descent for the test error on CIFAR-100.

To quantify this notion for deep networks trained in practice, we conduct an empirical exploration of smoothness of the loss landscape w.r.t. the input variable, in relationship to the double descent curve of the test error. We provide explicit measures of smoothness of the loss, and study deep networks trained on image classification.

Due to the inherently noisy nature of Euclidean estimators in pixel space, and following the *manifold hypothesis* Pope et al. (2020); Bengio (2013); Narayanan & Mitter (2010), postulating that natural data lies on a combination of manifolds of lower dimension than the input data's ambient dimension, we constrain our measures to the support of the data distribution, locally to each training point.

Our experiments show that smooth interpolation – emerging both for large overparameterization and prolonged training – results in large models confidently predicting the (noisy) training targets over large volumes around each training point.

Contributions

- We present the first systematic empirical study of smoothness of the loss landscape of deep networks in relation to overparameterization and interpolation for natural image datasets.
- Starting from infinitesimal smoothness measures from prior work, we introduce volumetric measures that capture loss smoothness when traveling away from training points.
- We develop a geodesic Monte Carlo integration method for constraining our measures to a local approximation of the data manifold, in proximity of each training point.
- We present an empirical study of model-wise and epoch-wise double descent for neural networks trained without confounders (explicit regularization, data augmentation, batch normalization), as well as for commonly-found training settings. By decoupling smoothness from generalization, we empirically show that overparameterization promotes input-space smoothness of the loss landscape. We produce practical examples in which smoothness of the learned function of deep networks does not result in improved generalization, highlighting that the implicit regularization effect of overparameterization should be studied in terms of reduced variation of the learned function.

2 Related work

Our work presents an empirical study of deep double descent in relationship to smoothness of the loss landscape, with respect to the input variable. Our methodology builds upon input-space sensitivity analyses for neural networks, presenting a first systematic study of the role of overparameterization in promoting smoothness of the network's learned function. The smoothness measures presented in section 3, are inspired by the vast body of work on the loss landscape of neural networks in parameter space. Due to the extensive theoretical literature on double descent in simplified controlled settings such as linear regression (Muthukumar et al., 2020; Bartlett et al., 2020; Belkin et al., 2018), in the following we mainly draw connections to prior work targeting deep networks.

Deep Double Descent The classical bias-variance interpretation states that small low-complexity models are characterized by high bias and low variance, while increasing model size reduces bias at the cost of increased variance. After a threshold of optimal bias-variance tradeoff, model variance grows too large, causing the test performance to degrade, showing the characteristic U-shaped test error curve Geman et al. (1992). However, by increasing model size even further, a second descent in the test error is observed (Belkin et al., 2019). This phenomenon was first observed for several machine learning algorithms with increasing model size (model-wise), but Nakkiran et al. (2019b) showed a similar trend during training of deep networks (epoch-wise), as well as for increasing dataset size (sample-wise).

Double descent has been studied from various perspectives: bias-variance decomposition (Yang et al., 2020; Neal et al., 2018), parameter norms (Belkin et al., 2019), and decision boundaries (Somepalli et al., 2022). In this work, we study model-wise and epoch-wise double descent in terms of smoothness of the loss landscape with respect to the input, and separate the analysis in terms of clean and noisily-labeled data points. The most related work to ours is the concurrent one of Somepalli et al. (2022) that studies decision boundaries in terms of reproducibility and double descent. Our works differ in that we study double descent in the loss landscape. Furthermore, our study is more detailed in terms of investigating both model-wise and epoch-wise trends, and takes a closer look at the impact of clean and noisily-labeled data points. Finally, we study the phenomenon without explicit regularization (batch norm, data augmentation) and a simpler optimization procedure (SGD with constant learning rate instead of Adam) to reduce confounding factors.

Loss Landscape of Neural Networks To understand the remarkable generalization ability of deep networks (Xie et al., 2020; Geiger et al., 2019; Keskar et al., 2017), as well as to design better training criteria (Foret et al., 2020), several works study the loss landscape of deep networks in *parameter space*, focusing on solutions obtained by SGD (Kuditipudi et al., 2019), as well as the optimization process (Arora et al., 2022; Li et al., 2021). Inspired by such literature, we quantify smoothness of the loss landscape by estimating the *sharpness* of the loss, as proposed by Foret et al. (2020) and Keskar et al. (2017) for the parameter-space, but we perform our analysis in *input-space*. Importantly, in this work we focus on image classification tasks, and study smoothness of interpolation of training data points.

Input Space Sensitivity and Smoothness Novak et al. (2018) present an empirical sensitivity study of fully-connected networks with piece-wise linear activation functions through the input-output Jacobian norm, which is shown to strongly correlate to the generalization ability of the networks considered. Their study proposes an infinitesimal analysis of the Jacobian norm at training and validation points, as well as the use of input-space trajectories in proximity of the data manifold to probe trained networks. LeJeune et al. (2019) analyse second-order information (the tangent Hessian of a neural network) by using weak data augmentation to constrain their measure to the proximity of the data manifold. Lastly, Gamba et al. (2022) introduce a nonlinearity measure for neural networks with piece-wise linear activations, that strongly correlates with the test error in the second descent for large overparameterized models. Similar to the first two works, we study smoothness of neural networks, using the Jacobian and Hessian norm of neural networks trained in practice, and similar to the latter work, we provide a systematic study of model-wise double descent, which we further extend to epoch-wise trends. Importantly, our study focuses on the loss landscape and average-case robustness, rather than the network's learned function.

Finally, Rosca et al. (2020) investigate how to encourage smoothness, either on the entire input space or around training data points. Interestingly, they postulate a connection between model-wise double descent and smoothness: during the first ascent the model size is large enough to fit the training data at the expense of smoothness, while the second descent happens as the model size becomes large enough for smoothness to increase. Later, Bubeck & Sellke (2021) theoretically prove a universal law of robustness highlighting a trade-off between the model size (number of parameters) and smoothness (Lipschitz constant) of a learning algorithm w.r.t. its input variable. Our work provides empirical evidence supporting the postulate of Rosca et al. (2020) and the law of robustness of Bubeck & Sellke (2021).

3 Methodology

Our leading research question is to quantify smoothness of interpolation of training data for deep networks trained on classification tasks, as the number of model parameters is increased. We interpret a network as a function with input variable $\mathbf{x} \in \mathbb{R}^d$ and learnable parameter $\boldsymbol{\theta}$, incorporating all weights and biases. Our study focuses on the landscape of the loss $\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}, y) := \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y)$ treated as a function of the input $\mathbf{x}$, with target y. Inspired by the literature on the loss landscape of neural networks in parameter space (Foret et al., 2020; Keskar et al., 2017), we quantify (the lack of) smoothness by devising explicit measures of loss sharpness in a neighbourhood of training points $(\mathbf{x}_n, y_n)$, for $n = 1, \ldots, N$. Crucially, for any given network, we focus on sharpness w.r.t. the input variable $\mathbf{x}$, keeping the parameter $\boldsymbol{\theta}$ fixed.

We begin by describing infinitesimal sharpness in section 3.1, which we compute in proximity of the data manifold local to each training point in section 3.2. Finally, we introduce a method for estimating sharpness over data-driven volumes by exploiting data augmentation in section 3.3, and in section 3.4 we detail the chosen data augmentation strategies. The proposed methodology enables us to measure sharpness of interpolation of the training data, by restricting our study near the support of the data distribution.

3.1 Sharpness at Data Points

To estimate how sharply the loss changes w.r.t. infinitesimal perturbations of the input variable $\mathbf{x}$, we study the Jacobian of the loss,

$$\mathbf{J}(\mathbf{x}, y) := \frac{\partial}{\partial \mathbf{x}} \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}, y) \tag{1}$$

To measure sharpness at a point $(\mathbf{x}_n, y_n)$, we follow Novak et al. (2018), and compute the Frobenius norm of $\mathbf{J}(\mathbf{x}_n, y_n)$, which we take in expectation over the training set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$,

$$J = \mathbb{E}_{\mathcal{D}} \|\mathbf{J}(\mathbf{x}, y)\|_F \tag{2}$$

assuming that the loss is differentiable one time at the points considered. Intuitively, sharpness is measured by how fast the loss $\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}, y)$ changes in infinitesimal neighbourhoods of the training data, and a network is said to smoothly interpolate a data point $\mathbf{x}_n$ if the loss is approximately flat locally around the point and the point is classified correctly according to the corresponding target y_n. Throughout our experiments, the Jacobian $\mathbf{J}$ is computed using a backward pass w.r.t. the input variable $\mathbf{x}$.

Equation 2 provides first-order information about the loss landscape. To gain knowledge about curvature, we also study the Hessian of the loss w.r.t. the input variable,

$$\mathbf{H}(\mathbf{x}, y) := \frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^T} \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{x}, y) \tag{3}$$

whose Frobenius norm again we take in expectation over the training set

$$H = \mathbb{E}_{\mathcal{D}} \|\mathbf{H}(\mathbf{x}, y)\|_F \tag{4}$$

The Hessian tensor in equation 3 depends quadratically on the input space dimensionality d, providing a noisy Euclidean estimator of loss curvature in proximity of the input data. Following the *manifold hypothesis* (Bengio, 2013; Narayanan & Mitter, 2010), stating that natural data lies on subspaces of dimensionality lower than the ambient dimension d, we restrict Hessian computation to the tangent subspace of each training point $\mathbf{x}_n$. Starting from equation 1, throughout our experiments, equation 3 is estimated by computing the tangent Hessian, as outlined in the next section.

3.2 Tangent Hessian Estimation

To constrain equation 3 to the support of the data distribution and reduce computational complexity, we adapt the method proposed by LeJeune et al. (2019) and estimate the Hessian norm of the loss projected onto the data manifold local to each training point.

For any input data point $(\mathbf{x}_n, y_n)$ and corresponding Jacobian $\mathbf{J}(\mathbf{x}_n, y_n)$, we generate M augmented data points $\mathbf{x}_n + \mathbf{u}_m$ by randomly sampling a displacement vector $\mathbf{u}_m$ using weak data augmentation. For each sampled $\mathbf{u}_m$, we then estimate the Hessian $\mathbf{H}(\mathbf{x}_n, y_n)$ projected along the direction $\mathbf{x}_n + \mathbf{u}_m$, by computing the finite difference $\frac{1}{\delta}\mathbf{J}(\mathbf{x}_n, y_n) - \mathbf{J}(\mathbf{x}_n + \delta\mathbf{u}_m, y_n)$. Then, following Donoho & Grimes (2003) we estimate the Hessian norm directly by computing

$$H = \frac{1}{M^2\delta^2}\mathbb{E}_{\mathcal{D}}\Big(\sum_{m=1}^{M} \|\mathbf{J}(\mathbf{x}_n, y_n) - \mathbf{J}(\mathbf{x}_n + \delta\mathbf{u}_m, y_n)\|_F^2\Big)^{\frac{1}{2}} \tag{5}$$

which is equivalent to a rescaled version of the rugosity measure of LeJeune et al. (2019). Importantly, different from rugosity, augmentations $\mathbf{x}_n + \mathbf{u}_m$ are generated using weak colour transformations in place of affine transformations (1-pixel shifts), since weak photometric transformations are guaranteed to be fully on-manifold. Details about the specific colour transformations are presented in appendix C.

3.3 Sharpness over Data-Driven Volumes

The measures introduced in equations 2 and 5, capture local sharpness over infinitesimal neighbourhoods of input data points. To study how different networks fit the training data, we devise a method for estimating loss sharpness over volumes centered at each training point $\mathbf{x}_n$, as one moves away from the point. Essentially, we exploit a variant of Monte Carlo (MC) integration to capture sharpness over data-driven volumes, by applying two steps. First, we integrate the Jacobian and Hessian norms along geodesic paths $\boldsymbol{\pi}_p \subset \mathbb{R}^d$ based at $\mathbf{x}_n$, on the data manifold local to each training point, for $p = 1, \ldots, P$. Second, we compute a MC estimation of sharpness over the volume covered by the P paths. The following details each step.

Sharpness along geodesic paths For each training point $(\mathbf{x}_n, y_n) \in \mathcal{D}$, we aim to estimate loss sharpness as we move away from $\mathbf{x}_n$, while traveling on the support of the data distribution. To do so, we exploit a sequence of weak data augmentations of increasing strength to generate paths $\boldsymbol{\pi}_p \subset \mathbb{R}^d$ in the input space, formed by connecting augmentations of $\mathbf{x}_n$.

Formally, let $\mathcal{T}_{\mathbf{s}} : \mathbb{R}^d \to \mathbb{R}^d$, represent a family of smooth transformations (data augmentation) acting on the input space and governed by parameter $\mathbf{s}$, controlling the strength $S = \|\mathbf{s}\|_F$ as well as the direction of the augmentation in $\mathbb{R}^d$. In general, the parameter $\mathbf{s}$, interpreted as a suitably distributed random variable, models the randomness of the transformation. Randomly sampling $\mathbf{s}$, yields a value $\mathbf{s}^{p,k}$ corresponding to a fixed transformation $\mathcal{T}_{\mathbf{s}^{p,k}}$ of strength S^k. For instance, for affine translations, $\mathbf{s}^{p,k}$ models a random radial direction sampled from a hypersphere centered at $\mathbf{x}_n$, with strength S^k denoting the magnitude of the translation (e.g. 4-pixel shift). For photometric transformations, $\mathbf{s}^{p,k}$ may model the change in brightness, contrast, hue, and saturation, with total strength S^k.

To generate on-manifold paths $\boldsymbol{\pi}_p$ starting from $\mathbf{x}_n$, we proceed as follows. First, we fix a sequence of $K+1$ strengths $S^0 < S^1 < \ldots < S^K$, with $S^0 = 0$ denoting the identity transformation $\forall\ p$. Then, for each strength S^k, with $k \geq 1$, p random directions $\mathbf{s}^{p,k}$ are sampled, each with respective fixed magnitude $\|\mathbf{s}^{p,k}\|_F = S^k$. This yields P sequences of transformations $\{\mathcal{T}_{\mathbf{s}^{p,k}}\}_{k=0}^{K}$, each producing augmented versions $\mathbf{x}_n^{p,k}$ of $\mathbf{x}_n$, ordered by strength, $\mathbf{x}_n^{p,1} \prec \ldots \prec \mathbf{x}_n^{p,K}$, and forming a path $\boldsymbol{\pi}_p \subset \mathbb{R}^d$. Specifically, each path $\boldsymbol{\pi}_p$ approximates an on-manifold trajectory by a sequence of Euclidean segments $\mathbf{x}_n^{p,k+1}\mathbf{x}_n^{p,k}$, for $k = 0, \ldots, K$. The maximum augmentation strength S^K controls the distance traveled from $\mathbf{x}_n$, while the number K of strengths used controls how fine-grained the Euclidean approximation is.

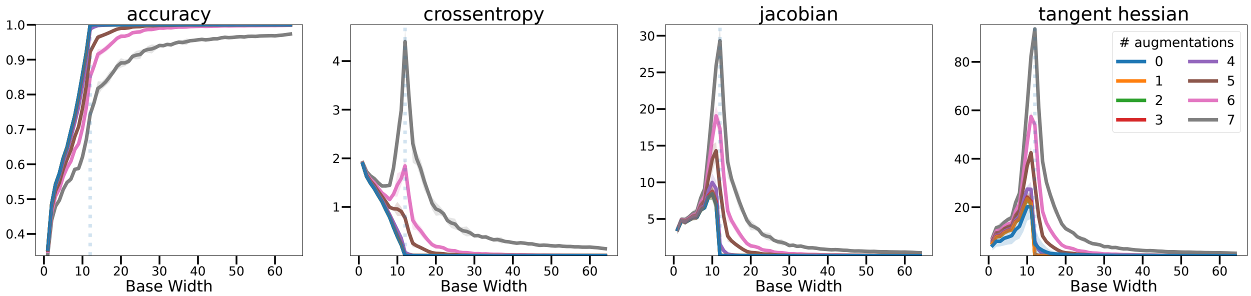


Figure 2: Average metrics integrated over volumes of increasing size (number of data augmentations per path) for a family of ConvNets of increasing base width trained on CIFAR-10 with 20% noisy labels. From left to right: average training accuracy, training loss, Jacobian norm and Hessian norm, each plotted against model size. The dotted vertical line marks the model width that achieves zero train error (i.e. the *interpolation threshold*). All models are trained for 4k epochs. We observe how accuracy over volumes increases monotonically with model size, while crossentropy follows double descent. Combined, the two observations suggest that large networks *confidently* predict the training targets over increasingly large volumes around the training data (for increasing model size). Intringuingly, interpolation is *sharp* at the interpolation threshold, even infinitesimally at each training point (blue curves), while increasing overparameterization produces *smooth* interpolation, contrary to existing intuition. Shaded areas depict standard deviations over 3 seeds.

Volume integration Once a sequence of paths $\{\boldsymbol{\pi}_p\}_{p=1}^P$ is generated for $\mathbf{x}_n$, volume-based sharpness is computed by integrating over all P paths, and normalizing the measure by the length $\mathrm{len}(\boldsymbol{\pi}_p)$ of each path:

$$\frac{1}{P}\sum_{p=1}^{P}\frac{1}{\mathrm{len}(\boldsymbol{\pi}_p)}\int_{\boldsymbol{\pi}_p}\boldsymbol{\sigma}(\mathbf{x}, y_n)d\mathbf{x} \tag{6}$$

where $\boldsymbol{\sigma}$ represents an infinitesimal sharpness measure, namely the Jacobian and tangent Hessian norms at $(\mathbf{x}_n, y_n)$.

Finally, volume-based sharpness is obtained by averaging over the training set $\mathcal{D}$

$$\frac{1}{P}\mathbb{E}_{\mathcal{D}}\sum_{p=1}^{P}\frac{1}{\mathrm{len}(\boldsymbol{\pi}_p)}\int_{\boldsymbol{\pi}_p}\boldsymbol{\sigma}(\mathbf{x}, y_n)d\mathbf{x} = \frac{1}{NP}\sum_{n=1}^{N}\sum_{p=1}^{P}\frac{1}{\mathrm{len}(\boldsymbol{\pi}_p)}\int_{\boldsymbol{\pi}_p}\boldsymbol{\sigma}(\mathbf{x}, y_n)d\mathbf{x} \tag{7}$$

Importantly, extending LeJeune et al. (2019), we replace Euclidean integration by geodesic integration over a local approximation of the data manifold, by generating augmentations of increasing strength.

Crucially, the proposed MC integration captures average-case sharpness in proximity of the training data and is directly related to the generalization ability of the studied networks, as opposed to worst-case sensitivity, as typically considered in adversarial settings (Moosavi-Dezfooli et al., 2019). In fact, the random sampling performed in equation 7 is unlikely to hit adversarial directions, which are commonly identified by searching the input space through an optimization process (Goodfellow et al., 2014; Szegedy et al., 2013).

To conclude our methodology, in section 3.4 we present the family of transformations $\mathcal{T}_{\mathbf{s}}$ used for generating trajectories $\boldsymbol{\pi}_p$ throughout our experiments.

3.4 Weak Data Augmentation Strategies

Computing sharpness of interpolation via equation 6 for each data point $\mathbf{x}_n$ requires generating P trajectories $\boldsymbol{\pi}_p$ composed of augmentations of $\mathbf{x}_n$ of controlled increasing strength. Furthermore, the augmented data points $\{\mathbf{x}_n^{p,k}\}_{k=0}^K$ should lie in proximity of the base point $\mathbf{x}_n$ in order for the Euclidean approximation to be meaningful. Finally, to correctly estimate correlation between smoothness and the generalization ability of the networks considered, volume-based sharpness should not rely on validation data points, i.e. the augmentations $\mathbf{x}_n^{p,k}$ should be strongly correlated to $\mathbf{x}_n$, for each p, k.

To satisfy the above, we modify a weak data augmentation algorithm introduced by Yu et al. (2018), which allows to efficiently generate augmentations that lie in close proximity to the base training point $\mathbf{x}_n$, for image data. Specifically, each base image $\mathbf{x}_n$, consisting of C input channels (e.g. $C = 3$ for RGB images) and $h \times w$ spatial dimensions, is interpreted as C independent matrices $\mathbf{x}_n[c,:,:] \in \mathbb{R}^{h\times w}$, each factorized using Singular Value Decomposition (SVD), yielding a decomposition $\mathbf{x}_n[c,:,:] = U^c\Sigma^c V^{cT}$, where Σ^c is a diagonal matrix whose entries are the singular values of $\mathbf{x}_n[c,:,:]$ sorted by decreasing magnitude. In the original method, Yu et al. (2018) produce weak augmentations by randomly erasing one singular value from the smallest ones, thereby obtaining a modified matrix $\tilde{\Sigma}^c$, and then reconstructing each channel of the base sample via $U^c\tilde{\Sigma}^c V^{cT}$. In this work, in order to generate P random augmentations of strength k, $\tilde{\Sigma}^c$ is obtained by erasing k singular values $\Sigma^c_{i,i}$, for $i = w-k-p+1, \dots, w-p$, and $p = 0, \dots, P-1$[1]. Essentially, the augmentation strength is given by the number k of singular values erased, and P augmentations of similar strength are generated by erasing P subsets of size k from the smallest singular values, for each channel c.

We note that this method produces augmented images that are highly correlated with the corresponding base training sample, and as such they do not directly amount to producing validation data points. We refer the reader to appendix D for further details.

4 Experiments

In this section, we present our empirical exploration of input-space smoothness of the loss landscape of deep networks as the model size and number of training epochs vary. Focusing on *implicit regularization* (Neyshabur et al., 2015) promoted by optimization and model architecture, we evaluate our sharpness measures on a series of networks of increasing number of parameters, trained without any form of explicit regularization (e.g. weight decay, batch normalization, dropout), and later extend our analysis to common training settings in section 4.4.

Experimental setup We reproduce deep double descent by following the experimental setup of Nakkiran et al. (2019b). Specifically, we train a family of ConvNets formed by 4 convolutional stages of controlled base width $[w, 2w, 4w, 8w]$, for $w = 1, \dots, 64$, on the CIFAR-10 dataset with 20% noisy training labels and on CIFAR-100. All models are trained for 4k epochs using SGD with momentum 0.9 and fixed learning rate. Following Arpit et al. (2019), to stabilize prolonged training, we use a learning rate warmup schedule. Furthermore, we extend our empirical results to training settings more commonly found in practice, and validate our main findings on a series of ResNet18s (He et al., 2015) of increasing base width $w = 1, \dots, 64$, with batch normalization, trained with the Adam optimizer for 4k epochs using data augmentation. We refer the reader to section B for a full description of our experimental setting.

We begin our experiments by reproducing double descent for the test error for the ConvNets (Figure 1c). Starting with small models and by increasing model size, a U-shaped curve is observed whereupon small models underfit the training data, as indicated by high train and test error. As model size increases, the optimal bias/variance trade-off is reached (Geman et al., 1992). Mid-sized models increasingly overfit training data – as shown by increasing test error for decreasing train error and loss – until zero training error is achieved, and the training data is interpolated. The smallest interpolating model size is typically referred to as *interpolation threshold* (Belkin et al., 2019). Near said threshold, the test error peaks. Finally, large overparameterized models achieve improved generalization, as marked by decreasing test error, while still interpolating the training set.

4.1 Loss Landscape Smoothness Follows Double Descent

In this section, we establish a strong correlation between double descent of the test error and smooth interpolation of noisy training data. Figure 2 studies fitting of training data for models at convergence (training for 4k epochs) as model size increases. Starting with (infinitesimal) sharpness at training points (blue curve), we observe that training accuracy at convergence monotonically increases with model size, with 100% accuracy reached at the interpolation threshold and maintained therefrom. At the same time, crossentropy loss

[1] Assuming square spatial dimensions $h = w$.

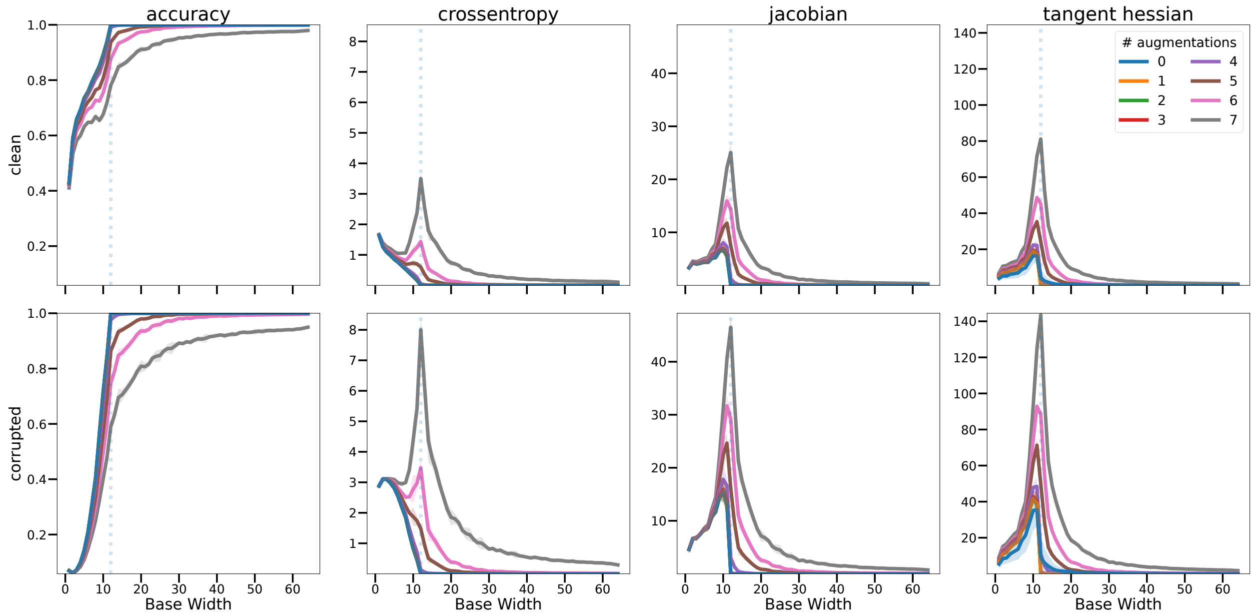


Figure 3: Average accuracy, crossentropy, Jacobian and Hessian norms integrated over volumes of increasing size (augmentations per path) around clean (top) and noisy (bottom) subsets of the CIFAR-10 training set with 20% noisy labels. For models near the interpolation threshold, we observe a large increase in the loss for increasing neighborhood size. At the interpolation threshold, sharp interpolation is observed for both clean and noisy samples, with crossentropy, sensitivity (Jacobian norm) and curvature peaking over all volumes considered. Larger models present a smoother loss landscape around training points, with the largest models expressing a locally flat landscape around each point. This finding shows that large networks are confidently and smoothly *predicting the noisy labels* around data points whose label was corrupted, suggesting that smoothness emerging from overparameterization in fact hinders generalization locally to those points.

undergoes double descent, showing a peak near the interpolation threshold, and then decreasing as model size grows. Similarly, the Jacobian and Hessian norms peak at the interpolation threshold and then rapidly decrease, showing that all training points become stationary for the loss, and that the landscape becomes flatter as model size grows past the interpolation threshold. When all measures are integrated over volumes of increasing size (number of augmentations per path), we observe how large overparameterized models are able to smoothly fit the training data over large volumes. This finding suggests that – in contrast to the polynomial intuition of Figure 1a) – overparameterized networks interpolate training data *smoothly* (as intuitively depicted in Figure 1b).

Our finding extends the observations of Novak et al. (2018) and LeJeune et al. (2019) from fixed-size networks to a spectrum of model sizes, and establishes a clear correlation with the test error peak in double descent. Finally, the results substantiate the universal law of robustness (Bubeck & Sellke, 2021), showing that at the interpolation threshold highest sensitivity to input perturbations is observed, while overparameterization beyond the threshold promotes smoothness. Intriguingly, the peak in input-space sensitivity is observed for much smaller models than what predicted by Bubeck & Sellke (2021). In the following section, we study this behaviour in proximity of cleanly- and noisily-labeled training samples. We refer the reader to section 4.4 for analogous results on ResNets trained with Adam.

4.2 Smooth Interpolation of Noisy Labels

In this section, we break down the noisily labeled training set into two subsets: cleanly-labeled points, and training points with corrupted labels, and explore how fitting is affected by the training labels.

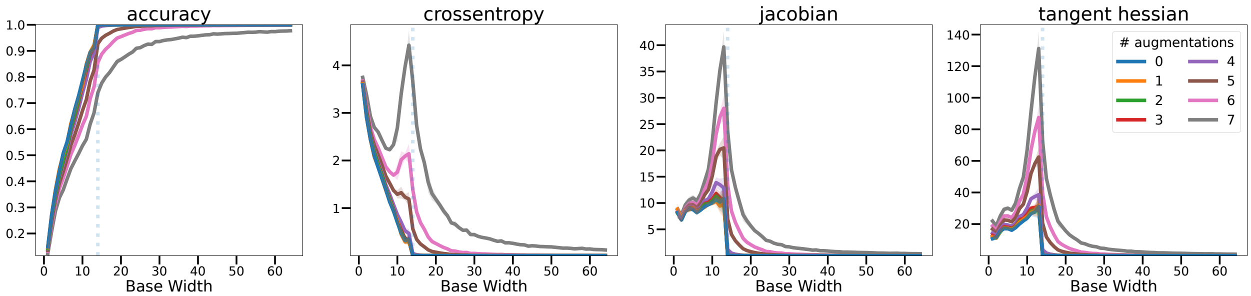


Figure 4: Average metrics integrated over volumes of increasing size (number of data augmentations per path) for a family of ConvNets of increasing base width trained on CIFAR-100. From left to right: average training accuracy, training loss, Jacobian norm and Hessian norm, each plotted against model size. All models are trained for 4k epochs. For relatively complex datasets (i.e. with few samples per class), our findings hold even without artificially corrupted labels, suggesting that the trends reported in this work are not caused by synthetic noise. Shaded areas depict standard deviations over 3 seeds.

Figure 3 reports accuracy, crossentropy, as well as sharpness measures computed on the clean subset of CIFAR-10 (top), as well as the corrupted subset (bottom). We begin by noting that small models fit mostly the cleanly labeled data points, and show close to zero accuracy on the noisily labeled data points, showing a bias towards learning simple patterns. We hypothesize that most cleanly labeled samples act as "simple examples", while noisily labeled ones provide "hard examples", akin to support vectors. This behaviour is aligned with prior observations, reporting that deep networks admit support vectors (Toneva et al., 2018) and that deep networks are biased towards learning training samples in approximately the same order (Hacohen et al., 2020).

As model size grows toward the interpolation threshold, networks fit both clean and noisy samples (as marked by increasing accuracy on both subsets), with large models consistently predicting the clean and noisy labels over large volumes. At the same time, crossentropy local to each training point (blue curve) approaches zero past the interpolation threshold, while volume-based crossentropy undergoes double descent. Interestingly, this trend is observed both around cleanly- and noisily-labeled training samples, with peaks at the interpolation threshold which are considerably more marked for noisy labels.

Our sharpness measures follow double descent for all volumes considered, even when no Monte Carlo integration is performed (blue curve). Importantly, curvature as measured by the Hessian norm rapidly decreases as model size grows, showing that large networks smoothly interpolate both clean and noisy samples. Importantly, we observe how the second descent in test error corresponds to improved fitting of cleanly-labeled samples, while the network lose their generalization ability locally to noisy labeled points.

In Figure 1d we extend the observations to CIFAR-100, where model-wise double descent is observed even on the standard dataset without artificially corrupted labels. Similarly to what observed for the noisy version of CIFAR-10, Figure 4 shows the loss landscape peaking in sharpness at the interpolation threshold, and then rapidly decreasing as model size grows, with large networks smoothly fitting the training set over increasingly large volumes. This finding suggests that double descent is tied to dataset complexity, and that the trends reported in this work are not caused by artificially corrupted labels.

4.3 Epochwise Double-Descent

We now turn our attention to epoch-wise double descent, first reported for the test error of deep networks by (Nakkiran et al., 2019b). Figure 5 shows the test error (left), train crossentropy (middle), as well as Jacobian norm (right) for ConvNets trained on CIFAR-10 with 20% noisy labels. We consolidate our observations for each metric with heatmaps, in which the y-axis represents training epochs, and the x-axis denotes the models' base width. We observe that models past the interpolation threshold (base width $w = 15$) undergo epoch-wise double descent for each metric. At the same time, models with base width $w < 15$ are unable to reduce their test error within 4k training epochs, and this is associated to non-decreasing training

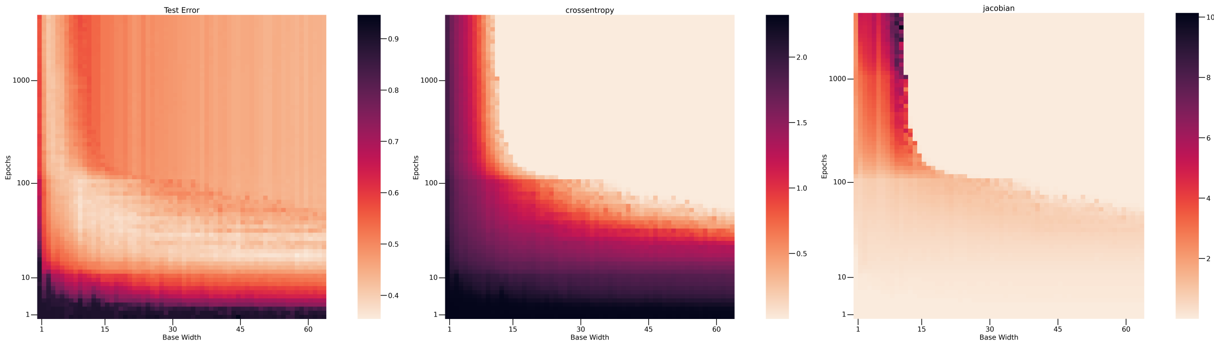
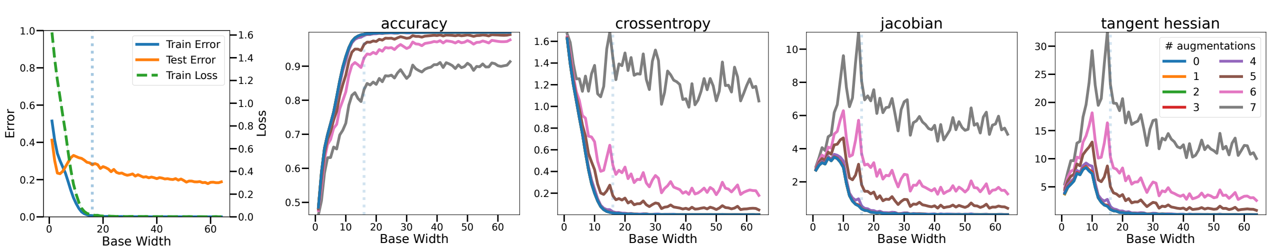


Figure 5: (Left) Test error (Middle) Train crossentropy (Right) Jacobian norm for ConvNets trained on CIFAR-10 with 20% noisy labels. The heatmaps show each metric for increasing training epochs (y-axis) and base width (x-axis). Models past the interpolation threshold (base width $w = 15$) undergo epoch-wise double descent for each metric. Similar trends are observed for curvature, as measured by the Hessian norm. All metrics are computed on the training set only, without geodesic Monte Carlo integration.

loss as well as Jacobian norm. We hypothesize that the model size affects a model's ability to interpolate the training data, and therefore affects the training dynamics and the occurrence of epoch-wise double descent.

4.4 Practical Training Settings

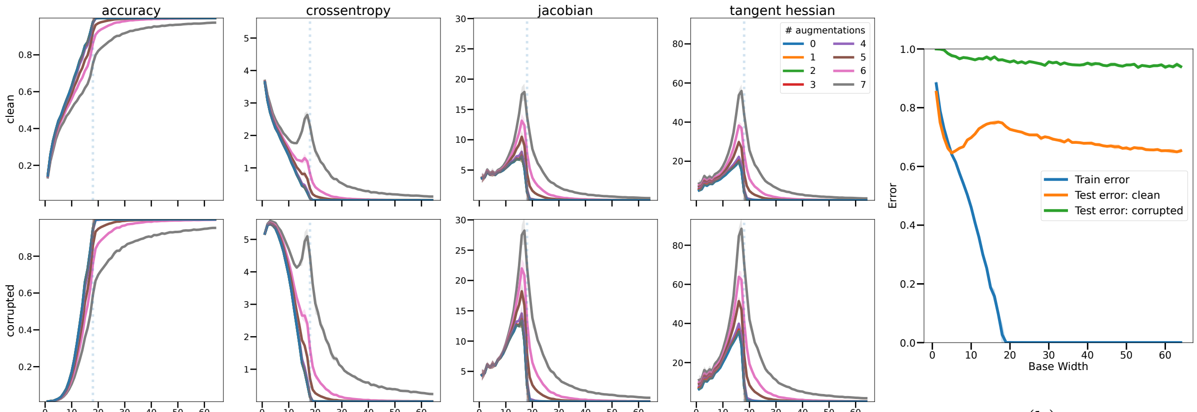
The training set up considered so far was designed to include the least amount of confounders (e.g. adaptive learning rates, explicit regularization, skip connections, normalization layers) and focus on implicit regularization. In Figure 6, we extend our findings to a family of ResNets trained on CIFAR-10 with 20% noisy labels, with the Adam optimizer, data augmentation (4-pixel shifts and random horizontal flips), as well as batch normalization layer (see appendix B for details). Both model-wise and epoch-wise trends reported for the ConvNets also hold for this setup, with the interpolation threshold occurring at base width $w = 18$. We consolidate our model-wise and epoch-wise findings with heatmaps in Figure 11 and 13. Interestingly, data augmentation causes the peak in test error to occur earlier than the interpolation threshold. We hypothesize that the mismatch – which can also be observed in related works (Nakkiran et al., 2019b) – is due to a lack of fine grained control over model size as base width w varies. Importantly, for large volumes around training points (7 augmentations per path), training accuracy degrades and loss sharpness increases. However, all sharpness metrics undergo double descent as model size grows, confirming the trends reported in simpler training settings.



(a)

(b)

Figure 6: (a) Double descent for the test error the CIFAR-10 with 20% noisy labels for a family of ResNet18s of increasing base width w, trained with data augmentation. (b) Accuracy, crossentropy, Jacobian and Hessian norms over volumes. All models are trained for 4k epochs. Analogous trends as observed for the ConvNets holds in this case. However, the largest integration volume considered, consisting of 7 augmentations per training sample, now shows considerably increased sharpness and loss curvature, while still undergoing double descent.



(b)

(a)

Figure 7: **Decoupling smoothness from generalization**. We present an experiment in which 20 randomly selected classes of the CIFAR-100 training split are corrupted with asymmetric label noise, perturbing 80% training labels within each class, for a total of 20% corrupted training samples. At test time, this enables splitting the test set into classes that have been corrupted at train time, and unperturbed classes. (a) Overparameterization promotes a smooth and flat loss landscape around both cleanly-labeled as well as noisy training samples under asymmetric label noise. (b) Confirming our hypothesis, double descent for the test error can still be observed for the unperturbed classes, while the trend disappears for the corrupted classes. This finding shows that overparameterization promotes smoothness in the input variable, which is aligned with generalization only around cleanly labeled points.

4.5 Towards Decoupling Smoothness from Generalization

Our experimental results suggest that double descent in the test error is closely related to input-space smoothness. One possible interpretation is that models at the interpolation threshold learn small and irregular decision regions, marked by high loss sharpness, while large models learn more regular decision regions with wider margins, supporting the observations of Jiang et al. (2019).

As consistently observed in our experiments, on the one hand, models near the interpolation threshold fail to smoothly interpolate all clean samples, while on the other hand large models can smoothly interpolate the entire training set. This effectively enforces a trade-off for which large models lose generalization ability around noisy samples, but can correctly classify all clean samples. Assuming the train and test distributions are similar (i.e. excluding covariate shifts), this would in turn result in improved average test error past the interpolation threshold, as indeed observed in practice.

To corroborate our interpretation, in principle, one would need to construct a nearest neighbour classifier (either in input or in feature space), and test whether predictions for each test sample are affected by proximity to corrupted samples. In the following, we conclude by proposing a simple experiment to decouple smoothness from generalization, without knowledge the relative density of training and test samples.

We consider CIFAR-100 and corrupt 20% of the training set with asymmetric label noise, such that 80% samples of 20 randomly selected classes are perturbed. At test time, this enables us to split the test set into samples whose classes have been corrupted, and samples belonging to unperturbed classes. Figure 7 shows that, even under strong asymmetric noise, overparameterization promotes input-space smoothness over increasingly large volumes around both clean and noisy training samples. Perhaps surprisingly, at test time double descent is still observed for test samples belonging to unperturbed classes, while the trend disappears for the corrupted classes. This confirms our interpretation and shows that double descent should be understood in terms of input-space smoothness, and its relation to generalization.

5 Conclusions

In this work, we develop simple geodesic Monte Carlo integration tools to study the input space of neural networks, providing intuition – built on extensive experiments – on how neural networks fit training data. We present a strong correlation between epoch-wise, model-wise double descent for the test error and smoothness of the loss landscape in input space. Our experiments show that overparameterization promotes input space regularity via smooth interpolation of noisy training data, which is aligned with improved generalization for datasets with relatively low ratio of label noise. Crucially, contrary to intuitions on polynomial regression, deep networks uniformly predict noisy training targets over volumes around noisily-labeled training samples – a behaviour which may have severe negative impact in practical applications with imbalanced training sets or where the population distribution differs (e.g. over time) from the training distribution.

Consistently throughout our experiments, we observe a peak in test error and loss sharpness near the interpolation threshold, which decreases for better generalizing models. Finally, for increasing volumes around each training point, we observe that overparametrization promotes flatter minima of the loss in input space, providing some initial clues as to why large overparameterized models generalize better, and corroborating the concurrent findings of Somepalli et al. (2022) on regularity of decision boundaries of overparameterized classifiers.

Our analysis substantiates the universal law of robustness of Bubeck & Sellke (2021), and extends the findings of Novak et al. (2018) to experimental settings with controlled model size. We hypothesize that overparameterization affects optimization dynamics, promoting a relatively flat loss landscape. An interesting open problem, which we leave for future work, is characterizing the impact of each model parameter on interpolation, as model size grows in depth and width.

References

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*, 2022.

Devansh Arpit, Víctor Campos, and Yoshua Bengio. How to initialize your network? robust initialization for weightnorm & resnets. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yoshua Bengio. Deep learning of representations: Looking forward. In *International conference on statistical language and speech processing*, pp. 1–37. Springer, 2013.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6974–6983, 2021.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021.

David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003. doi: 10.1073/pnas.1031596100.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020.

Matteo Gamba, Adrian Chmielewski-Anders, Josephine Sullivan, Hossein Azizpour, and Mårten Björkman. Are all linear regions created equal? In *International Conference on Artificial Intelligence and Statistics*, pp. 6573–6590. PMLR, 2022.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, 01 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.1.1.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guy Hacohen, Leshem Choshen, and Daphna Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3950–3960. PMLR, 13–18 Jul 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019.

Daniel LeJeune, Randall Balestriero, Hamid Javadi, and Richard G Baraniuk. Implicit rugosity regularization via data augmentation. *arXiv preprint arXiv:1905.11639*, 2019.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16805–16817. Curran Associates, Inc., 2021.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent, 2019a. URL `https://windowsontheory.org/2019/12/05/deep-double-descent/`.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019b.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations Workshop Track*, 2015.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2018.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8093–8104. PMLR, 13–18 Jul 2020.

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. *NeurIPS Workshops*, 2020.

Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13699–13708, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Tao Yu, Huan Long, and John E Hopcroft. Curvature-based comparison of two neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 441–447. IEEE, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2018.