Benign, Tempered, or Catastrophic: A Taxonomy of Overfitting

Anonymous Author(s)
Affiliation
Address
email

Abstract

The practical success of overparameterized neural networks has motivated the recent scientific study of *interpolating methods*– learning methods which are able to fit training data perfectly. Empirically, certain interpolating methods can fit noisy training data without catastrophically bad test performance, which defies standard intuitions from statistical learning theory. Aiming to explain this, a large body of recent work has studied *benign overfitting*, a behavior seen in certain asymptotic settings under which interpolating methods approach Bayes-optimality, even in the presence of noise. In this work, we argue that, while benign overfitting has been instructive to study, real interpolating methods like deep networks do not fit benignly. That is, noise in the train set leads to suboptimal generalization, suggesting that these methods fall in an intermediate regime between benign and catastrophic overfitting, where asymptotic risk is neither unbounded nor Bayes-optimal. We call this behavior *tempered overfitting*. We first study kernel (ridge) regression, obtaining conditions on the ridge parameter and kernel eigenspectrum under which KR exhibits each of the three behaviors, thus obtaining a relatively complete picture of overfitting for kernel regression. We find that kernels with powerlaw spectra, which are ubiquitous in practice, display tempered overfitting. We then provide empirical evidence that deep neural networks trained to interpolation are tempered, while those stopped early are benign.

1 Introduction

In the last decade, the dramatic success of overparameterized deep neural networks (DNNs), in defiance of classical intuitions from statistical learning theory, has inspired the field to reexamine the theoretical foundations of generalization. Classical statistical learning theory suggests that an algorithm which *interpolates* (i.e. perfectly fits) its training data will typically *catastrophically overfit* at test time, generalizing no better than a random function[1]. Figure 1c illustrates the catastrophic overfitting classically expected of an interpolating method. Yet DNNs can interpolate their training data (even in the presence of noise) and generalize well nonetheless [Neyshabur et al., 2015, Zhang et al., 2017]. This suggests the need for a new theoretical framework, replacing classical statistical learning theory, within which to understand this phenomenon.

[1]There are various ways to formalize this prediction depending on the setting: it is a consequence of the "bias-variance tradeoff" in statistics, the "bias-complexity tradeoff" in PAC learning, and "capacity control"-based generalization bounds in kernel ridge regression. For example, in supervised learning, generalization bounds for empirical risk minimization become vacuous when VC dimension is larger than the sample size, and many risk bounds for kernel ridge regression become vacuous at zero ridge [Christmann and Steinwart, 2007, Shalev-Shwartz and Ben-David, 2014].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

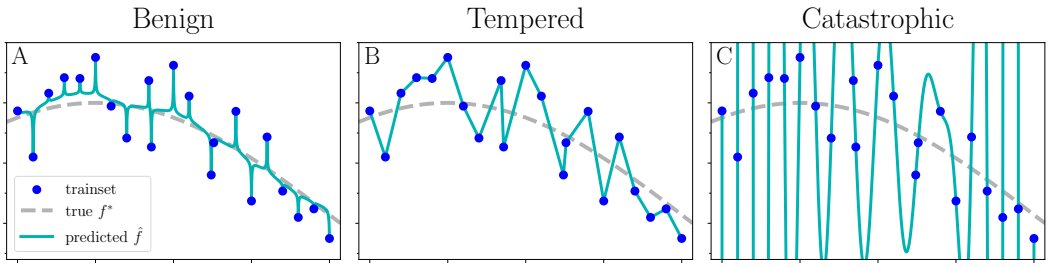


Figure 1: **As $n \to \infty$, interpolating methods can exhibit three types of overfitting. (A)** In *benign overfitting*, the predictor asymptotically approaches the ground-truth, Bayes-optimal function. Nadaraya-Watson kernel smoothing with a singular kernel, shown here, is asymptotically benign. **(B)** In *tempered overfitting*, the regime studied in this work, the predictor approaches a constant test risk greater than the Bayes-optimal risk. Piecewise-linear interpolation is (asymptotically) tempered. **(C)** In *catastrophic overfitting*, the predictor generalizes arbitrarily poorly. Rank-n polynomial interpolation is asymptotically catastrophic.

This need eventually led to the identification and study of *benign overfitting* [Bartlett et al., 2020] (also called "harmless interpolation" Muthukumar et al. [2020]), a phenomenon in which certain methods that perfectly fit the training data still approach *Bayes-optimal* generalization in the limit of large trainset size. Intuitively speaking, benignly-overfitting methods fit the target function globally, yet fit the noise only locally, and the addition of more label noise does not asymptotically degrade generalization. Figure 1a illustrates a simple method that is asymptotically benign. The study of benign overfitting has proven fruitful, leading to rich mathematical insights into high-dimensional learning[2]. And benign overfitting is certainly closer to the real behavior of DNNs compared to the catastrophic overfitting expected from classical intuition.

That said, simple experiments reveal that many standard DNNs *do not overfit benignly*: when training on noisy data, DNNs do not diverge catastrophically, but *neither* do they approach Bayes-optimal risk. Instead, they converge to a predictor that is neither catastrophic nor optimal, but rather somewhere in between— with error that increases as the noise in the data increases. Figure 2 depicts such an experiment: a ResNet is trained on a binary variant of CIFAR-10 with varying amounts of training label noise and increasing n, revealing that more train noise indeed causes more test error, even for large n. The observation that interpolating DNNs are harmed by label noise is well-known (e.g. Zhang et al. [2017]), and the linear shape of noise profile in Figure 2 is consistent with the prediction of the Feature Calibration Conjecture of Nakkiran and Bansal [2020]. These observations suggest that, in the search for a paradigm to understand modern interpolating methods, we should identify and study a regime intermediate between benign and catastrophic.

1.1 Summary of Contributions

In this work we formally identify an intermediate regime between benign and catastrophic overfitting, which we call *tempered overfitting*, because the noise's harmful effect is tempered but still nonzero. We find that, *even as the number of training examples n approaches infinity*, both DNNs trained to interpolation and (ridgeless) kernel regression (KR) using certain common kernels fall into this intermediate regime, as do common methods like 1-nearest-neighbors and piecewise linear interpolation (as in Figure 1b). Our tempered regime completes the taxonomy of overfitting: any learning procedure is either benign, tempered, or catastrophic in the asymptotic limit.

We begin in Section 2 with preliminaries, formal definitions of the three regimes of overfitting, and a taxonomy of some common ML methods according to these regimes. We generally consider an algorithm's limiting behavior as $n \to \infty$, as in the study of statistical consistency[3]. We are generally

[2] A partial list of works here include [Bartlett et al., 2020, 2021, Belkin et al., 2018a,b, 2019a, Cao et al., 2022, Chatterji and Long, 2021, Chatterji et al., 2021, Frei et al., 2022, Hastie et al., 2019, Koehler et al., 2021, Liang and Rakhlin, 2018, 2020, Muthukumar et al., 2020, Rakhlin and Zhai, 2019, Tsigler and Bartlett, 2020, Zhang et al., 2017, 2021].

[3] We note that statistical consistency is simply benign overfitting without the requirement of interpolation.

more interested in algorithms' asymptotic test risk than in the fact that they interpolate, and we consider non-interpolating methods as well as interpolating methods.

In Section 3, we study these three regimes for kernel regression (KR). Using recent spectral theories of the generalization of KR, we obtain conditions on the ridge parameter and kernel eigenspectrum under which KR falls into each of the three regimes. Importantly, we find that ridgeless kernels with powerlaw spectra, including the Laplace kernel and ReLU fully-connected NTKs, are asymptotically *tempered*, not benign. We confirm our theory with experiments on real and synthetic data.

In Section 4, we empirically study overfitting for DNNs. We give evidence that standard interpolating DNNs exhibit tempered overfitting, and not benign overfitting, which highlights the relevance of our framework for understanding modern methods. We conclude with a discussion in Section 5.

1.2 Relation to Prior Work

Our work is inspired by recent developments in the theory and empirics of interpolating methods [Bartlett et al., 2020, Belkin et al., 2019c,d, Ji et al., 2021, Koehler et al., 2021, Liang and Rakhlin, 2018, 2020, Muthukumar et al., 2020, Rakhlin and Zhai, 2019, Tsigler and Bartlett, 2020]. Many of these works prove that certain interpolating estimators are statistically consistent in certain settings— demonstrating benign overfitting . In contrast, we argue that a wide range of interpolators, including DNNs used in practice, are *not benign*. Our work is compatible with prior work because our formalism differs in several crucial ways, which leads us to different conclusions.

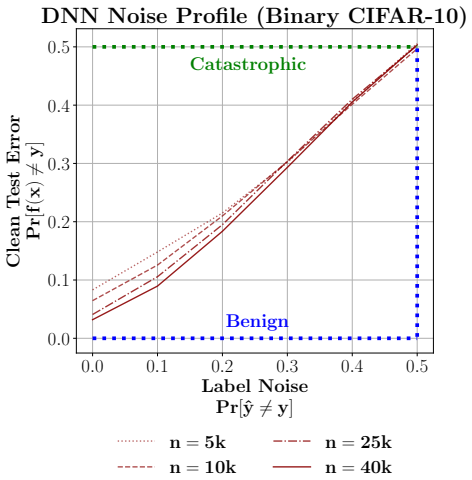


Figure 2: **DNNs trained on image data exhibit *tempered overfitting*, not benign overfitting.** Curves show test classification error vs. training label flip probability for a Wide ResNet 28×10 trained to interpolation on binary CIFAR-10 (animals vs. vehicles) for different training sizes n. These curves are *noise profiles*, as discussed in Section 2.

Briefly, many prior works assume that the target task always lies in "high enough dimension" relative to the number of samples n, while our work considers the limit $n \to \infty$ for tasks of fixed dimension d— as is the realistic asymptotic in practice. For example, in the linear regression setting, Bartlett et al. [2020] highlight that benign overfitting occurs most robustly when *input dimension grows faster than the sample size*. Other works explicitly scale the ambient problem dimension and sample size to infinity together, at a proportional rate [Hastie et al., 2019, Liang and Rakhlin, 2018, Mei and Montanari, 2019]. Taking a different approach, Frei et al. [2022] prove that interpolating two-layer networks can achieve close to optimal test error on certain distributions— but require an assumption that $n \leq \Omega(d)$. And Koehler et al. [2021] state a generalization bound for interpolators that decays to 0 with n, but also require $n \leq d$. In summary, a bulk of prior benign overfitting results apply in the regime where n is large, but still restricted to be smaller than the dimension of the problem, whereas we do not consider such restrictions.

Our work is compatible with prior works which observe that excess risk of certain interpolating methods decays with ambient dimension— thus high-dimensional problems enjoy a "blessing of dimensionality." For example, the simplicial interpolation scheme of Belkin et al. [2018a] has excess risk that decays as $O(2^{-d})$ for ambient dimension d. And Rakhlin and Zhai [2019] show that kernel ridgeless regression with the Laplace kernel is inconsistent in any fixed dimension — but with a lower-bound that decays with dimension. This agrees with our conclusions, that many interpolating methods are *tempered* on real distributions, which have fixed dimension.

2 The Three Types of Overfitting

Here we formally present our taxonomy of overfitting, delineating the three types of asymptotic behaviors which learning procedures can exhibit.

2.1 Definitions

We consider the standard setting of supervised learning. For simplicity, we present definitions for regression, but these are readily extended to classification. We wish to learn a function $\widehat{f} : \mathcal{X} \to \mathbb{R}$ from a size-n dataset $\mathcal{D}_n \equiv \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, where $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \mathbb{R}$, and $\mathcal{X}$ is the input domain. We evaluate the generalization performance of $\widehat{f}$ by the *mean squared risk*: $R(\widehat{f}) := \mathbb{E}_{x,y\sim\mathcal{D}}\left[(\widehat{f}(x) - y)^2\right]$. The Bayes-optimal regression function is given by $f^* := \operatorname{argmin}_f R(f)$, where the minimization is over all measurable functions, and has risk R^* which is called the *irreducible risk*. The *excess risk* of any function $\widehat{f}$ is given by $\overline{R}(\widehat{f}) := R(\widehat{f}) - R^*$. We say an estimator achieves *interpolation* if $\widehat{f}_n(x_i) = y_i$ for all $(x_i, y_i) \in \mathcal{D}_n$. These definitions are readily generalized to classification, using classification error in place of MSE.

Learning Procedure. The central objects in our taxonomy are *learning procedures*. Our definition of learning procedure is fully general, which allows us to discuss anything from neural networks to 1-nearest-neighbors. Informally, a learning procedure is a fully specification of which model to output on a given train set, specified for all sizes of train sets.

Formally, a *learning procedure* $\mathcal{A} := \{A_n\}_n$ is a sequence of (potentially stochastic) functions, indexed by sample size $n \in \mathbb{N}$. At each n, the function $A_n : \mathcal{D}_n \mapsto \widehat{f}_n$ inputs a train set $\mathcal{D}_n$ and outputs a "model" $\widehat{f}_n : \mathcal{X} \to \mathbb{R}$. Note that this n-dependence allows learning procedures to be *non-uniform*, and use different algorithms for each sample size n. For example, it allows procedures which scale-up the DNN model size along with the sample size n, or which vary the kernel bandwidth with n. The *expected risk* of a learning procedure on n examples from distribution $\mathcal{D}$ is $\mathcal{R}_n := \mathbb{E}_{A_n,\mathcal{D}_n}[R(A_n(\mathcal{D}_n))]$.

2.2 The Taxonomy

We shall categorize learning procedures in terms of their *asymptotic expected risk*. We handle regression and K-class classification settings separately, due to their different loss scalings. As the number of samples $n \to \infty$, the sequence of expected risks $\{\mathcal{R}_n\}$ can behave in one of three ways:

	Regression	**Classification**
Benign	$\lim_{n\to\infty} \mathcal{R}_n = R^*$	$\lim_{n\to\infty} \mathcal{R}_n = R^*$
Tempered	$\lim_{n\to\infty} \mathcal{R}_n \in (R^*, \infty)$	$\lim_{n\to\infty} \mathcal{R}_n \in (R^*, 1 - \frac{1}{K})$
Catastrophic	$\lim_{n\to\infty} \mathcal{R}_n = \infty$	$\lim_{n\to\infty} \mathcal{R}_n = 1 - \frac{1}{K}$

This defines our taxonomy. There is technically a fourth option – that the limit does not exist – but, to our knowledge, this does not describe any non-pathological algorithms. All together, this set of behaviors is exhaustive, describing every possible learning procedure. Note that the definitions for classification and regression are identical, except for the bounds at ∞ replaced by the error of the predictor choosing a uniformly random label $(1 - \frac{1}{K})$.

2.3 Applying Our Taxonomy

Our taxonomy handles general learning procedures, even those which do not interpolate their train sets[4]. Thus, we can apply it to describe many existing methods in machine learning. In doing so, we *refine the language* around statistical consistency: Many methods were known to be statistically inconsistent (i.e. not benign), but we highlight that there are two ways to be inconsistent: tempered or catastrophic.

[4]We use "overfitting" to describe interpolating methods, and "fitting" to describe general methods.

Benign	Tempered	Catastrophic
• Early-stopped DNNs • KR with ridge • k-NN ($k \sim \log n$) • Nadaraya-Watson kernel smoothing w/Hilbert kernel	• **Interpolating DNNs** • **Laplace KR** • **ReLU NTKs** • k-NN (constant k) • Simplicial interpolation	• **Gaussian KR** • Critically-parameterized regression

Table 1: A taxonomy of models under the three types of fitting identified in this work. **BOLD** are results from our work, others are known or folklore results.

To illustrate our taxonomy, we give several examples of known results in Table 1. Any statistically consistent method is by definition *benign*: this includes non-interpolating methods such as early-stopped DNNs [Ji et al., 2021] and k-nearest-neighbors with $k \sim \log n$ [Chaudhuri and Dasgupta, 2014, Cover and Hart, 1967], as well as interpolating methods such as Nadarya-Watson kernel smoothing [Devroye et al., 1998] and k-NN schemes [Belkin et al., 2018a, 2019c].

Catastrophic learning procedures are also abundant. Parametric models which are "critically-parameterized" (i.e. at the double descent peak) overfit catastrophically: this includes Random Feature regression with number of features $p = n$ [Mei and Montanari, 2019], and, more generally, for a broad class of linear and random feature methods [Holzmüller, 2021]. Generic empirical-risk-minimization for a hypothesis class with VC-dimension $> n$ will also overfit catastrophically in the worst case. It turns out that RBF kernel regression under natural assumptions also catastrophically overfits, as we show in Section 3.

Finally, many methods, from classical to modern, exhibit tempered overfitting. First, it is well-known that 1-nearest-neighbor converges to an asymptotic risk that is finite, but bounded away from Bayes optimal[5]— thus, it is tempered. Further, any "underfitting" method, such as Empirical Risk Minimization over a hypothesis class with VC-dimension $\ll n$ will be tempered, as long as this hypothesis class does not contain the ground-truth function. In Section 4, we empirically demonstrate that many standard DNNs exhibit tempered overfitting, when trained to interpolation. We complement this with theoretical results in Section 3, showing that kernel regression with the Laplace kernel, as well as with the ReLU NTK [Bietti and Bach, 2020, Li et al., 2020, Nguyen et al., 2021] exhibits tempered overfitting. Thus, our new category of tempered overfitting is not merely a theoretical possibility, but rather captures many natural and practically-used learning methods.

2.4 Noise Profiles

To study asymptotic risk, we use a metric we call a *noise profile*. A noise profile characterizes the *sensitivity* of a learning procedure to *noise* in the training set. For a given learning procedure $\mathcal{A}$ and data distribution $\mathcal{D}$, a *noise profile* P_n describes how the asymptotic risk varies with respect to σ, the level of artificial noise added to training targets. Formally, the noise profile $\mathsf{P}_{\mathcal{A}}$ is

$$\mathsf{P}_{\mathcal{A}}(\sigma) = \lim_{n\to\infty} \mathop{\mathbb{E}}_{A_n;\mathcal{D}_n(\sigma)} R(A_n(D_n(\sigma))). \tag{1}$$

where $\mathcal{D}(\sigma)$ denotes n i.i.d. samples from the distribution $\mathcal{D}$, with σ-level of label noise added. The label noise $\sigma \in \mathbb{R}$ denotes a kind of noise that depends on the problem setting: σ is the variance of additive Gaussian noise for regression settings, or the label-flip probability for classification settings. We cannot empirically evaluate the $n \to \infty$ limit exactly, so we estimate it using asymptotics from finite but large sample sizes. As shown in Figure 2, noise profiles are easily plotted and reveal at a glance whether an learning procedure is benign, tempered, or catastrophic.

3 Overfitting in Kernel Regression

Before characterizing the asymptotic behavior for DNNs via experiment, we first turn to a simpler learning method for which these three regimes can be theoretically understood. We study *kernel*

[5] In fact, the excess test MSE converges to the variance of the observation noise.

regression (KR), a widely-used nonparameteric learning algorithm. Theoretical interest in this algorithm has increased significantly in recent years due to the discovery that trained neural networks converge to KR in the infinite-width and infinite-time limit Jacot et al. [2018], suggesting that insights into KR simultaneously shed light on overparameterized neural networks.

KR is fully specified by a positive-semidefinite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a ridge parameter $\delta \geq 0$. As described above, we allow the training set $\mathcal{D}_n$ to contain n samples $(x_i, y_i) \sim \mathcal{D}$, where here we assume that $y_i = f^*(x_i) + \eta_i$, where $\eta_i \sim \mathcal{N}(0, \sigma^2)$. KR returns the predicted function $\widehat{f}$ given by

$$\widehat{f}(x) = K(x, \mathcal{D}_n) \left(K(\mathcal{D}_n, \mathcal{D}_n) + \delta \mathbf{I}_n\right)^{-1} \mathcal{Y}, \tag{2}$$

where $K(\mathcal{D}_n, \mathcal{D}_n)$ is the "data-data kernel matrix" with components $K(\mathcal{D}_n, \mathcal{D}_n)_{ij} = K(x_i, x_j)$, and $K(x, \mathcal{D})$ is a row vector with components $K(x, \mathcal{D})_i = K(x, x_i)$, and $\mathcal{Y}$ is a column vector of target labels[6].

Existing literature already provides examples of KR exhibiting all three asymptotic behaviors. Benign overfitting has been analyzed for highly-overparameterized linear regression [Bartlett et al., 2020, Belkin et al., 2019b, Hastie et al., 2019, Muthukumar et al., 2020], a special case of KR. It is well-known that KR with a positive ridge value (with appropriate scaling conventions) is consistent [Christmann and Steinwart, 2007]. Furthermore, in 1D, a Laplace kernel approaches piecewise linear interpolation as $n \to \infty$ [Belkin et al., 2018a], which exhibits tempered overfitting, and Rakhlin and Zhai [2019] proved that the Laplace kernel does not overfit benignly in (fixed) dimension greater than one. Finally, it is known in experimental folklore (though not theoretically, to our knowledge) that KR with a Gaussian kernel and zero ridge tends to yield poorly-conditioned kernel matrices and catastrophic behavior. Here we derive *general conditions* under which KR falls into each regime, solidifying these various observations into a unified picture.

As our chief tool for obtaining these conditions, we use recently-derived closed-form expressions for the generalization MSE of KR in terms of the eigenspectrum of the kernel (with respect to the data measure) and the eigendecomposition of the target function. These eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and orthonormal eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$ are given by

$$\mathbb{E}_{x' \sim p}[K(x, x')\phi_i(x')] = \lambda_i \phi_i(x), \quad \text{where} \quad \mathbb{E}_{x \sim p}[\phi_i(x)\phi_j(x)] = \delta_{ij}. \tag{3}$$

We note that $\lambda_i \geq 0$ because a kernel must be positive-semidefinite and that $\sum_i \lambda_i = \mathrm{Tr}[K] \equiv \mathbb{E}_{x \sim p}[K(x, x)]$, which we assume is finite. We shall assume that the eigenvalues are indexed in decreasing order, with $\lambda_i \geq \lambda_j$ if $i < j$. Because the eigenfunctions form a complete basis, we can also decompose the target function as $f^*(x) = \sum_i v_i \phi_i(x)$, where $\{v_i\}_{i=1}^{\infty}$ are the eigencoefficients.

Recent results have shown that, remarkably, knowledge of the eigenvalues and eigencoefficients (but not the eigenfunctions) is typically sufficient to approximate test MSE. Bordelon et al. [2020] and Jacot et al. [2020] derived equivalent expressions for test MSE using the replica method and random matrix theory respectively, and Simon et al. [2021] then found that this result can be derived and expressed more simply in terms of *modewise learnabilities* $\{\mathcal{L}_i\}_{i=1}^{\infty}$, a set of scores in $[0, 1]$ which indicate how well each eigenmode is learned at a given n. These approximations are nonrigorous (relying on informal but well-accepted methods from statistical physics), but comparisons with true MSE using real and synthetic datasets generally reveal a close match, particularly at large n. Using the formulation of Simon et al. [2021], test MSE $\mathcal{R}_n$ is approximated by

$$\begin{aligned} &\mathcal{R}_n \approx \mathcal{E}_n \equiv \mathcal{E}_0 \left(\sum_i (1 - \mathcal{L}_i)^2 v_i^2 + \sigma^2 \right), \quad \text{where} \quad \mathcal{E}_0 \equiv \frac{n}{n - \sum_j \mathcal{L}_j^2}, \\ &\mathcal{L}_i \equiv \frac{\lambda_i}{\lambda_i + C}, \quad \text{and} \quad C \geq 0 \text{ satisfies } \sum_i \frac{\lambda_i}{\lambda_i + C} + \frac{\delta}{C} = n. \end{aligned} \tag{4}$$

This MSE includes noise on test labels; to compute the *excess* risk, one would simply subtract σ^2.

Here we study the asymptotic behavior of $\mathcal{E}_n$ as $n \to \infty$ for varying eigenspectra and ridge values. The fitting regime of the kernel is then given by $\mathcal{E}$: (a) if $\lim_{n \to \infty} \mathcal{E} = \sigma^2$ (the Bayes-optimal MSE), then fitting is *benign*, (b) if $\lim_{n \to \infty} \mathcal{E} \in (\sigma^2, \infty)$, then fitting is *tempered*, and (c) if $\lim_{n \to \infty} \mathcal{E} = \infty$,

[6] We note that some works use $n\delta$ instead of δ in Equation 2. This convention gives an asymptotically-biased estimator and changes some downstream conclusions about KR with ridge.

then fitting is *catastrophic*. We obtain conditions on the kernel eigenspectrum under which the kernel fits according to each of these three regimes.

Our proofs rely on several (quite weak) technical assumptions on $\{\lambda_i\}_i$ and $\{v_i\}_i$, the most important of which is that the target function does not place weight in zero-eigenvalue modes (i.e. outside the kernel's RKHS). We defer enumeration and discussion of these conditions to Appendix A, where they are captured by Assumption A.1.

Our result is the following:

Theorem 3.1 (KRR trichotomy)**.** *For $\{\lambda_i\}_{i=1}^{\infty}$ and $\{v\}_{i=1}^{\infty}$ satisfying Assumption A.1, $\sigma^2 > 0$, and $\mathcal{E}_n$ given by Equation 4,*

(a) If $\delta > 0$ or $\lambda_i = i^{-1}\log^{-\alpha} i$ for some $\alpha > 1$, then $\lim_{n\to\infty} \mathcal{E}_n = \sigma^2$.

(b) If $\delta = 0$ and $\lambda_i = i^{-\alpha}$ for some $\alpha > 1$, then $\lim_{n\to\infty} \mathcal{E}_n = \alpha\sigma^2$.

(c) If $\delta = 0$ and $\frac{\lambda_i}{\lambda_{i+1}} \geq \frac{i^{-\log i}}{(i+1)^{-\log(i+1)}}$ for all i, then $\lim_{n\to\infty} \mathcal{E}_n = \infty$.

We defer the proof to Appendix A. The proof proceeds by first showing that asymptotic MSE is dominated by the noise, and then finding $\lim_{n\to\infty} \mathcal{E}_0$ in each of the three cases. Theorem 3.1 can be summarized as follows: a ridge parameter leads to benign fitting, powerlaw decay of eigenvalues leads to tempered overfitting, and eigenvalue decay at least as fast as $i^{-\log i}$ leads to catastrophic overfitting[7]. We note that the three conditions of Theorem 3.1 are not exhaustive; in particular, spectra can asymptotically decay as powerlaws without matching the exact form in (b), or be $\mathcal{O}(i^{-\log i})$ without obeying the condition for (c), and there of course exist decay rates intermediate between powerlaws and $i^{-\log i}$. Nonetheless, we suggest these three conditions heuristically capture the spectral decay of virtually all kernels used in practice and conjecture that a similar result with exhaustive conditions could be proved.

Theorem 3.1 has several consequences for KR with familiar kernels and the training of infinite-width networks. For illustrative purposes, we contrast the Gaussian (RBF) kernel $K_G(x_1, x_2) = e^{-w^{-2}\|x_1-x_2\|_2^2}$ with the Laplace kernel $K_L(x_1, x_2) = e^{-w^{-1}\|x_1-x_2\|_2}$, where w is a width parameter. With data drawn from a d-dimensional manifold, the Gaussian and Laplace kernels have eigenspectra that decay like $\lambda_i \sim e^{-i^{2/d}}$ (as can be seen by taking a Fourier transform of K_G) and $\lambda_i \sim i^{-(d+1)/d}$, respectively [Bietti and Bach, 2020]. We note that ReLU NTKs, restricted to the hypersphere, have the same eigendecay as the Laplace kernel [Chen and Xu, 2020, Geifman et al., 2020]. The consequences of Theorem 3.1 include the following:

- KR with a ridge parameter and any sufficiently-expressive kernel will fit benignly as $n \to \infty$.
- Ridgeless KR with *Laplace kernels* or *ReLU NTKs* will exhibit *tempered overfitting*, approaching benignness as dimension grows[8].
- Ridgeless KR with *Gaussian kernels* will exhibit *catastrophic overfitting*.
- The asymptotic behavior of KR with non-ReLU NTKs *depends on the activation function*. Virtually any kernel on the d-sphere (including the Gaussian kernel) can be realized as the NTK of a wide network with a proper choice of activation function [Simon et al., 2022], and thus there exist choices of activation function that yield catastrophic overfitting.
- Early stopping is known to act as an effective ridge parameter for wide networks [Ali et al., 2019], and thus we should expect that early-stopped wide networks will fit benignly.

4 Experiments

Here we report experimental results confirming our theoretical conclusions for kernel regression, and showing that DNNs can be either benign or tempered, depending on whether they are stopped early. We provide complete experimental details in Appendix B.

[7] Compare with Theorem 6.1 of Bartlett et al. [2020], which found (in a slightly different setting) that a kernel spectrum $\lambda_i \sim (i \log^{\beta} i)^{-1}$ with $\beta > 1$ leads to benign behavior, which heuristically matches our result as this decay rate is like taking $\alpha \to 1$ while ensuring the sum remains finite. Their theorem also notes that powerlaw decays with $\alpha > 1$ are not benign; here we find that they are in fact *tempered*.

[8] This is compatible with the lower-bounds in Rakhlin and Zhai [2019].

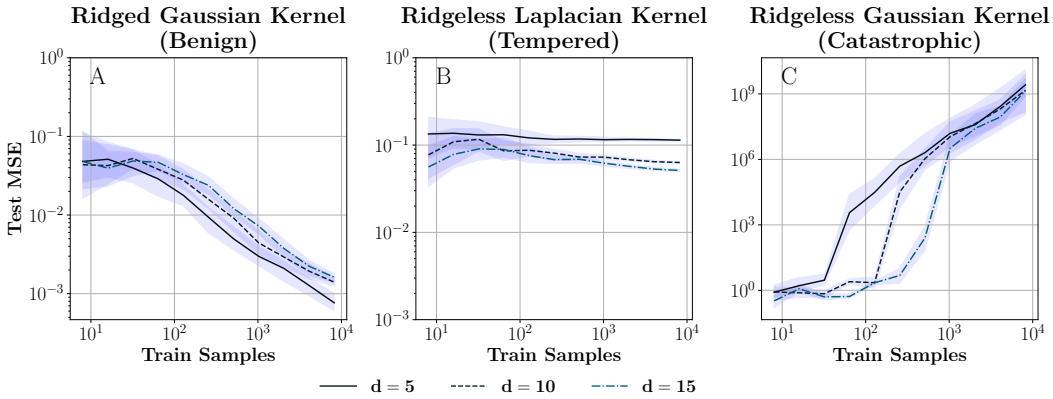
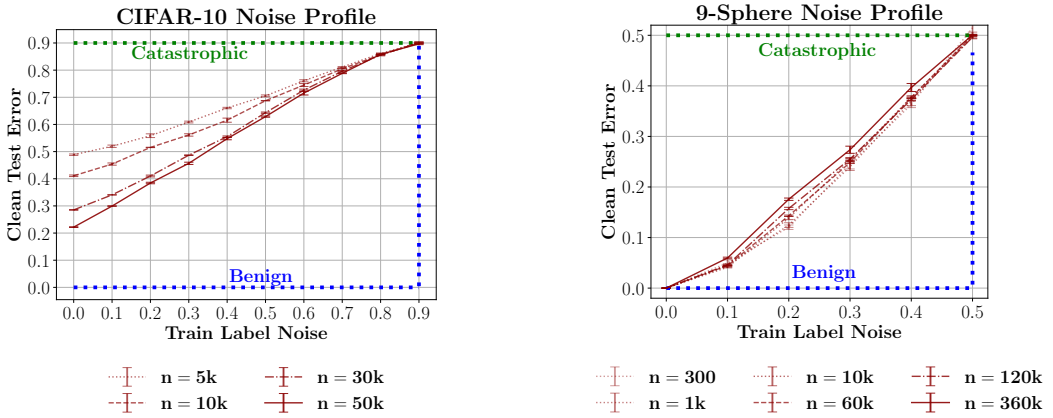


Figure 3: **Kernel Regression** can exhibit all three fitting regimes with proper choice of ridge parameter and kernel. Plots show learning curves for KR with data $\{x_i\}$ sampled uniformly from the unit sphere $\mathcal{S}^{d-1}$, trained with pure noise target labels $y_i \sim \mathcal{N}(0, 1)$. Test MSE is computed with respect to a clean test set. **(a)** KR with a Gaussian kernel and nonzero ridge is asymptotically benign. A ridge value of $\delta = 0.1$ was used. **(b)** Ridgeless KR with a Laplace kernel exhibits tempered overfitting. **(c)** Ridgeless KR with a Gaussian kernel exhibits catastrophic overfitting.



(a) WideResNet trained on CIFAR-10.

(b) MLP trained on the constant function.

Figure 4: Noise Profiles for DNNs on CIFAR-10 and synthetic data.

4.1 Experiments on Kernel Regression

We begin by running KR with (a) Gaussian kernel + ridge parmeter, (b) Laplace kernel + no ridge parameter, and (c) Gaussian kernel + no ridge parameter, which should exhibit benign, tempered, and catastrophic behavior, respectively, according to Theorem 3.1. The results, shown in Figure 3, confirm our theoretical predictions. Figure 3c also shows that increasing input dimension tends to both improve MSE at fixed n and increase the critical n at which MSE begins to diverge, suggesting that high dimensionality (which is known to encourage tempered methods to be benign) also acts to prevent catastrophic methods from diverging.

4.2 Experiments on Deep Neural Networks

We first present experimental results for Wide ResNets (28×10) trained on standard image datasets. We train on MSE loss and measure test classification error, and consider interpolation to mean a training accuracy of 100%. For each dataset, we train at varying n *and* with varying *label noise*, quantified by the probability with which each training label is uniformly resampled. We are primarily interested in asymptotic noise-fitting behavior at large n.

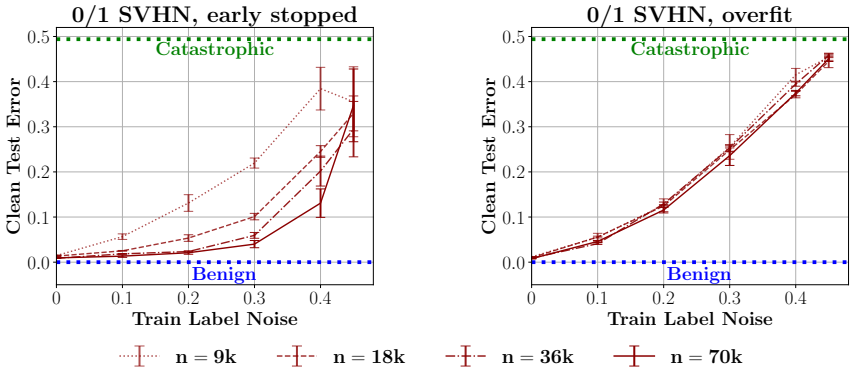


Figure 5: **Early-stopped DNNs exhibit fairly benign overfitting, while DNNs trained to interpolation exhibit tempered overfitting.** We show noise profiles for ResNets trained on SVHN, both (a) stopped early and (b) trained to interpolation.

Figure 2 and Figure 4a show noise profiles for ResNets trained to interpolation on *two-class* and *10-class* CIFAR-10, respectively. In both cases, fitting is tempered, and the noise profile rapidly approaches a limiting, fairly linear curve as $n \to \infty$.

Figure 5 shows noise profiles for ResNets trained on SVHN, both *stopped early* (Figure 5a) and *trained to interpolation* (Figure 5b). The early-stopped ResNets approaches benign fitting as n grows, with low error even at sizable noise levels, while the ResNets trained to interpolation quickly converge to a tempered noise profile. In Figure 7, we plot MSE vs. training time for an MLP trained on synthetic noisy data sampled uniformly on the hypersphere, showing this transition between regimes directly. Finally, Figure 4b shows the noise profile for an MLP trained to interpolation on an extremely simple synthetic dataset, where the ground-truth function is *constant*. Even in this simple setting, interpolating DNNs are not benign— they do not successfully learn the constant function (experimental details in Appendix B.3).

5 Conclusion

In this paper we study the nature of overfitting in learning methods which interpolate their training data. Much of contemporary theory and experimental work categorizes overfitting as either catastrophic or benign. In contrast, we observe that many natural learning procedures, including DNNs used in practice, overfit in a manner that is neither benign nor catastrophic— but rather in an intermediate regime. We identify and formally define this regime, which we call *tempered overfitting*.

We present empirical evidence of learning procedures that exhibit tempered overfitting, on both synthetic and natural data, using kernel machines and deep neural networks. We show tempered over-fitting can be quantified in terms of noise profiles, which measure how asymptotic performance depends on noise in the train distribution. For kernel regression, we provide a theoretical result in the form of a trichotomy: conditions on problem parameters which yield each of the three regimes of overfitting.

Our work presents an initial study of tempered overfitting, and lays the framework for future study of what we believe is a rich and relevant regime for modern learning. We hope this framework inspires further investigation into tempered overfitting for more complex models, both theoretically and experimentally. For example, it is open to understand which conditions on neural network architecture and data distribution lead to benign, tempered, or catastrophic overfitting— an analogue of our "kernel trichotomy" for deep networks.

References

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 1370–1378. PMLR, 2019. 3

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. 1, 2, 1.2, 3, 7

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021. 2

Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2306–2317, 2018a. URL https://proceedings.neurips.cc/paper/2018/hash/e22312179bf43e61576081a2f250f845-Abstract.html. 2, 1.2, 2.3, 3

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 540–548. PMLR, 2018b. URL http://proceedings.mlr.press/v80/belkin18a.html. 2

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, 2019a. 2

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b. 3

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 2019c. URL http://proceedings.mlr.press/v89/belkin19a.html. 1.2, 2.3

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019d. 1.2

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020. 2.3, 3

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020. 3

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022. 2

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*, 2020.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *arXiv preprint arXiv:2110.02914*, 2021. 2

Niladri S Chatterji, Philip M Long, and Peter L Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *arXiv preprint arXiv:2108.11489*, 2021. 2

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014. 2.3

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020. 3

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007. 1, 3

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967. 2.3

Luke Nicholas Darlow, Elliot J. Crowley, Antreas Antoniou, and Amos J. Storkey. Cinic-10 is not imagenet or cifar-10. *ArXiv*, abs/1810.03505, 2018. B.1

Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998. 2.3

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928*, 2022. 2, 1.2

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 33:1451–1461, 2020. 3

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019. 2, 1.2, 3

David Holzmüller. On the universality of the double descent peak in ridgeless regression. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IO5VdnSAaH. 2.3

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 3

Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020. 3

Ziwei Ji, Justin Li, and Matus Telgarsky. Early-stopped neural networks are consistent. *Advances in Neural Information Processing Systems*, 34, 2021. 1.2, 2.3, C.2

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 1.2

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. B.1

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. B.1

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/. B.1

Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pages 2613–2682. PMLR, 2020. 2.3

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018. 2, 1.2

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “ridgeless” regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020. 2, 1.2

Siyuan Ma and Mikhail Belkin. Diving into the shallows: a computational perspective on large-scale shallow learning. *Advances in neural information processing systems*, 30, 2017. B.2

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019. 1.2, 2.3

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 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*, 2020. 1, 2, 1.2, 3

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020. 1

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. B.1

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015. 1

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR, 2021. 2.3

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2595–2623. PMLR, 2019. URL http://proceedings.mlr.press/v99/rakhlin19a.html. 2, 1.2, 3, 8

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 1

James B Simon, Madeline Dickens, and Michael R DeWeese. A theory of the generalization and inductive bias of kernel ridge regression and wide neural networks. *arXiv preprint arXiv:2110.03922*, 2021. 3, A, A

James B. Simon, Sajant Anand, and Michael Robert DeWeese. Reverse engineering the neural tangent kernel. *arXiv preprint arXiv:2106.03186*, 2022. 3

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556. B.2

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020. 2, 1.2

Ke Alexander Wang, Niladri S Chatterji, Saminul Haque, and Tatsunori Hashimoto. Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*, 2021.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. URL http://www.bmva.org/bmvc/2016/papers/paper087/index.html. B.2

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx. 1, 1, 2, B.2

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 2

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We investigated a fundamental property of learning algorithms in a general sense.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We use entirely public datasets and provide citations and links in the appendix. All model code is included, when we use public code citations are provided. Information for reproduction is in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Except in one plot for which we ran into compute constraints.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Resources are listed in appendix, total compute is estimated.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Appendix
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We use commonly used computer vision datasets in the public domain: MNIST, CIFAR-10, SVHN
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]