NOISY INTERPOLATION LEARNING WITH SHALLOW UNIVARIATE RELU NETWORKS

Nirmit Joshi
TTI-Chicago
nirmit@ttic.edu

Gal Vardi
TTI-Chicago and Hebrew University
galvardi@ttic.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu

ABSTRACT

Understanding how overparameterized neural networks generalize despite perfect interpolation of noisy training data is a fundamental question. Mallinar et al. (2022) noted that neural networks seem to often exhibit "tempered overfitting", wherein the population risk does not converge to the Bayes optimal error, but neither does it approach infinity, yielding non-trivial generalization. However, this has not been studied rigorously. We provide the first rigorous analysis of the overfitting behavior of regression with minimum norm (ℓ_2 of weights), focusing on univariate two-layer ReLU networks. We show overfitting is tempered (with high probability) when measured with respect to the L_1 loss, but also show that the situation is more complex than suggested by Mallinar et al., and overfitting is catastrophic with respect to the L_2 loss, or when taking an expectation over the training set.

1 INTRODUCTION

A recent realization is that, although sometimes overfitting can be catastrophic as suggested by our classic learning theory understanding, in other cases overfitting may not be so catastrophic. In fact, even *interpolation learning*, which entails achieving zero training error with noisy data, can still allow for good generalization, and even consistency (Zhang et al., 2017; Belkin et al., 2018). This has led to efforts towards understanding the nature of overfitting: how *benign* or *catastrophic* it is, and what determines this behavior, in different settings and using different models.

Although interest in benign overfitting stems from the empirical success of interpolating large neural networks, theoretical study so far has been mostly limited to linear and kernel methods, or to classification settings where the data is already linearly separable, with very high data dimension (tending to infinity as the sample size grows)[1]. But what about noisy interpolation learning in low dimensions, using neural networks?

[1]Minimum ℓ_2 norm linear prediction (aka ridgeless regression) with noisy labels and (sub-)Gaussian features has been studied extensively (e.g. Hastie et al., 2020; Belkin et al., 2020; Bartlett et al., 2020; Muthukumar et al., 2020; Negrea et al., 2020; Chinot & Lerasle, 2020; Koehler et al., 2021; Wu & Xu, 2020; Tsigler & Bartlett, 2020; Zhou et al., 2022; Wang et al., 2022; Chatterji et al., 2021; Bartlett & Long, 2021; Shamir, 2022; Ghosh & Belkin, 2022; Chatterji & Long, 2021; Wang & Thrampoulidis, 2021; Cao et al., 2021; Muthukumar et al., 2021; Montanari et al., 2020; Liang & Recht, 2021; Thrampoulidis et al., 2020; Wang et al., 2021; Donhauser et al., 2022; Frei et al., 2023), and noisy minimum ℓ_1 linear prediction (aka Basis Persuit) has also been considered (e.g. Ju et al., 2020; Koehler et al., 2021; Wang et al., 2022). Either way, these analyses are all in the high dimensional setting, with dimension going to infinity, since to allow for interpolation the dimension must be high, higher than the number of samples. Kernel methods amount to a minimum ℓ_2 norm linear prediction, with very non-Gaussian features. But existing analyses of interpolation learning with kernel methods rely on "Gaussian Universality": either assuming as an ansatz the behavior is as for Gaussian features (Mallinar et al., 2022) or establishing this rigorously in certain high dimensional scalings (Hastie et al., 2019; Misiakiewicz, 2022; Mei & Montanari, 2022). In particular, such analyses are only valid when the input dimension goes to infinity (though possibly slower than the number of samples) and not for fixed low or moderate dimensions. Frei et al. (2022; 2023); Cao et al. (2022); Kou et al. (2023) study interpolation learning with neural networks, but only with high input dimension and when the data is interpolatable also with a linear predictor—in these cases, although non-linear neural networks are used, the results show they behave similarly to linear predictors. Manoj & Srebro (2023) take the other extreme and study interpolation learning with "short programs", which are certainly non-linear, but this is an abstract model that does not directly capture learning with neural networks.

Mallinar et al. (2022) conducted simulations with neural networks and observed *"tempered"* overfitting: the asymptotic risk does not approach the Bayes-optimal risk (there is no consistency), but neither does it diverge to infinity catastrophically. Such "tempered" behavior is well understood for 1-nearest neighbor, where the asymptotic risk is roughly twice the Bayes risk (Cover & Hart, 1967), and Mallinar et al. heuristically explain it also for some kernel methods. However, we do not have a satisfying and rigorous understanding of such behavior in neural networks, nor a more quantitative understanding of just how bad the risk might be when interpolating noisy data using a neural net.

In this paper, we begin rigorously studying the effect of overfitting in the noisy regression setting, with neural networks in low dimensions, where the data is *not* linearly interpolatable. Specifically, we study interpolation learning of univariate data (i.e. in one dimension) using a two-layer ReLU network (with a skip connection), which is a predictor $f_{\theta,a_0,b_0} : \mathbb{R} \to \mathbb{R}$ given by:

$$f_{\theta,a_0,b_0}(x) = \sum_{j=1}^{m} a_j (w_j x + b_j)_+ + a_0 x + b_0 \,, \tag{1}$$

where $\theta \in \mathbb{R}^{3m}$ denotes the weights (parameters) $\{a_j, w_j, b_j\}_{j=1}^m$. To allow for interpolation we do not limit the width m, and learn by minimizing the norm of the weights (Savarese et al., 2019; Ergen & Pilanci, 2021; Hanin, 2021; Debarre et al., 2022; Boursier & Flammarion, 2023):

$$\hat{f}_S = \arg\min_{f_{\theta,a_0,b_0}} \|\theta\|^2 \;\; \text{s.t.} \;\; \forall i \in [n],\; f_{\theta,a_0,b_0}(x_i) = y_i \;\; \text{where } S = \{(x_1, y_1), \ldots, (x_n, y_n)\}. \tag{2}$$

Following Boursier & Flammarion (2023) we allow an unregularized skip-connection in equation 1, where the weights a_0, b_0 of this skip connection are not included in the norm $\|\theta\|$ in equation 2. This skip connection avoids some complications and allows better characterizing $\hat{f}_S$ but does not meaningfully change the behavior (see Section 2).

Why min norm? Using unbounded size minimum weight-norm networks is natural for interpolation learning. It parallels the study of minimum norm high (even infinite) dimension linear predictors. For interpolation, we must allow the number of parameters to increase as the sample size increases. But to have any hope of generalization, we must choose among the infinitely many zero training error networks somehow, and it seems that some sort of explicit or implicit low norm bias is the driving force in learning with large overparameterized neural networks (Neyshabur et al., 2014). Seeking minimum ℓ_2 norm weights is natural, e.g. as a result of small weight decay. Even without explicit weight decay, optimizing using gradient descent is also related to an implicit bias toward low ℓ_2 norm: this can be made precise for linear models and for classification with ReLU networks (Chizat & Bach, 2020; Safran et al., 2022). For regression with ReLU networks, as we study here, the implicit bias is not well understood (see Vardi (2023)), and studying equation 2 is a good starting point for understanding the behavior of networks learned via gradient descent even without explicit weight decay. Interestingly, minimum-norm interpolation corresponds to the *rich regime*, and does not correspond to any kernel (Savarese et al., 2019). For the aforementioned reasons, understanding the properties of min-norm interpolators has attracted much interest in recent years (Savarese et al., 2019; Ongie et al., 2019; Ergen & Pilanci, 2021; Hanin, 2021; Debarre et al., 2022; Boursier & Flammarion, 2023).

Noisy interpolation learning. We consider a noisy distribution $\mathcal{D}$ over $[0,1] \times \mathbb{R}$:

$$x \sim \text{Uniform}([0,1]) \quad \text{and} \quad y = f^*(x) + \epsilon \;\; \text{with } \epsilon \text{ independent of } x, \tag{3}$$

where x is uniform for simplicity and concreteness[2]. The noise ϵ follows some arbitrary (but non-zero) distribution, and learning is based on an i.i.d. training set $S \sim \mathcal{D}^n$. Since the noise is non-zero, the "ground truth" predictor f^* has non-zero training error, seeking a training error much smaller than that of f^* would be overfitting (fitting the noise) and necessarily cause the complexity (e.g. norm) of the learned predictor to explode. The "right" thing to do is to balance between the training error and the complexity $\|\theta\|$. Indeed, under mild assumptions, this balanced approach leads to asymptotic consistency, with $\hat{f}_S \xrightarrow{n\to\infty} f^*$ and the asymptotic population risk of $\hat{f}_S$ converging to the Bayes risk. But what happens when we overfit and use the interpolating learning rule equation 2?

[2]All our results should also hold for any absolutely continuous distribution with bounded density and support. Roughly speaking, this can be achieved by dividing the support into disjoint intervals such that the distribution in each interval is well-approximated by a uniform distribution.

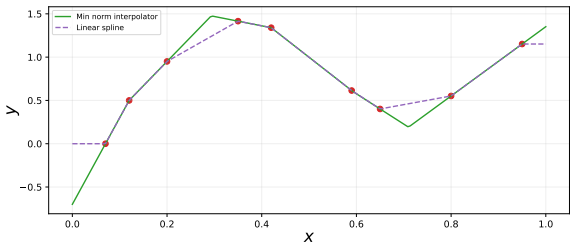

Figure 1: Comparison between linear-spline (purple) and min-norm (green) interpolators.

Linear Splines. At first glance, we might be tempted to think that two-layer ReLUs behave like linear splines (see Figure 1). Indeed, if minimizing the norm of weights w_i and a_i but *not* the biases b_i in equation 2, linear splines are a valid minimizer (Savarese et al., 2019; Ergen & Pilanci, 2021). As the number of noisy training points increases, linear splines "zig-zag" with tighter "zigs" but non-vanishing "amplitude" around f^*, resulting in an interpolator which roughly behaves like f^* plus some added non-vanishing "noise". This does not lead to consistency, but is similar to a nearest-neighbor predictor (each prediction is a weighted average of two neighbors). Indeed, in Theorem 1 of Section 3, we show that linear splines exhibit "tempered" behavior, with asymptotic risk proportional to the noise level.

From Splines to Min-Norm ReLU Nets. It turns out minimum norm ReLU networks, although piecewise linear, are not quite linear splines: roughly speaking, and as shown in Figure 1, they are more conservative in the number of linear "pieces". Because of this, in convex (conversely, concave) regions of the linear spline, minimum norm ReLU nets "overshoot" the linear spline in order to avoid breaking linear pieces. This creates additional "spikes", extending above and below the data points (see Figures 1 and 2) and thus potentially increasing the error. In fact, such spikes are also observed in interpolants reached by gradient descent (Shevchenko et al., 2022, Figure 1). How bad is the effect of such spikes on the population risk?

Our Contribution

Effect of Overfitting on L_p Risk. It turns out the answer is quite subtle and, despite considering the same interpolator, the nature of overfitting actually depends on how we measure the error. For a function $f : \mathbb{R} \to \mathbb{R}$, we measure its L_p population error and the reconstruction error respectively as

$$\mathcal{L}_p(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[|f(x) - y|^p] \quad \text{and} \quad \mathcal{R}_p(f) := \mathop{\mathbb{E}}_{x\sim\text{Uniform}([0,1])}[|f(x) - f^*(x)|^p].$$

We show in Theorems 2 and 3 of Section 4.2 that for $1 \le p < 2$,

$$\mathcal{L}_p(\hat{f}_S) \xrightarrow{n\to\infty} \Theta\left(\frac{1}{(2-p)_+}\right) \mathcal{L}_p(f^*). \tag{4}$$

This is an upper bound for any Lipschitz target f^* and any noise distribution, and it is matched by a lower bound for Gaussian noise. That is, for abs-loss (L_1 risk), as well as any L_p risk for $1 \le p < 2$, overfitting is **tempered**. But this tempered behavior explodes as $p \to 2$, and we see a sharp transition. We show in Theorem 4 of Section 4.3 that for any $p \ge 2$, including for the square loss ($p = 2$), in the presence of noise, $\mathcal{L}_p(\hat{f}_S) \xrightarrow{n\to\infty} \infty$ and overfitting is **catastrophic**.

Convergence vs. Expectation. The behavior is even more subtle, in that even for $1 \le p < 2$, although the risk $\mathcal{L}_p(\hat{f}_S)$ converges in probability to a tempered behavior as in equation 4, its *expectation* is infinite: $\mathbb{E}_S[\mathcal{L}_p(\hat{f}_S)] = \infty$. Note that in studying tempered overfitting, Mallinar et al. (2022) focused on this expectation, and so would have categorized the behavior as "catastrophic" even for $p = 1$, emphasizing the need for more careful consideration of the effect of overfitting.

I.I.D. Samples vs. Samples on a Grid. The catastrophic effect of interpolation on the L_p risk with $p \ge 2$ is a result of the effect of fluctuations in the *spacing* of the training points. Large, catastrophic, spikes are formed by training points extremely close to their neighbors but with different labels (see Figures 2 and 5). To help understand this, in Section 5 we study a "fixed design" variant

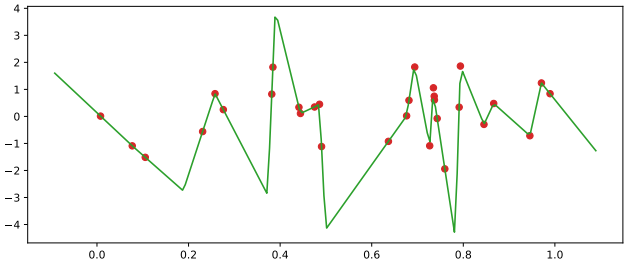

Figure 2: The min-norm interpolator for 30 random points with $f^* \equiv 0$ and $\mathcal{N}(0,1)$ label noise.

of the problem, where the training inputs lie on a uniform grid, $x_i = i/n$, and responses follow $y_i = f^*(x_i) + \epsilon_i$. In this case, interpolation is always tempered, with $\mathcal{L}_p(\hat{f}_S) \xrightarrow{n\to\infty} O(\mathcal{L}_p(f^*))$ for any constant $p \geq 1$ (Theorem 5 of Section 5).

Discussion and Takeaways. Our work is the first to study noisy interpolation learning with min-norm ReLU networks for regression. It is also the first to study noisy interpolation learning in neural networks where the input dimension does not grow with the sample size, and to consider non-linearly-interpolatable data distributions (see below for a comparison with concurrent work in a classification setting). The univariate case might seem simplistic, but is a rich and well-studied model in its own right (Shevchenko et al., 2022; Ergen & Pilanci, 2021; Hanin, 2021; Debarre et al., 2022; Boursier & Flammarion, 2023; Williams et al., 2019; Mulayoff et al., 2021; Safran et al., 2022), and as we see, it already exhibits many complexities and subtleties that need to be resolved, and is thus a non-trivial necessary first step if we want to proceed to the multivariate case.

The main takeaway from our work is that the transition from tempered to catastrophic overfitting can be much more subtle than previously discussed, both in terms of the details of the setting (e.g., sampled data vs. data on a grid) and in terms of the definition and notion of overfitting (the loss function used, and expectation vs. high probability). Understanding these subtleties is crucial before moving on to more complex models.

More concretely, we see that for the square loss, the behavior does not fit the "tempered overfitting" predictions of Mallinar et al. (2022), and for the L_1 loss we get a tempered behavior with high probability but not in expectation, which highlights that the definitions of (Mallinar et al., 2022) need to be refined. We would of course not get such strange behavior with the traditional non-overfitting approach of balancing training error and norm; in this situation the risk converges almost surely to the optimal risk, with finite expectation and vanishing variances. Moreover, perhaps surprisingly, when the input data is on the grid (equally spaced), the behavior is tempered for all losses even in the presence of label noise. This demonstrates that the catastrophic behavior for L_p losses for $p \geq 2$ is not just due to the presence of label noise; it is the combination of label noise and sampling of points that hurts generalization. We note that previous works considered benign overfitting with data on the grid as a simplified setting, which may help in understanding more general situations (Beaglehole et al., 2022; Lai et al., 2023). Our results imply that this simplification might change the behavior of the interpolator significantly. In summary, the nature of overfitting is a delicate property of the combination of how we measure the loss and how training examples are chosen.

Comparison with concurrent work. In a concurrent and independent work, Kornowski et al. (2023) studied interpolation learning in univariate two-layer ReLU networks in a classification setting, and showed that they exhibit tempered overfitting. In contrast to our regression setting, in classification only the output's sign affects generalization, and hence the height of the spikes do not play a significant role. As a result, our regression setting exhibits a fundamentally different behavior, and the above discussion on the delicateness of the overfitting behavior in regression does not apply to their classification setting.

2 Review: Min-Norm ReLU Networks

Minimum-norm unbounded-width univariate two-layer ReLU networks have been extensively studied in recent years, starting with Savarese et al. (2019), with the exact formulation equation 2 in-

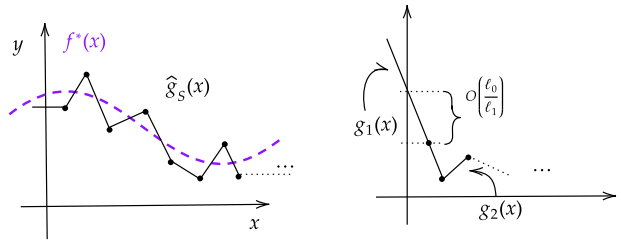

Figure 3: An illustration of the linear spline interpolator $\hat{g}_S$ (left), and of the variant $\hat{h}_S$ where linear pieces are extended beyond the endpoints (right).

corporating a skip connection due to Boursier & Flammarion (2023). Boursier & Flammarion, following prior work, establish that a minimum of equation 2 exists, with a finite number of units, and that it is also unique.

The problem in equation 2 is also equivalent to minimizing the "representation cost" $R(f) = \int_{\mathbb{R}} \sqrt{1+x^2}|f''(x)|dx$ over all interpolators f, although we will not use this characterization explicitly in our analysis. Compared to Savarese et al. (2019), where the representation cost is given by $\max\{\int |f''(x)|dx, |f'(-\infty) + f'(+\infty)|\}$, the weighting $\sqrt{1+x^2}$ is due to penalizing the biases b_i. More significantly, the skip connection in equation 1 avoids the "fallback" terms of $|f'(-\infty) + f'(+\infty)|$, which only kick-in in extreme cases (very few points or an extreme slope). This simplified the technical analysis and presentation, while rarely affecting the solution.

Boursier & Flammarion provide the following characterization of the minimizer[3] $\hat{f}_S$ of equation 2, which we will rely on heavily:

Lemma 2.1 (Boursier & Flammarion (2023))**.** *For $0 \le x_1 < x_2 < \cdots < x_n$, the problem in equation 2 admits a unique minimizer of the form:*

$$\hat{f}_S(x) = ax + b + \sum_{i=1}^{n-1} a_i(x - \tau_i)_+ \,, \tag{5}$$

where $\tau_i \in [x_i, x_{i+1})$ for every $i \in [n-1]$.

As in the above characterization, it is very convenient to take the training points to be sorted. Since the learned network $\hat{f}_S$ does not depend on the order of the points, we can always "sort" the points without changing anything. And so, throughout the paper, we will always take the points to be sorted (formally, the results apply to i.i.d. points, and the analysis is done after sorting these points).

3 Warm up: tempered overfitting in linear-spline interpolation

We start by analyzing tempered overfitting for linear-spline interpolation. Namely, we consider the piecewise-linear function obtained by connecting each pair of consecutive points in the dataset $S \sim \mathcal{D}^n$ (see Figures 1 and 3 left) and analyze its test performance.

Given a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, let $g_i : \mathbb{R} \to \mathbb{R}$ be the affine function joining the points (x_i, y_i) and (x_{i+1}, y_{i+1}). Thus, g_i is the straight line joining the endpoints of the i-th interval. Then, the linear spline interpolator $\hat{g}_S : [0,1] \to \mathbb{R}$ is given by

$$\hat{g}_S(x) := y_1 \cdot \mathbf{1}\{x < x_1\} + y_n \cdot \mathbf{1}\{x \ge x_n\} + \sum_{i=1}^{n-1} g_i(x) \cdot \mathbf{1}\{x \in [x_i, x_{i+1})\}. \tag{6}$$

[3] If the biases b_i are *not* included in the norm $\|\theta\|$ in equation 2, and this norm is replaced with $\sum_i (a_i^2 + w_i^2)$, the modified problem admits multiple non-unique minimizers, including a linear spline (with modified behavior past the extreme points) (Savarese et al., 2019). This set of minimizers was characterized by Hanin (2021). Interestingly, the minimizer $\hat{f}_S$ of equation 2 (when the biases are included in the norm) is also a minimizer of the modified problem (without including the biases). All our results apply also to the setting without penalizing the biases in the following sense: the upper bounds are valid for all minimizers, while some minimizer, namely $\hat{f}_S$ that we study, exhibits the lower bound behavior.

Note that in the intervals $[0, x_1]$ and $[x_n, 1]$ the linear-spline $\hat{g}_S$ is defined to be constants that correspond to labels y_1 and y_n respectively. The following theorem characterizes the asymptotic behavior of $\mathcal{L}_p(\hat{g}_S)$ for every $p \geq 1$:

Theorem 1. *Let f^* be any Lipschitz function and $\mathcal{D}$ be the distribution from equation 3. Let $S \sim \mathcal{D}^n$, and $\hat{g}_S$ be the linear-spline interpolator (equation 6) w.r.t. the dataset S. Then, for any $p \geq 1$ there is a constant C_p such that*

$$\lim_{n\to\infty} \mathop{\mathbb{P}}_{S}[\mathcal{R}_p(\hat{g}_S) \leq C_p\, \mathcal{L}_p(f^*)] = 1 \quad \textit{and} \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S}[\mathcal{L}_p(\hat{g}_S) \leq C_p\, \mathcal{L}_p(f^*)] = 1.$$

The theorem shows that the linear-spline interpolator exhibits tempered behavior, namely, w.h.p. over S the interpolator $\hat{g}_S$ performs like the predictor f^*, up to a constant factor. To understand why Theorem 1 holds, note that for all $i \in [n-1]$ and $x \in [x_i, x_{i+1}]$ the linear-spline interpolator satisfies $\hat{g}_S(x) \in [\min\{y_i, y_{i+1}\}, \max\{y_i, y_{i+1}\}]$. Moreover, we have for all $i \in [n]$ that $|y_i - f^*(x_i)| = |\epsilon_i|$, where ϵ_i is the random noise. Using these facts, it is not hard to bound the expected population loss of $\hat{g}_S$ in each interval $[x_i, x_{i+1}]$, and by using the law of large numbers it is also possible to bound the probability (over S) that the loss in the domain $[0, 1]$ is large. Thus, we can bound the L_p loss both in expectation and in probability.

Delicate behavior of linear splines. We now consider the following variant of the linear-spline interpolator:

$$\hat{h}_S(x) := g_1(x) \cdot \mathbf{1}\{x < x_1\} + g_{n-1}(x) \cdot \mathbf{1}\{x > x_n\} + \hat{g}_S(x) \cdot \mathbf{1}\{x \in [x_1, x_n]\}. \tag{7}$$

In words, $\hat{h}_S$ is exactly the same as $\hat{g}_S$ in the interval $[x_1, x_n]$, but it extends the linear pieces g_1 and g_{n-1} beyond the endpoints x_1 and x_n (respectively), as illustrated in Figure 3 (right). The interpolator $\hat{h}_S$ still exhibits tempered behavior in probability, similarly to $\hat{g}_S$. However, perhaps surprisingly, $\hat{h}_S$ is not tempered in expectation (see Appendix A for details). This delicate behavior of the linear-spline interpolator is important since in the next section we will show that the min-norm interpolator has a similar behavior to $\hat{h}_S$ in the intervals $[0, x_1], [x_n, 1]$, and as a consequence, it is tempered with high probability but not in expectation.

4 Min-norm interpolation with random data

In this section, we study the performance of the min-norm interpolator with random data. We first present some important properties of the min-norm interpolator in Section 4.1. In Sections 4.2 and 4.3 we use this characterization to study its performance.

4.1 Characterizing the min-norm interpolator

Our goal is to give a characterization of the min-norm interpolator $\hat{f}_S(x)$ (equation 5), in terms of linear splines as defined in equation 6. Recall the definition of affine functions $g_1(x), \ldots, g_{n-1}(x)$, which are piece-wise affine functions joining consecutive points. Let δ_i be the slope of the line $g_i(x)$, i.e. $\delta_i = g_i'(x)$. We denote $\delta_0 := \delta_1$ and $\delta_n := \delta_{n-1}$. Then, we can define the sign of the curvature of the linear spline $\hat{g}_S(x)$ at each point.

Definition 4.1. *For any $i \in [n]$,*

$$\mathsf{curv}(x_i) = \begin{cases} +1 & \delta_i > \delta_{i-1} \\ 0 & \delta_i = \delta_{i+1} \\ -1 & \delta_i < \delta_{i-1} \end{cases}$$

Based on the curvature, the following lemma geometrically characterizes $\hat{f}_S$ in any interval $[x_i, x_{i+1})$, in terms of the linear pieces g_{i-1}, g_i, g_{i+1}.

Lemma 4.2. *The function $\hat{f}_S$ can be characterized as follows:*

- $\hat{f}_S(x) = g_1(x)$ *for* $x \in (-\infty, x_2)$*;*
- $\hat{f}_S(x) = g_{n-1}(x)$ *for* $x \in [x_{n-1}, \infty)$*;*

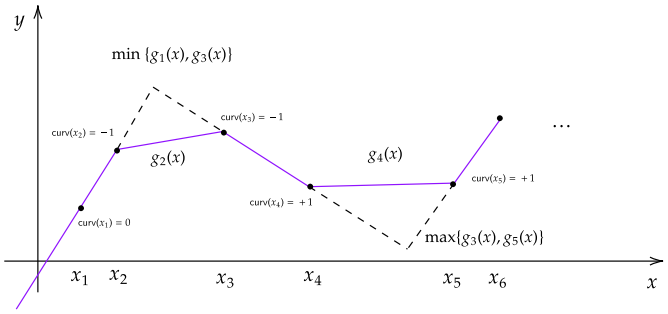

Figure 4: An illustration of the characterization of $\hat{f}_S$ from Lemma 4.2.

- *In each interval* $[x_i, x_{i+1})$ *for* $i \in \{2, \dots n-2\}$,
  1. *If* $\mathsf{curv}(x_i) = \mathsf{curv}(x_{i+1}) = +1$ *then*
  $$\max\{g_{i-1}(x), g_{i+1}(x)\} \le \hat{f}_S(x) \le g_i(x);$$
  2. *If* $\mathsf{curv}(x_i) = \mathsf{curv}(x_{i+1}) = -1$ *then*
  $$\min\{g_{i-1}(x), g_{i+1}(x)\} \ge \hat{f}_S(x) \ge g_i(x);$$
  3. *Else, i.e. either* $\mathsf{curv}(x_i) = 0$ *or* $\mathsf{curv}(x_{i+1}) = 0$ *or* $\mathsf{curv}(x_i) \ne \mathsf{curv}(x_{i+1})$,
  $$\hat{f}_S(x) = g_i(x).$$

The lemma implies that $\hat{f}_S$ coincides with $\hat{g}_S$ except in an interval $[x_i, x_{i+1})$ where the curvature of the two points are both $+1$ or -1 (see Figure 4). Intuitively, this property captures the worst-case effect of the spikes and will be crucial in showing the tempered behavior of $\hat{f}_S$ w.r.t. L_p for $p \in [1, 2)$. However, this still does not imply that such spikes are necessarily formed.

To this end, Boursier & Flammarion (2023, Lemma 8) characterized the situation under which indeed these spikes are formed. Roughly speaking, if the sign of the curvature changes twice within three points, then we get a spike. Formally, we identify special points from left to right recursively where the sign of the curvature changes.

Definition 4.3. *We define* $n_1 := 1$. *Having defined the location of the special points* $n_1, \dots, n_{i-1}$, *we recursively define*
$$n_i = \min\{j > n_{i-1} : \mathsf{curv}(x_j) \ne \mathsf{curv}(x_{n_i})\}.$$
If there is no such $n_{i-1} < j \le n$ *where* $\mathsf{curv}(x_j) \ne \mathsf{curv}(x_{n_i})$, *then* n_{i-1} *is the location of the last special point.*

Lemma 4.4 (Boursier & Flammarion (2023))**.** *For any* $k \ge 1$, *if* $\delta_{n_k-1} \ne \delta_{n_k}$ *and* $n_{k+1} = n_k + 2$, *then* $\hat{f}_S$ *has exactly one kink between* $(x_{n_k-1}, x_{n_{k+1}})$. *Moreover, if* $\mathsf{curv}(x_{n_k}) = \mathsf{curv}(x_{n_k+1}) = -1$ *then* $\hat{f}_S(x) = \min\{g_{n_k-1}(x), g_{n_k+1}(x)\}$ *in* $[x_{n_k}, x_{n_k+1})$.

This is a slight variation of (Boursier & Flammarion, 2023, Lemma 8), which we reprove in the appendix for completeness. See Figure 5 for an illustration of the above lemma. To show the catastrophic behavior of $\hat{f}_S$ for $p \ge 2$, we will consider events under which such configurations of points are formed. This will result in spikes giving catastrophic behavior.

4.2 Tempered overfitting for L_p with $p \in [1, 2)$

We now show the tempered behavior of the minimal norm interpolator w.r.t. L_p losses for $p \in [1, 2)$.

Theorem 2. *Let* f^* *be a Lipschitz function and* $\mathcal{D}$ *be the distribution from equation 3. Sample* $S \sim \mathcal{D}^n$, *and let* $\hat{f}_S$ *be the min-norm interpolator (equation 5). Then, for some universal constant* $C > 0$, *for any* $p \in [1, 2)$ *we have*
$$\lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{R}_p(\hat{f}_S) \le \frac{C}{2-p} \cdot \mathcal{L}_p(f^*)\right] = 1 \quad \textit{and} \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{L}_p(\hat{f}_S) \le \frac{C}{2-p} \cdot \mathcal{L}_p(f^*)\right] = 1\,.$$

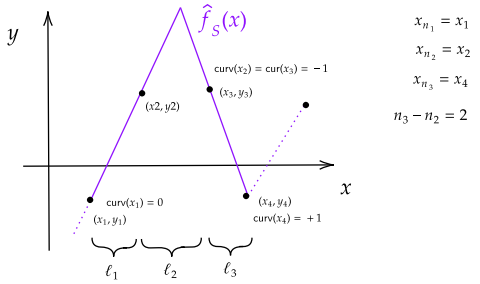

Figure 5: An illustration of the spike formed by Lemma 4.4. Here, x_2 and x_4 are two consecutive special points with exactly one point in between. There must be exactly one kink in (x_1, x_4). Thus, in $[x_2, x_3)$, the interpolator $\hat{f}_S$ must be $\min\{g_1(x), g_3(x)\}$.

The proof of Theorem 2 builds on Lemma 4.2, which implies that in an interval $[x_i, x_{i+1})$, a spike in the interpolator $\hat{f}_S$ must be bounded within the triangle obtained from g_{i-1}, g_i, g_{i+1} (see Figure 4). Analyzing the population loss of $\hat{f}_S$ requires considering the distribution of the spacings between data points. Let $\ell_0, \ldots, \ell_n$ be such that

$$\forall i \in [n-1] \;\; \ell_i = x_{i+1} - x_i, \quad \ell_0 = x_1, \quad \ell_n = 1 - x_n \; . \tag{8}$$

Prior works (Alagar, 1976; Pinelis, 2019) established that

$$(\ell_0, \ldots, \ell_n) \sim \left(\frac{X_0}{X}, \ldots, \frac{X_n}{X}\right), \text{ where } X_0, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(1), \text{ and } X := \sum_{i=0}^{n} X_i \; . \tag{9}$$

The slopes of the affine functions g_{i-1}, g_{i+1} are roughly $\frac{1}{\ell_{i-1}}, \frac{1}{\ell_{i+1}}$, where ℓ_j are the lengths as defined in equation 8. Hence, the spike's height is proportional to $\frac{\ell_i}{\max\{\ell_{i-1}, \ell_{i+1}\}}$. As a result, the L_p loss in the interval $[x_i, x_{i+1}]$ is roughly

$$\left(\frac{\ell_i}{\max\{\ell_{i-1}, \ell_{i+1}\}}\right)^p \cdot \ell_i = \frac{\ell_i^{p+1}}{\max\{\ell_{i-1}, \ell_{i+1}\}^p} \; .$$

Using the distribution of the ℓ_j's given in equation 9, we can bound the expectation of this expression. Then, similarly to our discussion on linear splines in Section 3, in the range $[x_1, x_n]$ we can bound the L_p loss both in expectation and in probability. In the intervals $[0, x_1]$ and $[x_n, 1]$, the expected loss is infinite (similarly to the interpolator $\hat{h}_S$ in equation 7), and therefore we have

$$\mathop{\mathbb{E}}_{S}\left[\mathcal{L}_p(\hat{f}_S)\right] = \infty \; . \tag{10}$$

Still, we can get a high probability upper bound for the L_p loss in the intervals $[0, x_1]$ and $[x_n, 1]$. Thus, we get a bound on L_p loss in the entire domain $[0, 1]$ w.h.p. We note that the definition of tempered overfitting in Mallinar et al. (2022) considers only the expectation. Theorem 2 and equation 10 imply that in our setting we have tempered behavior in probability but not in expectation, which demonstrates that tempered behavior is delicate.

We also show a lower bound for the population loss L_p which matches the upper bound from Theorem 2 (up to a constant factor independent of p). The lower bound holds already for $f^* \equiv 0$ and Gaussian label noise.

Theorem 3. *Let $f^* \equiv 0$, consider label noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some constant $\sigma > 0$, and let $\mathcal{D}$ be the corresponding distribution from equation 3. Let $S \sim \mathcal{D}^n$, and let $\hat{f}_S$ be the min-norm interpolator (equation 5). Then, for some universal constant $c > 0$, for any $p \in [1, 2)$ we have*

$$\lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{R}_p(\hat{f}_S) \geq \frac{c}{2-p} \cdot \mathcal{L}_p(f^*)\right] = 1 \quad \textit{and} \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{L}_p(\hat{f}_S) \geq \frac{c}{2-p} \cdot \mathcal{L}_p(f^*)\right] = 1 \; .$$

The proof of the above lower bound follows similar arguments to the proof of catastrophic overfitting for $p \geq 2$, which we will discuss in the next section.

4.3 CATASTROPHIC OVERFITTING FOR L_p WITH $p \geq 2$

Next, we prove that for the L_p loss with $p \geq 2$, the min-norm interpolator exhibits catastrophic overfitting. We prove this result already for $f^* \equiv 0$ and Gaussian label noise:

Theorem 4. *Let $f^* \equiv 0$, consider label noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some constant $\sigma > 0$, and let $\mathcal{D}$ be the corresponding distribution from equation 3. Let $S \sim \mathcal{D}^n$, and let $\hat{f}_S$ be the min-norm interpolator (equation 5). Then, for any $p \geq 2$ and $b > 0$,*

$$\lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{R}_p(\hat{f}_S) > b\right] = 1 \quad \textit{and} \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{L}_p(\hat{f}_S) > b\right] = 1\,.$$

To obtain some intuition on this phenomenon, consider the first four samples $(x_1, y_1), \ldots, (x_4, y_4)$, and let ℓ_i be the lengths of the intervals as defined in equation 8. We show that with constant probability, the configuration of the labels of these samples satisfies certain properties, which are illustrated in Figure 5. In this case, Lemma 4.4 implies that in the interval $[x_2, x_3]$ the interpolator $\hat{f}_S$ is equal to $\min\{g_1(x), g_3(x)\}$, where g_1 (respectively, g_3) is the affine function that connects x_1, x_2 (respectively, x_3, x_4). Now, as can be seen in the figure, in this "unfortunate configuration" the interpolator $\hat{f}_S$ spikes above $f^* \equiv 0$ in the interval $[x_2, x_3]$, and the spike's height is proportional to $\frac{\ell_2}{\max\{\ell_1,\ell_3\}}$. As a result, the L_p loss in the interval $[x_2, x_3]$ is roughly $\frac{\ell_2^{p+1}}{\max\{\ell_1,\ell_3\}^p}$. Using equation 9, we can show that $\mathbb{E}_S\left[\frac{\ell_2^{p+1}}{\max\{\ell_1,\ell_3\}^p}\right] = \infty$ for any $p \geq 2$.

We then divide divide the n samples in S into $\Theta(n)$ disjoint subsets and consider the events that labels are such that the 4 middle points exhibit an "unfortunate configuration" as described above. Using the fact that we have $\Theta(n)$ such subsets and the losses in these subsets are only mildly correlated, we are able to prove that $\hat{f}_S$ exhibits a catastrophic behavior also in probability.

We note that the proof of Theorem 3 follows similar arguments, except that when $p < 2$ the expectation of the L_p loss in each subset with an "unfortunate configuration" is finite, and hence we get a finite lower bound.

5 MIN-NORM INTERPOLATION WITH SAMPLES ON THE GRID

In this section, we analyze the population loss of the min-norm interpolator, when the n data-points in S are uniformly spaced, instead of i.i.d. uniform sampling considered in the previous sections. Namely, consider the training set $S = \{(x_i, y_i) : i \in [n]\}$, where

$$x_i = \frac{i}{n} \;\text{ and }\; y_i = f^*(x_i) + \epsilon_i \;\text{ for i.i.d. noise }\; \epsilon_i\,. \tag{11}$$

Note that the randomness in S is only in the label noises ϵ_i. It can be interpreted as a *non-adaptive active learning* setting, where the learner can actively choose the training points, and then observe noisy measurements at these points, and the query points are selected on an equally spaced grid. We show that in this situation the min-norm interpolator exhibits tempered overfitting with respect to any L_p loss:

Theorem 5. *Let f^* be any Lipschitz function. For the size-n dataset S given by equation 11, let $\hat{f}_S$ be the min-norm interpolator (equation 5). Then for any $p \geq 1$, there is a constant C_p such that*

$$\lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{R}_p(\hat{f}_S) \leq C_p\,\mathcal{L}_p(f^*)\right] = 1 \quad \textit{and} \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S}\left[\mathcal{L}_p(\hat{f}_S) \leq C_p\,\mathcal{L}_p(f^*)\right] = 1\,.$$

An intuitive explanation is as follows. Since the points are uniformly spaced, whenever spikes are formed, they can at most reach double the height without the spikes. Thus, the population loss of $\hat{f}_S(x)$ becomes worse but only by a constant factor. We remark that in this setting the min-norm interpolator exhibits tempered overfitting both in probability (as stated in Theorem 5) and in expectation. From Theorem 5 we conclude that the catastrophic behavior for L_p with $p \geq 2$ shown in Theorem 4 stems from the non-uniformity in the lengths of the intervals $[x_i, x_{i+1}]$, which occurs when the x_i's are drawn at random.

ACKNOWLEDGEMENTS

This research was done as part of the NSF-Simons Sponsored Collaboration on the Theoretical Foundations of Deep Learning. N. J. would like to thank Surya Pratap Singh for his generous time in helping resolve Python errors.

REFERENCES

Vangalur S Alagar. The distribution of the distance between random points. *Journal of Applied Probability*, 13(3):558–566, 1976.

Peter L Bartlett and Philip M Long. Failures of model-dependent generalization bounds for least-norm interpolation. *The Journal of Machine Learning Research*, 22(1):9297–9311, 2021.

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.

Daniel Beaglehole, Mikhail Belkin, and Parthe Pandit. Kernel ridgeless regression is inconsistent for low dimensions. *arXiv preprint arXiv:2205.13525*, 2022.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. *arXiv preprint arXiv:2303.01353*, 2023.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

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.

Geoffrey Chinot and Matthieu Lerasle. On the robustness of the minimum ℓ_2 interpolator. *arXiv preprint arXiv:2003.05838*, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory (COLT)*, 2020.

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

Thomas Debarre, Quentin Denoyelle, Michael Unser, and Julien Fageot. Sparsest piecewise-linear regression of one-dimensional data. *Journal of Computational and Applied Mathematics*, 406: 114044, 2022.

W Edwards Deming and Clarence G Colcord. The minimum in the gamma function. *Nature*, 135 (3422):917–917, 1935.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning (ICML)*, 2022.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *Journal of machine learning research*, 2021.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory (COLT)*, 2022.

Spencer Frei, Gal Vardi, Peter L Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. *arXiv preprint arXiv:2303.01462*, 2023.

Nikhil Ghosh and Mikhail Belkin. A universal trade-off between the model size, test loss, and training loss of linear predictors. *arXiv preprint arXiv:2207.11621*, 2022.

Boris Hanin. Ridgeless interpolation with shallow relu networks in $1d$ is nearest neighbor curvature extrapolation and provably generalizes on lipschitz functions. *arXiv preprint arXiv:2109.12960*, 2021.

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

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Preprint, arXiv:1903.08560*, 2020.

Peizhong Ju, Xiaojun Lin, and Jia Liu. Overfitting can be harmless for basis pursuit, but only to a degree. *Advances in Neural Information Processing Systems*, 33:7956–7967, 2020.

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting. *arXiv preprint arXiv:2106.09276*, 2021.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in relu neural networks. *arXiv preprint arXiv:2305.15141*, 2023.

Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting for two-layer relu networks. *arXiv preprint arXiv:2303.04145*, 2023.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$. *arXiv preprint arXiv:2302.05933*, 2023.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

Naren Sarayu Manoj and Nathan Srebro. Interpolation learning with minimum description length. *arXiv preprint arXiv:2302.07263*, 2023.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

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

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761, 2021.

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.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pp. 7263–7272, 2020.

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

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Iosif Pinelis. Order statistics on the spacings between order statistics for the uniform distribution. *arXiv preprint arXiv:1909.06406*, 2019.

Itay Safran, Gal Vardi, and Jason D Lee. On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pp. 2667–2690. PMLR, 2019.

Ohad Shamir. The implicit bias of benign overfitting. In *Conference on Learning Theory*, pp. 448–478. PMLR, 2022.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide relu networks. *The Journal of Machine Learning Research*, 23(1): 5660–5714, 2022.

Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Advances in Neural Information Processing Systems*, 33:8907–8920, 2020.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Preprint, arXiv:2009.14286*, 2020.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. *Preprint, arXiv:2011.09148*, 2021.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. In *Advances in Neural Information Processing Systems*, pp. 8378–8387, 2019.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J Sutherland, and Nathan Srebro. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. *arXiv preprint arXiv:2210.12082*, 2022.