DO HIGHLY OVER-PARAMETERIZED NEURAL NETWORKS
GENERALIZE SINCE BAD SOLUTIONS ARE RARE?

A PREPRINT

Julius Martinetz
Machine Learning Group
Technical University Berlin
Berlin, Germany
j.martinetz@tu-berlin.de

Thomas Martinetz
Institute for Neuro- and Biocomputing
University of Lübeck
Lübeck, Germany
martinetz@inb.uni-luebeck.de

December 5, 2023

ABSTRACT

We study over-parameterized classifiers where Empirical Risk Minimization (ERM) for learning leads to zero training error. In these over-parameterized settings there are many global minima with zero training error, some of which generalize better than others. We show that under certain conditions the fraction of "bad" global minima with a true error larger than $\epsilon$ decays to zero exponentially fast with the number of training data $n$. The bound depends on the distribution of the true error over the set of classifier functions used for the given classification problem, and does not necessarily depend on the size or complexity (e.g. the number of parameters) of the classifier function set. This insight may provide a novel perspective on the unexpectedly good generalization even of highly over-parameterized neural networks. We substantiate our theoretical findings through experiments on synthetic data and a subset of MNIST. Additionally, we assess our hypothesis using VGG19 and ResNet18 on a subset of Caltech101.

1 Introduction

Extreme upscaling and deepening of neural networks has led to a quantum leap in many real-world object recognition tasks [1]. The same is true for neural networks in Natural Language Processing [2] and reinforcement learning [3]. By enlarging the networks appropriately, the level of performance increases, and even entering regimes with many more network parameters than training samples does not harm generalization. This is usually attributed to an appropriate regularization. However, generalization is also observed in highly over-parameterized regimes and without explicit regularization, where even random label assignments or even random image data can be memorized [4]. Belkin called this the "modern" interpolation regime' [5], although this effect has been studied for some time [6].

The generalization capabilities of large neural networks do not seem to be harmed by the millions of parameters in today’s popular architectures. Due to their large, inherent capacity the training error is zero after training by Empirical Risk Minimization (ERM), i.e. the training data is memorized. Nevertheless, the test error can be unexpectedly low [4], even after only a few handful of training samples [7]. This is in contrast to traditional machine learning wisdom, where one would not expect any generalization in these highly over-parameterized regimes, where the data is highly over-fitted. Statistical learning theory based on uniform generalization bounds requires a high probability that there is no solution with a large deviation between empirical and expected risk, i.e. the training and the true error. However, this is not the case in over-parameterized regimes. There is no uniform generalization bound. In these highly over-parameterized regimes always also "bad" solutions with zero training but large true error do exist. Interestingly, they hardly seem to occur in practice.

In this paper, we investigate over-parameterized classifiers with training error zero solutions. A classifier receives inputs $x$ from an input distribution $P(x)$ and assigns a class label using a classifier function $h(x)$. The input $x$ has a true label $y$ with probability $P(y|x)$. For each input $x$, the classifier produces a loss $L(y, h(x)) \in \{0, 1\}$, 0 if the classification is correct and 1 if the classification is incorrect. The error of the classifier $h$ on the whole data distribution
$P(x, y) = P(y|x)P(x)$ is given by the expected risk

$$E(h) = \int L(y, h(x))P(x, y)\, dx\, dy.$$ 

$E(h)$ is also called true error and always lies between 0 and 1. The classifier or predictor $h$ is usually chosen from a function set $\mathcal{H}$. For example, $\mathcal{H}$ is determined by the architecture of a neural network, and the allocated $h$ is determined by the network parameters. Learning means selecting an $h$ from the set $\mathcal{H}$ so as to minimize $E(h)$. This is often done by Empirical Risk Minimization (ERM). Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a so-called training set of $n$ input samples $x$ together with their labels $y$ which are drawn i.i.d. from $P(x, y)$. The empirical risk is defined as the average loss on this training set

$$E_S(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i))$$

and is also called training error. Learning via ERM chooses an $h \in \mathcal{H}$ that minimizes $E_S(h)$. However, $E(h)$ and $E_S(h)$ may deviate. This deviation is called the generalization gap.

If the generalization gap is small, a small training error provides a good solution with a small true error. Statistical learning theory based on VC-dimension [8] or Rademacher complexity [9] provides uniform bounds for the generalization gap. The generalization gap becomes small with increasing $n$ for each $h \in \mathcal{H}$. The objective is to completely rule out "bad" solutions with large generalization gaps, at least with high probability. However, these bounds do not exist in highly over-parameterized regimes. There are always $h \in \mathcal{H}$ left with large generalization gaps. Nevertheless, in practice learning takes place and ERM provides good solutions that generalize well also in these over-parameterized regimes. There have been theoretical and empirical attempts to understand this "mystery". It has been argued, that this might be due to implicit regularization of (stochastic) gradient descent [10], [11], [12], [13], [14], [15]. Further, a number of novel algorithmic-dependent uniform generalization bounds based on norm or compression measures have been supposed for explanation [16], [17], [18], [19], [20], [21], [22], [23], [24]. The Neural Tangent Kernel (NTK) with its linearization at initialization gives further insights for very wide neural networks [25], [26], [27], [28], [29], [30], [31]. Another approach applies the concepts of algorithmic stability [32], [33], [34], [35]. But still there is empirically based scepticism that these bounds already help [36] or that uniform convergence bounds are fundamentally the right approach [37].

As "bad" solutions cannot be ruled out in over-parameterized regimes, we conjecture that within the set of ERM solutions, the proportion of "bad" classifiers nevertheless remains small. This hypotheses offers a potential alternative explanation for the enigma of good generalization in over-parameterized regimes and corresponds to and supports the empirical studies in [38]. If the fraction of "bad" classifiers is small among the global minima with zero training error, the ERM algorithm is more likely to converge to a favorable solution rather than an unfavorable one. This perspective introduces a novel angle, challenging the conventional wisdom in machine learning that typically assumes the prevalence of "bad" solutions in highly over-parameterized regimes and corresponds to and supports the empirical studies in [38].

2 Preliminaries, Notations and Definitions

We assume $\mathcal{H} = \mathcal{H}_W$ to be parameterized on a compact set $W \subseteq \mathbb{R}^N$ with non-zero Lebesgue measure, and each $w \in W$ determines a classifier function $h_w \in \mathcal{H}_W$. If a neural network is used as classifier, then $w$ are the weights of the neural network. For each $w$ we obtain a true error $E(h_w) = E(w)$. We assume $E(w)$ to be integrable on $W$ and $\mathcal{H}_W$ to have a large but finite VC-dimension. $E_{\min} = \min_{w \in W} E(w)$ denotes the minimum true error of the classifier function set $\mathcal{H}_W$ on the given classification problem. We work with the following subsets of $W$:

\[
\begin{align*}
\mathcal{W}_\varepsilon & : \text{ set of all } w \in W \text{ with } E(w) \geq E_{\min} + \varepsilon \quad (\varepsilon \geq 0) \\
\mathcal{W}(S) & : \text{ set of all } w \in W \text{ with } E_S(w) = 0 \\
\mathcal{W}_\varepsilon(S) & : \text{ set of all } w \in W \text{ with } E_S(w) = 0
\end{align*}
\]

$\mathcal{W}(S)$ is the set of all global minima on the training error $E_S(w)$ for a given training set $S$. Since we are interested in the over-parameterized regime, i.e. the training set size $n = |S|$ is small compared to the complexity of the classifier, we assume $\mathcal{W}(S)$ to be of non-zero measure for almost all $S$ with $|S| = n$. We can take it as a definition for over-parameterization. Hence, the ERM algorithm ends up in one of these global minima. $\mathcal{W}_\varepsilon(S)$ is what we call the set of "bad" global minima with true errors larger than $E_{\min} + \varepsilon$. We have $\mathcal{W}_\varepsilon(S) \subseteq \mathcal{W}(S) \subseteq W$. Further, we use the

1Convergence to global minima of ERM algorithms such as gradient descent or its variants is a topic on its own rights. However, in highly over-parameterized settings convergence to zero training error is usually not a problem.
Do highly over-parameterized neural networks generalize since bad solutions are rare?

Figure 1: The density of classifiers (DOC) for a neural network from our experiments (A) and a binary classification problem with random labels (B). In (A) the DOC is shown as a histogram of the true error for one million randomly chosen network weights. \( g_\varepsilon \) gives the fraction of weights with true errors lower than 0.2. In a binary classification problem with random labels (B), for each classifier \( w \) the true error \( E(w) \) is always 0.5 and, hence, the DOC is a peak at \( E = 0.5 \).

notations \( \Omega = |W|, \Omega_\varepsilon = |W_\varepsilon|, \omega(S) = |W(S)| \) and \( \omega_\varepsilon(S) = |W_\varepsilon(S)| \) for the size (volume) of these parameter sets. In the following the fractions

\[
\begin{align*}
g_\varepsilon &= \frac{\Omega - \Omega_\varepsilon}{\Omega} : \text{fraction of "good" classifiers within } W \\
\phi_\varepsilon(S) &= \frac{\omega_\varepsilon(S)}{\omega(S)} : \text{fraction of "bad" classifiers within the set of global minima } W(S)
\end{align*}
\]

are important. \( g_\varepsilon \) as the fraction of "good" classifiers within \( W \) is a measure for the "appropriateness" of \( H \) for the given classification problem. It is a measure for the bias of the classifier function set towards the given classification problem. \( \phi_\varepsilon(S) \) is the main property we are looking at. If \( \phi_\varepsilon(S) \) is small, the ERM algorithm should more likely end up in a good solution. We show that under certain conditions \( \phi_\varepsilon(S) \) is very likely to become small with increasing \( n \) and relate it to \( g_\varepsilon \).

2.1 Density of Classifiers (DOC)

For quantifying the parameter volume of good and bad classifiers we introduce the "density of classifiers" \( D(E) \) at true error \( E \) defined such that

\[
\Omega_\varepsilon = \int_{E_{\min}+\varepsilon}^{1} D(E) \, dE \quad \text{for each} \quad 0 \leq \varepsilon \leq 1 - E_{\min}.
\]

\( D(E)\,dE \) is the volume of parameters \( w \in W \) with \( E \leq E(w) \leq E + dE \). For \( 0 \leq E < E_{\min} \) we set \( D(E) = 0 \).

In Fig. 1 we illustrate the density of classifiers (DOC) for two scenarios. On the left (A) we see the DOC of a neural network from our experiment in Section 4.1. For one million network weights \( w \in W \) uniformly chosen from \( W \) their corresponding true error \( E(w) \) is determined (details in Section 4.1). The DOC in A shows the distribution of the true errors. The part left of the red line illustrates \( g_\varepsilon = \int_{E_{\min}+\varepsilon}^{E_{\min}} D(E) \, dE/\Omega \), the fraction of good classifiers with a true error smaller than \( E_{\min} + \varepsilon \). On the right (B) we show the DOC in case of a binary classification problem with random class labels (each class with equal probability). For each \( w \), the true error is \( E(w) = 0.5 \) and, hence, the DOC is a peak at \( E = 0.5 \).

3 Results

For \( S \) randomly drawn from \( P(x, y)^n \), the parameter set size \( \omega_\varepsilon(S) \) of "bad" global minima is a random variable with \( 0 \leq \omega_\varepsilon(S) \leq \Omega \). The first and main Theorem, for which the proof is given in the Appendix, looks at its mean value:
Theorem 1. For each $0 \leq \varepsilon \leq 1 - E_{\min}$

$$
\langle \omega_\varepsilon(S) \rangle_S = \int_{\mathcal{W}_\varepsilon} (1 - E(w))^n \, dw
$$

$$
= \int_{E_{\min} + \varepsilon}^1 (1 - E)^n D(E) \, dE
$$

$$
\leq \Omega_\varepsilon (1 - (E_{\min} + \varepsilon))^n.
$$

Theorem 1 tells us, that the mean parameter set size (parameter volume) of "bad" global minima decreases to zero exponentially fast with increasing training set size $n$. However, this might not suffice for $\phi_\varepsilon(S)$ to become small, since the mean parameter set size volume of ALL global minima, $\langle \omega(S) \rangle_S$, decreases exponentially fast as well. $\langle \omega(S) \rangle_S$ is given by (1) for $\varepsilon = 0$. Either it converges to zero, e.g. if $E_{\min} > 0$, or to the mean set size of solutions with true error zero in case $E_{\min} = 0$. If the classification problem is separable, this set size can be of non-zero measure.

However, the mean parameter set size of ALL global minima converges much slower than the mean set size of "bad" global minima and, hence, the mean set size of "bad" global minima will become very small relative to the mean set size of all global minima. The following corollary quantifies it:

Corollary 1. For each $0 \leq \varepsilon < 1 - E_{\min}$

$$
\frac{\langle \omega_\varepsilon(S) \rangle_S}{\langle \omega(S) \rangle_S} = \frac{\int_{E_{\min} + \varepsilon}^1 (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE}
$$

$$
\leq \frac{\Omega_\varepsilon}{\Omega (1 - g_{\varepsilon/2}) + g_{\varepsilon/2} e^{\varepsilon n}}
$$

$$
\leq \frac{1}{g_{\varepsilon/2}} e^{-\varepsilon n} \quad \text{if} \quad g_{\varepsilon/2} > 0.
$$

The mean set size of "bad" global minima relative to mean set size of all global minima diminishes at least exponentially fast with $n$. Note, that the prefactor in (1) is determined by the inverse of the fraction of "good" classifiers within $\mathcal{W}$ (better than $\varepsilon/2$). The larger the fraction of good classifiers to begin with, i.e., the better the bias of the classifier set, the smaller the prefactor. It does not depend on the size of $\mathcal{W}$ and does not increase necessarily with $\Omega$, e.g., with the number of parameters. We will demonstrate this later in our experiments. The exponent is linear in $\varepsilon$.

Corollary 1 tells us, that the mean set size of "bad" global minima is small relative to the mean set size of all global minima. In a concrete setting with a given training set $S$, $\phi_\varepsilon(S) = \omega_\varepsilon(S)/\omega(S)$, i.e., the given fraction of "bad" global minima should be small relative to the given fraction of all global minima. In general

$$
\langle \phi_\varepsilon(S) \rangle_S \neq \frac{\langle \omega_\varepsilon(S) \rangle_S}{\langle \omega(S) \rangle_S},
$$

but with the following corollary we get a condition for the l.h.s. to be smaller than the r.h.s., which would be sufficient:

Corollary 2. If and only if $\phi_\varepsilon(S)$ and $\omega(S)$ do not correlate negatively, then

$$
\langle \phi_\varepsilon(S) \rangle_S \leq \frac{\langle \omega_\varepsilon(S) \rangle_S}{\langle \omega(S) \rangle_S}.
$$

$$
\leq \frac{\int_{E_{\min} + \varepsilon}^1 (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE}.
$$

Equality is given if and only if $\phi_\varepsilon(S)$ and $\omega(S)$ do not correlate.

Corollary 2 directly follows from its condition, since then

$$
\langle \phi_\varepsilon(S) \omega(S) \rangle_S - \langle \phi_\varepsilon(S) \rangle_S \langle \omega(S) \rangle_S \geq 0,
$$

and from the definition of $\phi_\varepsilon(S)$ we have $\phi_\varepsilon(S)\omega(S) = \omega_\varepsilon(S)$.

With the Markov-inequality the probability, that $\phi_\varepsilon(S)$ is larger than a $\gamma > 0$, is given by

$$
Prob \{ \phi_\varepsilon(S) \geq \gamma \} \leq \frac{1}{\gamma} \frac{\int_{E_{\min} + \varepsilon}^1 (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE} \leq \frac{1}{\gamma g_{\varepsilon/2}} e^{-\varepsilon n}.
$$

The probability, that the fraction $\phi_\varepsilon(S)$ of "bad" ERM solutions within the set of all ERM solutions is not smaller (smaller than $\gamma$) for a given $S$, decreases exponentially fast to zero with the number of training data $n$. "Bad" solutions become rare with high probability. It is not the absolute number (density) of classifiers, which determines the r.h.s., but only the shape of $D(E)$, i.e., the normalized $D(E)$. The shape of $D(E)$ does not necessarily depend on the number of parameters of our classifier functions.
3.1 Mean true error

In the following we derive an expression for the mean true error $E_n$ over all global minima $\mathcal{W}_S$ and all $S$. For a given $S$ we define $Q_S(E)$ such that

$$\phi_\varepsilon(S) = \frac{\omega_\varepsilon(S)}{\omega(S)} = \int_{E_{\min}+\varepsilon}^{1} Q_S(E) \, dE \quad \text{for each} \quad 0 \leq \varepsilon \leq 1 - E_{\min}. \quad (6)$$

$Q_S(E)$ is the normalized density of global minima at true error $E$ for the given $S$. One will rather end up with a classifier at true errors where the density $Q_S(E)$ is large. The mean true error $E_S$ over all global minima of a given $S$ is

$$E_S = \int_0^1 E \, Q_S(E) \, dE.$$

With $Q_n(E) = \langle Q_S(E) \rangle_S$ we obtain from (6)

$$\langle \phi_\varepsilon(S) \rangle_S = \int_{E_{\min}+\varepsilon}^{1} Q_n(E) \, dE \quad \text{for each} \quad 0 \leq \varepsilon \leq 1 - E_{\min}, \quad (7)$$

and the mean true error $E_n$ over all global minima $\mathcal{W}_S$ and all $S$ is accordingly given by

$$E_n = \int_0^1 E \, Q_n(E) \, dE. \quad (8)$$

If (and only if) Corollary 2 with equality in (5) is valid for each $0 \leq \varepsilon \leq 1 - E_{\min}$, then comparing (5) and (7) gives

$$Q_n(E) = \frac{(1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE},$$

and with (8) we obtain for the mean true error

$$E_n = \int_0^1 \frac{E(1 - E)^n D(E)}{\int_0^1 (1 - E)^n D(E) \, dE} \, dE. \quad (9)$$

In this case the mean true error over all ERM solutions is exactly determined by the density of classifiers $D(E)$. Again, it is not the absolute number (density) of classifiers, which determines $E_n$, but only the shape of $D(E)$.

In case that equality in (5) is not valid, we have to extend Corollary 2. For each $\varepsilon, \varepsilon'$ with $0 \leq \varepsilon < \varepsilon' \leq 1 - E_{\min}$ we denote by $\mathcal{W}_{\varepsilon < \varepsilon'}(S)$ the set of all $w \in \mathcal{W}(S)$ with $E_{\min} + \varepsilon \leq E(w) \leq E_{\min} + \varepsilon'$ and with $\omega_{\varepsilon < \varepsilon'}(S)$ its size. Since $\omega_{\varepsilon < \varepsilon'}(S) = \omega_\varepsilon(S) - \omega_{\varepsilon'}(S)$ and with (2) we obtain

$$\langle \omega_{\varepsilon < \varepsilon'}(S) \rangle_S = \frac{\int_{E_{\min}+\varepsilon}^{E_{\min}+\varepsilon'} (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE}.$$

With $\phi_{\varepsilon < \varepsilon'}(S) = \omega_{\varepsilon < \varepsilon'}(S) / \omega(S)$ and analog to Corollary 2, if and only if $\phi_{\varepsilon < \varepsilon'}(S)$ and $\omega(S)$ do not correlate negatively for each $0 \leq \varepsilon < \varepsilon' \leq 1 - E_{\min}$, then

$$\langle \phi_{\varepsilon < \varepsilon'}(S) \rangle_S \leq \frac{\int_{E_{\min}+\varepsilon}^{E_{\min}+\varepsilon'} (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE} \quad \text{for each} \quad 0 \leq \varepsilon < \varepsilon' \leq 1 - E_{\min}. \quad (10)$$

But then

$$Q_n(E) \leq \frac{(1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE}$$

and with (9)

$$E_n \leq \int_0^1 \frac{E(1 - E)^n D(E)}{\int_0^1 (1 - E)^n D(E) \, dE} \, dE. \quad (10)$$

Interestingly, in our experiments in the next section $E_n$ is not only bounded by (10) but in most cases exactly determined by (9).
Do highly over-parameterized neural networks generalize since bad solutions are rare?

A P

Figure 2: Illustration of the structure of the distribution $P(\omega, \omega_\varepsilon)$ of the two random variables $\omega$ and $\omega_\varepsilon$. For each $S$ the corresponding pair $(\omega, \omega_\varepsilon)$ always lies in the grey shaded area. $\bar{\omega}_\varepsilon(\omega)$ goes to zero for $\omega \to 0$ and to $\Omega_\varepsilon$ for $\omega \to \Omega$. $\bar{\omega}_\varepsilon(\omega)$ never lies strictly above the straight line from the origin to $(\Omega, \Omega_\varepsilon)$, hence, it is never concave. If it is convex, than Corollary 2 is valid.

3.2 Discussion of results

The central question is: Under what conditions does Corollary 2 hold, meaning when do $\bar{\phi}_\varepsilon(S)$ and $\omega(S)$ not correlate negatively? The proportion of "bad" global minima, $\bar{\phi}_\varepsilon(S)$, should not have a tendency of becoming smaller for larger global minima set sizes $\omega(S)$.

We take a look at the random variables $0 < \omega(S) \leq \Omega$ and $0 \leq \omega_\varepsilon(S) \leq \Omega_\varepsilon$ generated by random drawings of $S$ and their joint probability distribution $P(\omega, \omega_\varepsilon)$. With $P(\omega, \omega_\varepsilon) = P(\omega_\varepsilon | \omega) P(\omega)$ we can define $\bar{\phi}_\varepsilon(\omega)$ and $\bar{\omega}_\varepsilon(\omega)$ according to

$$\bar{\phi}_\varepsilon(\omega) = \int_0^{\Omega_\varepsilon} \frac{\omega_\varepsilon}{\omega} P(\omega_\varepsilon | \omega) d\omega_\varepsilon$$

$$= \frac{1}{\omega} \int_0^{\Omega_\varepsilon} \omega_\varepsilon P(\omega_\varepsilon | \omega) d\omega_\varepsilon$$

$$= \bar{\omega}_\varepsilon(\omega).$$

If $\bar{\phi}_\varepsilon(\omega)$ is monotonically non-decreasing with $\omega$, then

$$\langle \phi_\varepsilon(S) \omega(S) \rangle_S - \langle \bar{\phi}_\varepsilon(S) \rangle_S \langle \omega(S) \rangle_S = \langle \phi_\varepsilon \omega \rangle_{\omega, \omega_\varepsilon} - \langle \bar{\phi}_\varepsilon \rangle_{\omega, \omega_\varepsilon}$$

$$= \langle \bar{\phi}_\varepsilon(\omega) \omega \rangle_* - \langle \bar{\phi}_\varepsilon(\omega) \rangle_* \omega_*$$

$$\geq 0,$$

which is sufficient for Corollary 2 to be valid.

When is this the case? In Fig. 2 we illustrate some scenarios. With a random $S$, the corresponding pair $(\omega, \omega_\varepsilon)$ always falls into the grey area, since always $\omega_\varepsilon \leq \Omega_\varepsilon$ and $\omega \leq \Omega$. For $\omega \to 0$, $\bar{\omega}_\varepsilon(\omega) \to 0$, and for $\omega \to \Omega$, $\bar{\omega}_\varepsilon(\omega) \to \Omega_\varepsilon$. A simple course of $\bar{\omega}_\varepsilon(\omega)$ between these two points would be convex or concave, in the sense that the first derivative is monotone. However, the case that $\omega_\varepsilon(\omega)$ is strictly above the line given by $\omega_\varepsilon = (\Omega_\varepsilon / \Omega) \omega$, which includes being
strictly concave, can be ruled out, since otherwise
\[
\langle \omega_\varepsilon(S) \rangle_S = \frac{\langle \omega_\varepsilon \rangle_{\omega, \omega}}{\langle \omega \rangle_{\omega, \omega}} = \frac{1}{\langle \omega \rangle_{\omega}} \int_0^\Omega \tilde{\omega}_\varepsilon(\omega) P(\omega) d\omega > \frac{1}{\langle \omega \rangle_{\omega}} \int_0^\Omega \frac{\Omega_\varepsilon}{\Omega} \omega P(\omega) d\omega > \frac{\Omega_\varepsilon}{\Omega} \omega
\]
which contradicts \(^3\) of Corollary 1. But if \(\tilde{\omega}_\varepsilon(\omega)\) is convex, then per definition of convexity
\[
\frac{\tilde{\omega}_\varepsilon(\omega)}{\omega} = \phi_\varepsilon(\omega)
\]
is monotonically non-decreasing in \(\omega\) and, hence, Corollary 2 is valid.

Thus, the trajectory of \(\tilde{\omega}_\varepsilon(\omega)\) must at least exhibit a complexity beyond that of a simple convex shape for Corollary 2 to be rendered invalid. While a higher complexity is a plausible scenario, it is not yet sufficient and raises intriguing research questions. Our experiments in the subsequent section consistently affirm the validity of Corollary 2.

4 Experiments

To experimentally verify our theoretical results, we study experimental setups in which we can determine \(D(E), \omega(S)\) and \(\phi_\varepsilon(S)\) by random sampling. We take multi-layer-perceptrons with leaky ReLUs (10% leakiness) and no bias. The output of the classifier is then given by
\[
y = \text{ReLU}(W_K \text{ReLU}(W_{K-1}(... \text{ReLU}(W_1 x))))
\]
with \(W_k, k = 1, ..., K\) as the weight matrix of layer \(k, x \in \mathbb{R}^D\) as the input vector of dimension \(D\) and \(y \in \mathbb{R}^2\) as the output vector, in our case for binary classification. The ReLU of the output layer with the largest output determines the predicted class membership for the given input \(x\). In this setting, it is easy to see and well known that the class membership is invariant to the lengths of the weight vectors due to the positive homogeneity of the leaky ReLUs (for \(s \geq 0, \text{ReLU}(sx) = s\text{ReLU}(x)\)). Thus, each classifier function is determined by a weight vector (comprising all weights of the network) of unit length. \(\mathcal{W}\) is given by the unit sphere in dimension \(\mathbb{R}^N\), with \(N\) as the number of weights of the network. In our experiments, we will sample uniformly from this unit sphere. With the leakiness of the ReLUs, we avoid too many trivial solutions with many "dead" ReLUs.

4.1 Synthetic data set

For the first experiment we take a synthetic data set of two isotropic Gaussians in 10 dimensions. The Gaussian of class A is centered at \((+1, 0, ..., 0)\) and of class B at \((-1, 0, ..., 0)\). Both Gaussians have a variance of 0.5 and, hence, overlap with 2.28\%. This is the minimal possible classification error \(E_{\text{min}}\). We apply three different networks. The smallest has one hidden layer with ten hidden units, which leads to 120 weights. The next has again one hidden layer, but 100 hidden units, hence, with 1, 200 ten times as many weights. The third network has ten hidden layers with ten units each, hence, 1, 020 weights.

The first row of Fig. \[\[\]\] shows the results with the smallest network. On the left we see its density of classifiers \(D(E)\) as a histogram of 100 bins (A1). We determined \(D(E)\) by randomly sampling weight vectors from a uniform probability distribution on the 120-dimensional unit hypersphere. For each weight vector the true error of the corresponding classifier is measured with a balanced test set consisting of 10,000 randomly chosen test data from the two classes. The result is put into the corresponding bin. This was repeated 1,000,000 times. As expected, \(D(E)\) is symmetric and has its peak at \(E = 0.5\), and good classifiers with a small true error are rare within \(\mathcal{W}\), i.e., on the hypersphere.

In the middle (B1) we see the distribution of the test (true) errors at randomly found global minima of randomly chosen training sets. For each \(n\), the distribution of the test error along the y-axis corresponds to the mean normalized density of global minima \(Q_n(E)\). We obtained this by taking a randomly chosen training data set, and then randomly sampled weight vectors from the unit hypersphere until we found one with training error zero on this training set. Then we measured the corresponding true error (red dot) with the test data set and then took the next randomly chosen training data set. For each \(n\) this was repeated 1,000 times (1,000 red dots for each \(n\)). The box-plots show, that already for
Do highly over-parameterized neural networks generalize since bad solutions are rare?

Figure 3: Experiments with two Gaussians in 10 dimensions, classified by three different networks from top to bottom. The shape of the density of classifiers $D(E)$ (left column) of the first two networks hardly deviate, in spite of 10 times more weights in A2 than in A1. The column in the middle shows the distribution of test errors of training error zero solutions for increasing $n$. The average fraction of "bad" solutions converges to zero in these highly over-parameterized scenarios. In the right column the average of the test errors from the middle column (red x) are compared with the values from our bound determined by the $D(E)$s in the left column (blue crosses). In all three cases, the bound is valid, even (almost) exactly in C1 and C2.

$n = 30$ more than 75% of the training error zero solutions provide test (true) errors of less than 0.2, i.e., for $n = 30$ we have $\langle \phi_{\epsilon=0.2}(S) \rangle = 0.25$. In only one of the 1,000 trials there is a test error slightly above 0.4. "Bad" global minima are quickly becoming rare on the unit hypersphere.

On the right (C1) we show the average test error (red x) for each $n$, the average over the 1,000 red dots from (B1). This corresponds to $E_n$ of the l.h.s. of (9). The r.h.s. we can determine with the measured $D(E)$ from the left column, with the integral as sums over the bins. The results are shown by the blue crosses. Red and blue crosses match surprisingly well. Indeed, (5) is valid, there is no negative correlation. Even more, rather equality seems to be given, which means there is no correlation.

The second row of Fig. 3 shows the results of the network with one hidden layer and 100 hidden units. This network has ten times as many weights as the previous network with 10 hidden units. Nevertheless, the shape of the density of classifiers $D(E)$ (A2) hardly deviates from the $D(E)$ of the much smaller network (A1). Also the distribution of the test (true) errors of training errors zero solutions (B1 and B2) are more or less identical. As for the ten times smaller network, already for $n = 30$ in about 75% of the cases one ends up with a test error of less than 0.2, and in none of the
Do highly over-parameterized neural networks generalize since bad solutions are rare?

A PREPRINT

Figure 4: Classification of the two digits 1 and 2 of the MNIST data set with two output ReLUs (first row) and with one hidden layer with 10 hidden ReLUs and two output ReLUs (second row). Left (A1, A2) the density of classifiers $D(E)$ and in the middle (B1, B2) the distribution of the test errors (red dots) of training error zero solutions for increasing number of training data $n$. On the right (C1, C2) the average of the test errors (red x) compared with values of the bound (blue crosses) derived from the respective $D(E)$. The bound holds and is (almost) even tight in both cases.

1,000 trials there is a test error above 0.4. Accordingly, also Figs. C1 and C2 look identical. Again, the red and blue crosses match in C2 and, hence, equality in (5) is valid and there is no correlation.

In the third row of Fig. 3 we see the density of classifiers $D(E)$ for the network with ten hidden layers with ten neurons each and 1,020 weights (A3). The large peak at $E = 0.5$ stems from weight configurations where the classifier output is always the same class, independent from the input. The classifiers $h$ which correspond to these weight configurations are ruled out immediately as soon as training data from both classes occur, since then these $h$ do not provide zero training error and do not belong to the global minimum. With $D(E)$ having a different shape as in the two cases before, also Fig. B3 looks differently now. It converges more slowly, and the peak of $D(E)$ resolves with increasing $n$, as can be seen from the increasing box. For example, for $n = 2$ in half of the training data instances the training data are from the same class and there still are solutions from the peak at $E = 0.5$. We stopped with $n = 22$, since the time to find zero training error solutions by random sampling started to take too much time. In Fig. C3 the red and blue crosses do not match for every $n$, but the blue crosses always lie above the red ones. The bound (5) is valid, but now for some $n$ there is a positive correlation between $\phi_\varepsilon(S)$ and $\omega(S)$.

4.2 Data set from MNIST

In Fig. 4 we show the results of the same experiments with real data. We took the digits 1 and 2 of the MNIST data set as the two classes to be classified. Each digit image has a size of $28 \times 28$ pixels and thus the input dimension is 784. We used the same network structure as before, in one case with no hidden layer, i.e., only two leaky output ReLUs, one for each class, and in the other case a one-hidden layer MLP with 10 hidden leaky ReLUs and two leaky output ReLUs. In the first case, the classifier has 1,568 weights, and in the second case 7,860 weights. The training data sets were randomly drawn from a set of 6,000 images of each digit, respectively. These images were taken from the MNIST training set. For testing, 900 images of each digit were taken from the MNIST test set.

The results look very much like in Fig. 3 for the synthetic data. In the first row of Fig. 4 the results without hidden layer and 1,568 weights and in the second row the results with ten hidden units and 7,860 weights are shown. From left to right we see the density of classifiers $D(E)$, the distribution of test errors at training error zero for different $n$, and the average test errors compared to values of the bound (5). Again, the bound holds and even equality is given, i.e., there is no correlation between the fraction of "bad" global minima $\phi_\varepsilon(S)$ and the size of the set of global minima $\omega(S)$.
Do highly over-parameterized neural networks generalize since bad solutions are rare?

A Preprint

Figure 5: The left plots are adapted from [7] and show the mean test accuracies (and variances) in classifying motorcycles and airplanes on images from the Caltech101 dataset with VGG19 and ResNet18, respectively. Both networks are trained with stochastic gradient descent up to training error zero. To the right we see the mean test accuracies and their variances of solutions with training error zero obtained by random sampling. Good solutions quickly dominate within the set of solutions with training error zero, and the more than ten times larger VGG19 converges faster than ResNet18, which corresponds to gradient descent.

In both scenarios, the classifiers are highly over-parameterized for these small training data sets, but nevertheless good solutions quickly dominate. For both networks, already for \( n = 24 \) about 75\% of all test errors are below 0.2. Instead of a fifty-fifty chance of incorrect predictions, 75\% of the solutions with zero training error make fewer than one mistake in five cases.

4.3 VGG and ResNet and Caltech101

Finally, we take a look at very large Deep Networks. In [7] was shown that a VGG19 net with 140 million weights is able to learn to discriminate airplanes and motorcycles on images from the Caltech101 dataset [39] up to 95\% accuracy trained with only 20 examples from each class. This extremely over-parameterized network was initialized randomly and then trained by stochastic gradient descent up to training error zero without any regularization. Fig. 5A shows these results adapted from [7]. With 200 training data the test accuracy is almost 100\% despite complete over-fitting and memorizing the training data.

We hypothesize that the above results can be explained by “bad” solutions being rare in the set of training error zero solutions. Therefore, an ERM algorithm like stochastic gradient descent should very likely end up in a good solution also without regularization. We test our hypotheses like in the previous sections. Again, due to the positive homogeneity of the VGG19 net each classifier function realized by VGG19 corresponds to a point on the unit hypersphere in the weight space. We uniformly draw random weights from this sphere, and if the corresponding classifier function has zero training error we determine its test accuracy. The training examples are drawn from 798 motorbike and aircraft images from Caltech101, always the same number for each class, analog to the experiments in Fig. 5A. The rest of the images are used for determining the test accuracy. Like in Fig. 5A we perform the same experiment also for ResNet18, which has 11, 2 million parameters - more than ten times less than VGG19.

Fig. 5B shows the results for up to 10 samples per class (for larger \( n \) it became too computationally expensive). For each \( n \) we randomly drew random training sets and weights until we reached 100 zero training error solutions. Fig. 5B shows the mean of the corresponding test accuracies and their variances. To find zero training error solutions required several million trials for larger \( n \). Interestingly, on average it required much more trials for ResNet18 than for VGG19. Already for \( n = 10 \) the training error zero solutions of VGG19 have a mean test accuracy of 75\%. A rough extrapolation to \( n = 20 \) indeed leads to the range of VGG19 in Fig. 5A. Also ResNet18 being worse than VGG19 in spite of being ten times smaller corresponds to Fig. 5A. Both is in line with our conjecture.

5 Discussion

Experiments consistently demonstrate that highly over-parameterized classifiers can learn effectively even without explicit regularization and despite “memorizing” the training data up to zero training error. This contradicts traditional
We thank Christoph Linse for discussions and providing the graphs in Figure 5.

We experimentally validated our theoretical results using a synthetic dataset and a subset of MNIST with three and two different network architectures and sizes, respectively. Through random sampling, we determined the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We experimentally validated our theoretical results using a synthetic dataset and a subset of MNIST with three and two different network architectures and sizes, respectively. Through random sampling, we determined the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.

We proved mathematically that within the over-parameterized regime, the mean set size or parameter volume of "bad" ERM solutions diminishes at a much faster rate with the increase in the number of training samples compared to the mean set size of all possible ERM solutions. If the fraction of "bad" classifiers within the set of ERM solutions does not exhibit a negative correlation with the volume of ERM solutions, the fraction of "bad" classifiers will be small with high probability.

We introduced the density of classifiers (DOC) at a true error $E$, which depends on the set of classifier functions and the classification problem. The DOC determines the bound for the fraction of "bad" classifiers and serves as the primary characteristic of a given classifier/problem case. Since its shape (i.e., the normalized DOC) is important, it does not necessarily depend on the classifier's capacity or its number of parameters. Increasing the absolute density values by expanding the capacity of the classifier function set may keep the normalized DOC invariant, or may even improve it, as observed in numerous experiments.
References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” *Neural Information Processing Systems*, vol. 25, 01 2012.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” 2017.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, p. 201903070, 07 2019.

[6] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. J. Tax, “A brief prehistory of double descent,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 20, pp. 10625–10626, 2020.

[7] C. Linse and T. Martinetz, “Large neural networks learning from scratch with very few data and without explicit regularization,” 15th International Conference on Machine Learning and Computing (ICMLC), 2023.

[8] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[9] P. L. Bartlett and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results.,” *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2002.

[10] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, “Exploring generalization in deep learning,” in *NIPS*, 2017.

[11] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz, “SGD learns over-parameterized networks that provably generalize on linearly separable data,” in *International Conference on Learning Representations*, 2018.

[12] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, “The implicit bias of gradient descent on separable data,” *Journal of Machine Learning Research*, vol. 19, no. 70, pp. 1–57, 2018.

[13] K. Lyu and J. Li, “Gradient descent maximizes the margin of homogeneous neural networks,” in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 2020.

[14] M. S. Advani, A. M. Saxe, and H. Sopomolinsky, “High-dimensional dynamics of generalization error in neural networks,” *Neural Networks*, vol. 132, pp. 428–446, 2020.

[15] G. Vardi, “On the implicit bias in deep-learning algorithms,” *Commun. ACM*, vol. 66, p. 86–93, may 2023.

[16] B. Neyshabur, R. Tomioka, and N. Srebro, “Norm-based capacity control in neural networks,” in *Proceedings of the 28th Conference on Learning Theory (COLT)*, vol. PMLR 40, pp. 1376–1401, 2015.

[17] P. Bartlett, D. Foster, and M. Telgarsky, “Spectrally-normalized margin bounds for neural networks,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 6241–6250, 2017.

[18] N. Golowich, A. Rakhlin, and O. Shamir, “Size-independent sample complexity of neural networks,” in *COLT*, 2018.

[19] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, “Stronger generalization bounds for deep nets via a compression approach,” in *Proceedings of the 35th International Conference on Machine Learning*, pp. 254–263, 2018.

[20] Y. Li and Y. Liang, “Learning overparameterized neural networks via stochastic gradient descent on structured data,” in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[21] V. Nagarajan and Z. Kolter, “Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience,” in *International Conference on Learning Representations*, 2019.

[22] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, “The role of over-parametrization in generalization of neural networks,” in *International Conference on Learning Representations*, 2019.

[23] C. Wei and T. Ma, “Data-dependent sample complexity of deep neural networks via lipschitz augmentation,” in *NeurIPS*, 2019.

[24] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, “Fisher-rao metric, geometry, and complexity of neural networks,” in *Proceedings of Machine Learning Research*, pp. 888–896, 2019.
Do highly over-parameterized neural networks generalize since bad solutions are rare?

[25] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, p. 8580–8589, 2018.

[26] L. Chizat, E. Oyallon, and F. R. Bach, “On lazy training in differentiable programming,” in *Neural Information Processing Systems*, 2018.

[27] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” in *Neural Information Processing Systems*, 2018.

[28] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, “On exact computation with an infinitely wide neural net,” in *Neural Information Processing Systems*, 2019.

[29] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks,” in *International Conference on Machine Learning*, 2019.

[30] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, “Wide neural networks of any depth evolve as linear models under gradient descent*,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2020, p. 124002, dec 2020.

[31] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, “On the explicit role of initialization on the convergence and implicit bias of overparameterized linear networks,” *Proceedings of Machine Learning Research*, 2021.

[32] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1225–1234, PMLR, 20–22 Jun 2016.

[33] W. Mou, L. Wang, X. Zhai, and K. Zheng, “Generalization bounds of sgd for non-convex learning: Two theoretical viewpoints,” in *Proceedings of the 31st Conference On Learning Theory* (S. Bubeck, V. Perchet, and P. Rigollet, eds.), vol. 75 of *Proceedings of Machine Learning Research*, pp. 605–638, PMLR, 06–09 Jul 2018.

[34] Y. Lei, R. Jin, and Y. Ying, “Stability and generalization analysis of gradient methods for shallow neural networks,” in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[35] L. Oneto, S. Ridella, and D. Anguita, “Do we really need a new theory to understand over-parameterization?,” *Neurocomput.*, vol. 543, jul 2023.

[36] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, “Fantastic generalization measures and where to find them,” in *International Conference on Learning Representations*, 2020.

[37] V. Nagarajan and J. Z. Kolter, “Uniform convergence may be unable to explain generalization in deep learning,” in *Neural Information Processing Systems*, 2019.

[38] P. yeh Chiang, R. Ni, D. Y. Miller, A. Bansal, J. Geiping, M. Goldblum, and T. Goldstein, “Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent,” in *The Eleventh International Conference on Learning Representations*, 2023.

[39] F.-F. Li, M. Andreito, M. Ranzato, and P. Perona, “Caltech 101,” Apr 2022.

[40] V. N. Vapnik and A. Y. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities,” *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.
A Appendix

Proofs

Theorem 1:

Proof. For a given \( w \in W \), the loss \( L(y, h_w(x)) \) is a binomial random variable assuming the values 0 or 1 for training data drawn i.i.d. from \( P(x, y) \). The probability that \( L(y, h_w(x)) = 0 \) for a \((x, y)\) drawn from \( P(x, y) \) is \( 1 - E(w) \). Hence, for a \( S \) randomly drawn from \( P(x, y)^n \), the probability for \( h_w \) to be a consistent classifier with \( L(y_i, h_w(x_i)) = 0 \) for each \((x_i, y_i) \in S \) is \( (1 - E(w))^n \).

We define the indicator function \( 1_S(h_w) \), which is one for \( w \in W_\epsilon(S) \) and otherwise zero. For a fixed \( w \in W_\epsilon \) and \( S \) randomly drawn from \( P(x, y)^n \), the probability for \( 1_S(h_w) \) assuming the value one is \( (1 - E(w))^n \). With this we obtain

\[
\langle \omega_\epsilon(S) \rangle_S = \left( \int_{W_\epsilon} 1_S(h_w) \, dw \right)_S
= \int_{W_\epsilon} \langle 1_S(h_w) \rangle_S \, dw
= \int_{W_\epsilon} (1 - E(w))^n \, dw
= \int_{E_{\min} + \epsilon}^1 (1 - E)^n \, dE
\leq \Omega_\epsilon (1 - (E_{\min} + \epsilon))^n.
\]

With the second step we interchanged integrals. This can be done if for a random sequence \( S_1, \ldots, S_m \) of training sets

\[
\frac{1}{m} \sum_{i=1}^m 1_{S_i}(h_w) \to \langle 1_S(h_w) \rangle_S \text{ uniformly in probability}
\]

on \( W_\epsilon \) for \( m \to \infty \). But this is the case due to the finite VC-dimension of our classifier set \( \mathcal{H}_W \). \( \mathcal{H}_W \) loses the capacity to shatter \( S_1, \ldots, S_m \) with \( m \) becoming sufficiently large and, hence, the Uniform Convergence Theorem [40] applies.

\[\square\]

Corollary 1:

Proof. We show that for any \( a > 1 \) and \( 0 < \epsilon < 1 - E_{\min} \)

\[
\frac{\langle \omega_\epsilon(S) \rangle_S}{\langle \omega(S) \rangle_S} \leq \frac{\Omega_\epsilon}{\Omega} \frac{1}{(1 - g_\epsilon/a) + g_\epsilon/a e^{(1-1/a)\epsilon n}} \leq \frac{1}{g_\epsilon/a} e^{- (1/a) \epsilon n}.
\]

Corollary 1 is the special case for \( a = 2 \). The r.h.s. of course requires \( g_\epsilon/a > 0 \).

With Theorem 1 and dividing the integral in the denominator into two parts we have

\[
\frac{\langle \omega_\epsilon(S) \rangle_S}{\langle \omega(S) \rangle_S} = \frac{\int_{E_{\min}}^{1 - \epsilon} (1 - E)^n D(E) \, dE}{\int_0^1 (1 - E)^n D(E) \, dE}
\]

\[
= \frac{\int_{E_{\min} - \epsilon}^1 (1 - E)^n D(E) \, dE}{\int_0^{E_{\min} + \epsilon} (1 - E)^n D(E) \, dE + \int_{E_{\min} + \epsilon}^1 (1 - E)^n D(E) \, dE}
\]

\[
\leq \frac{1}{1 + \frac{\int_{E_{\min} + \epsilon}^1 (1 - E)^n D(E) \, dE}{\Omega_\epsilon (1 - (E_{\min} + \epsilon))^n}}.
\]
Do highly over-parameterized neural networks generalize since bad solutions are rare?

For any $a > 1$ and dividing the integral into two parts, we obtain for the nominator

$$\int_0^{E_{\text{min}} + \varepsilon} (1 - E)^n D(E) \, dE = \int_0^{E_{\text{min}} + \varepsilon/a} (1 - E)^n D(E) \, dE + \int_{E_{\text{min}} + \varepsilon/a}^{E_{\text{min}} + \varepsilon} (1 - E)^n D(E) \, dE \geq (\Omega - \Omega \varepsilon/a)(1 - (E_{\text{min}} + \varepsilon/a))^n + (\Omega \varepsilon/a - \Omega \varepsilon)(1 - (E_{\text{min}} + \varepsilon))^n.$$

But then with some basic transformations

$$\frac{\langle \omega_{\varepsilon}(S) \rangle}{\langle \omega(S) \rangle} \leq 1 + \frac{\Omega \varepsilon/a - \Omega \varepsilon}{\Omega \varepsilon} + \frac{1}{\Omega \varepsilon(1 - (E_{\text{min}} + \varepsilon/a))^n} \leq \frac{\Omega \varepsilon}{1 - g \varepsilon/a} + \frac{g \varepsilon/a}{(1 - (E_{\text{min}} + \varepsilon))^n}.$$

For the factor in the denominator with $n$ in the exponent we can write

$$\frac{(1 - (E_{\text{min}} + \varepsilon/a))^n}{(1 - (E_{\text{min}} + \varepsilon))^n} = e^{n(\ln(1 - (E_{\text{min}} + \varepsilon/a)) - \ln(1 - (E_{\text{min}} + \varepsilon)))} \geq e^{(1 - 1/a)\varepsilon}.$$

The last step uses the Taylor-expansion $\ln(1 - x) = -x - x^2/2 - x^3/3 - \ldots$. With this expansion it is easy to see that

$$\ln(1 - (E_{\text{min}} + \varepsilon/a)) - \ln(1 - (E_{\text{min}} + \varepsilon)) \geq (1 - 1/a)\varepsilon,$$

which concludes the proof.