Learning with Gradient Descent and Weakly Convex Losses

Dominic Richards
University of Oxford

Mike Rabbat
Facebook AI Research

Abstract

We study the learning performance of gradient descent when the empirical risk is weakly convex, namely, the smallest negative eigenvalue of the empirical risk's Hessian is bounded in magnitude. By showing that this eigenvalue can control the stability of gradient descent, generalisation error bounds are proven that hold under a wider range of step sizes compared to previous work. Out of sample guarantees are then achieved by decomposing the test error into generalisation, optimisation and approximation errors, each of which can be bounded and traded off with respect to algorithmic parameters, sample size and magnitude of this eigenvalue. In the case of a two layer neural network, we demonstrate that the empirical risk can satisfy a notion of local weak convexity, specifically, the Hessian's smallest eigenvalue during training can be controlled by the normalisation of the layers, i.e., network scaling. This allows test error guarantees to then be achieved when the population risk minimiser satisfies a complexity assumption. By trading off the network complexity and scaling, insights are gained into the implicit bias of neural network scaling, which are further supported by experimental findings.

1 Introduction

A standard task in machine learning is to fit a model on a collection of training data in order to predict some future observations. With a loss function evaluating the performance of a model on a single data point, a popular technique is to choose a model that minimises the empirical risk, i.e., the loss with respect to all of the training data, with a form of regularisation to control model complexity and avoid over fitting. Due to the complexity of modern models (e.g., neural networks), many empirical risk problems encountered in practice are non-convex, resulting in an optimisation problem where finding a global minimizer is generally computationally infeasible. Naturally, this has motivated the adoption of tractable methods that incrementally improve the model utilising first order gradients of the empirical risk. In the case of gradient descent, model parameters are iteratively updated by taking a step in the direction of the negative gradient, and there is a source of implicit regularization controlled by algorithmic parameters such as the step size and number of iterations.

A model's predictive performance is often measured by the population risk (i.e., expected loss on a new data point), which can be decomposed into optimisation and generalisation errors (Bousquet and Bottou, 2008). The optimisation error accounts for how well the model minimises the empirical risk, and thus, decreases with iterations of gradient descent. Meanwhile, the generalisation error accounts for the discrepancy between the empirical and population risk, and thus, intuitively increases with the iterations of gradient descent as the model fits to noise within the training data. Guarantees for the population risk are then achieved by considering these two errors simultaneously and, in our case, trading them off each other by choosing the number of iterations and step size appropriately, i.e., early stopping.

The optimisation and generalisation errors for gradient descent have been investigated under a variety of different structural assumptions on the loss function. For the optimisation error a rich literature on convex optimisation can be leveraged, see for instance (Nesterov, 2013), while a broader range of non-convex settings can be considered that include weak convexity (Nurminskii, 1973) as well as the Polyak-Łojasiewicz and quadratic growth conditions (Karimi et al., 2016). Meanwhile for the generalisation error, the technique of stability (Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002) has been applied to variants of gradient descent in both

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

the convex and non-convex settings (Hardt et al., 2016; Lin et al., 2016a; Kuzborskij and Lampert, 2018; Chen et al., 2018; Madden et al., 2020) with differing degrees of success. Specifically, although near-optimal rates are achieved in the convex case, for general non-convex losses restrictive step size conditions are currently required as bounds grow exponentially with the step size, number of iterations and the loss's smoothness (Hardt et al., 2016; Kuzborskij and Lampert, 2018; Yuan et al., 2019; Madden et al., 2020). This is in contrast to applications with non-convex losses where gradient descent is routinely applied with a variety of step sizes. One possible explanation for this difference is that problems encountered in typical applications are not arbitrarily non-convex; rather, the curvature of losses involving neural networks, as measured empirically by the Hessian spectrum, can often be more benign (LeCun et al., 2012; Sagun et al., 2016, 2017; Yao et al., 2018; Ghorbani et al., 2019; Yuan et al., 2019). This leads to the question of whether there are natural curvature assumptions that yield improved generalisation error bounds for gradient descent in the non-convex setting.

In this work we study the generalisation performance of gradient descent on loss functions that satisfy a notion of weak convexity (Nurminskii, 1973). The assumption relates to the loss's curvature as encoded by the Hessian spectrum, specifically, the magnitude of the most negative Eigenvalue. Our first main result (Theorem 1) shows that the magnitude of this Eigenvalue can control the stability of gradient descent, and thus, the generalisation error. Precisely, provided the magnitude of this Eigenvalue is sufficiently small with respect to the step size and number of iterations, and the sample size is sufficiently large, then generalisation error bounds on the same order as the convex setting hold with a wider range of step sizes than previously known (Hardt et al., 2016; Kuzborskij and Lampert, 2018).

Building upon our first result, guarantees on the population risk for gradient descent with weakly convex losses are then achieved by combining our generalisation error bounds with optimisation error bounds. In short, we note that the objective is convex when adding ℓ_2 regularisation scaled by the magnitude of the smallest negative Eigenvalue. Utilising this as a proof device, the optimisation error becomes tractable at the cost of an additional statistical bias, resulting from the regularisation, which can be interpreted as an approximation error. Our population risk bounds then hold provided the approximation error is below the statistical precision, which then requires an assumption on the complexity (as encoded by the ℓ_2 norm) of a population risk minimiser.

Utilising our population risk bounds, insights are then gained into the influence of neural network scaling on learning. Specifically, we consider a loss that is the composition of a convex function and a two layer neural network. The final layer of the network then being scaled by a coefficient that is decreasing with the network width to a polynomial power, allowing us to interpolate between what are referred to as the kernel (power 1/2 scaling) and mean field (power 1 scaling) regimes (Chizat et al., 2019). By showing that the magnitude of the Hessian's smallest negative Eigenvalue can be controlled by the network scaling, we show gradient descent can generalise provided the neural network is sufficiently wide. In particular, this suggests that *wider* networks with *larger scalings* are *more stable*, and thus, can be trained for longer. Moreover, by trading off the complexity of the population risk minimiser with the network scaling, we argue (theoretically and empirically) that networks with larger scaling (higher power) are biased to have weights with larger norms. We now summarise our primary contributions.

- **Generalisation Error Bound under Pointwise Weak Convexity.** Prove generalisation error bounds for gradient descent with point wise weak convexity (Assumption 2). When the weak convexity parameter is small enough with respect to the step size and number of iterations, and sample size is sufficiently large, bounds hold under milder step sizes conditions than in previous work (Hardt et al., 2016; Kuzborskij and Lampert, 2018) (Theorem 1).

- **Test Error Bounds under Weak Convexity.** Prove test error bounds for gradient descent with weak convexity (Assumption 3). When the weak convexity parameter is small enough and the minimiser of the population risk has small norm, bounds on the same order of the convex setting are achieved. (Theorem 2)

- **Influence of Neural Network Scaling.** When the loss is a composition of a convex function and a neural network, we prove that the weak convexity parameter is on the order of the network scaling. Utilising this in conjunction with our theoretical results and empirical evidence, we argue that networks with smaller scaling are biased towards weights with larger norm (Section 4).

The remainder of this work proceeds as follows. Section 1.1 summarises related works. Section 2 introduces the setting we consider as well as the generalisation error bounds. Section 3 presents bounds on the population risk for gradient descent with weakly convex losses. Section 4 considers the particular case of two layer neural networks.

1.1 Related Work

In this section we discuss a number of related work. For clarity, we adopt standard big $O(\cdot)$ notation so $a = O(b)$ if there is a constant $c > 0$ independent of dimension, sample size, iterations and stepsize such that $a \leq cb$.

Learning with First Order Gradient Methods. Guarantees on the population risk for gradient descent typically fall into one of two categories: single-pass or multi-pass. In the single-pass setting, each sample is used once to gain an unbiased estimate of the population risk gradient, and thus, guarantees following from studying the optimisation performance of stochastic gradient descent. Weak convexity has then be previously investigated within the optimisation community (Poliquin and Rockafellar, 1992, 1996; Rockafellar, 1981; Davis and Drusvyatskiy, 2019), with the most relevant work to ours being (Davis and Drusvyatskiy, 2019) which showed, for non-smooth objectives, a first order stochastic proximal algorithm converges to a stationary point at the rate $O(t^{-1/4})$ within t iterations. In contrast, we focus on smooth losses, standard gradient descent and guarantees for the function value.

In the multi-pass setting of this work, optimisation and generalisation errors are considered as each data point is used multiple times. To bound the generalisation error, stability (Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002) was applied within (Hardt et al., 2016; Lin et al., 2016a; Kuzborskij and Lampert, 2018; Yuan et al., 2019; Madden et al., 2020) for stochastic gradient descent. In the general non-convex setting these works then require an $O(1/t)$ step size after t iterations. In contrast, with *pointwise* weak convexity and a sufficiently large sample size N, our bounds hold with an $O(1/(\epsilon t^{\alpha}))$ step size where $0 \leq \alpha \leq 1$ and ϵ is an upper bound on the magnitude of the Hessian's (evaluated at the points of gradient descent only) smallest negative Eigenvalue.

A number of other works have investigated the generalisation of first-order gradient methods, which we now briefly discuss. The work (Lin et al., 2016b) considers the multi-pass setting for gradient descent on convex losses. The works (Charles and Papailiopoulos, 2018; Yuan et al., 2019; Madden et al., 2020) considers the stability on non-convex loss functions which satisfy Polyak-Łojasiewicz and/or quadratic growth conditions. Their setting is different to ours as their curvature conditions require that the gradient norm grows unbounded on unbounded domains (hence, e.g., excluding the logistic loss). The work (Mou et al., 2018) gives stability bounds for Stochastic Gradient Langevin Dynamics (SGLD), while (Chen et al., 2018) demonstrate a trade off between the stability and the optimisation error. Finally, we note works studying multi-pass gradient methods for non-parametric regression (Bauer et al., 2007; Rosasco and Villa, 2015; Pillaud-Vivien et al., 2018; Pagliana and Rosasco, 2019).

Scaling and Spectral Properties of Neural Networks. The scaling of neural networks has been shown to influence the inductive bias in a number of settings (Chizat et al., 2019; Woodworth et al., 2020), with two popular choices of scaling being "Neural Tanget Kernel" (NTK) (Jacot et al., 2018; Du et al., 2019b,a; Allen-Zhu et al., 2019; Zou et al., 2020; Arora et al., 2019a,b; Zou and Gu, 2019; Cao and Gu, 2019; Ji and Telgarsky, 2019; Jacot et al., 2019) and mean field (Chizat and Bach, 2018; Mei et al., 2018, 2019; Chizat et al., 2019). Specifically, it was found that a larger network scaling (i.e., mean field), can yield a richer implicit bias (Chizat et al., 2019; Woodworth et al., 2020). This aligns with our findings (Section 4.3) where, in short, a larger network scaling allows gradient descent to learn parameters with larger norms. A number of works have also investigated the loss's Hessian when using neural networks, which we now discuss. By decomposing the loss's Hessian into two matrices, the work (Jacot et al., 2019) studies the asymptotic moments of the Hessian with each of the aforementioned scalings. Similarly, (Pennington and Bahri, 2017) utilise random matrix theory to study the spectra of this decomposition by modelling each matrix individually. In each case, focus is placed upon the spectral distribution of the sum, which, as suggested by (Jacot et al., 2019), can be dominated by the first matrix. In contrast, our work is purely focused on the negative Eigenvalues, for which the second matrix primarily contributes since the first matrix is positive semi-definite in our case.[1] More generally, studies investigating the loss's Hessian spectrum can be traced back to (Bourrely, 1989; Bottou, 1991) with a number of follow-up works (LeCun et al., 2012; Sagun et al., 2016, 2017; Yao et al., 2018; Ghorbani et al., 2019; Yuan et al., 2019). Within (Sagun et al., 2017; Yuan et al., 2019) in particular, it was demonstrated empirically that the negative Eigenvalues of the Hessian decreased in magnitude during training. This observation is then explicitly leveraged within our work through a pointwise weak convexity assumption (Assumption 2 below).

Notation: For vectors $\omega, x \in \mathbb{R}^p$, denote the ith co-ordinate as ω_i, as well as standard Euclidean inner product $\langle \omega, x \rangle = \omega^\top x = \sum_{i=1}^p \omega_i x_i$ and ℓ_2 norm $\|\omega\|_2 = \langle \omega, \omega \rangle^{1/2}$. For matrices $A \in \mathbb{R}^{p \times q}$ denote the i,jth entry by A_{ij} and the Euclidean operator norm,

[1] The work (Jacot et al., 2019) also demonstrates each matrix within the decomposition becomes orthogonal within the limit, in which case, it suffices to study the spectra of each matrix individually.

equivalently spectral norm, as $\|A\|_2$. The Frobenius norm (i.e., entrywise ℓ_2 norm) of a matrix A is denoted as $\|A\|_F$. For square matrices $A \in \mathbb{R}^{p\times p}$ let $A \succeq 0$ denote that A is positive semi-definite; i.e., for any $u \in \mathbb{R}^p$ we have $u^\top A u \geq 0$. For $B \in \mathbb{R}^{p\times p}$ denote $A \succeq B$ if $A - B$ is positive semi-definite. For a function $f : \mathbb{R}^p \to \mathbb{R}$ denote the gradient with respect to the ith co-ordinate as $\partial f(\omega)/\partial\omega_i$, and the vector of gradients $\nabla f(\omega) \in \mathbb{R}^p$ so that $(\nabla f(\omega))_i = \partial f(\omega)/\partial\omega_i$. Denote the second derivative of a function with respect to the i, jth co-ordinates as $\partial f(\omega)/\partial\omega_i\partial\omega_j$ as well as the Hessian of the function $\nabla^2 f(\omega) \in \mathbb{R}^{p\times p}$ such that $(\nabla^2 f(\omega))_{ij} = \partial f(\omega)/\partial\omega_i\partial\omega_j$.

2 Setup and Generalisation Error Bounds

In this section we formally introduce the learning setting as well as the generalisation error bounds. Section 2.1 formally introduces the setting. Section 2.2 presents the assumptions and generalisation error bound. Section 2.3 presents a sketch proof of the generalisation error bound.

2.1 Generalisation, Stability and Gradient Descent

We consider a standard learning setting (Vapnik, 2013) which we now introduce. Let models be parameterised by Euclidean vectors $\omega \in \mathbb{R}^p$ and data points be denoted by $Z \in \mathcal{Z}$. A loss function $\ell : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}$ then maps a model ω and data point Z to a real number $\ell(\omega, Z) \in \mathbb{R}$. Observations are random variables following an unknown population distribution $Z \sim \mathbb{P}$, and the objective is to produce a model ω that minimises the expected loss with respect to the observations i.e. *population risk* $r(w) := \mathbf{E}_Z[\ell(w, Z)]$. To produce a model, a collection of independently and identically distributed samples $Z_i \sim \mathbb{P}$ for $i = 1, \ldots, N$ are observed and an approximation to the population risk is considered; i.e., the *empirical risk* $R(w) := \sum_{i=1}^N \ell(w, Z_i)/N$. In our case, an algorithm maps the observations to a model $\mathcal{A} : \mathcal{Z}^N \to \mathbb{R}^p$ so that $\widehat{\omega} = \mathcal{A}(Z_1, \ldots, Z_N)$. We then begin by considering the *generalisation error* $\mathbf{E}[R(\widehat{w}) - r(\widehat{w})]$.

To study the generalisation error we consider its stability (Bousquet and Elisseeff, 2002). Specifically, for $i = 1, \ldots, N$ consider the estimator with the ith data point resampled independently from the population, that is, $\widehat{\omega}^{(i)} = \mathcal{A}(Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_N)$ where $Z'_i \sim \mathbb{P}$. The expected generalisation error can then be written as (Bousquet and Elisseeff, 2002)

$$\underbrace{\mathbf{E}[R(\widehat{w}) - r(\widehat{w})]}_{\text{Generalisation Error}} = \frac{1}{N}\sum_{i=1}^N \mathbf{E}[\underbrace{\ell(\widehat{w}^{(i)}, Z'_i) - \ell(\widehat{w}, Z'_i)}_{\text{Stability}}]. \tag{1}$$

The stability of an estimator aligns with its sensitivity to individual data points in the training set. Intuitively, a stable estimator does not change much if a data point is resampled, and thus, generalises better owing to not depending in a strong way on any individual datapoint. This is captured within the above equality, where the difference on the right hand side involves the change in loss when resampling each training data point.

The algorithm considered will be gradient descent applied to the empirical risk. This produces a sequence of estimates indexed by $s \geq 1$ and denoted $\{\widehat{\omega}_s\}_{s\geq 0}$. For a sequence of non-negative step sizes $\{\eta_s\}_{s\geq 0}$ and an initialisation $\widehat{\omega}_0 \in \mathbb{R}^p$, the iterates of gradient descent are defined recursively for $s \geq 0$ as

$$\widehat{\omega}_{s+1} = \widehat{\omega}_s - \eta_s \nabla R(\widehat{\omega}_s).$$

For $i = 1, \ldots, N$ let us denote the sequence of estimators produced with the ith observation resampled by $\{\widehat{\omega}^{(i)}_s\}_{s\geq 0}$. These estimators are initialised at the same point $\widehat{\omega}^{(i)}_0 = \widehat{\omega}_0$ and are updated using the gradient of the empirical risk with the resampled data point $\nabla R^{(i)}(\omega) := \nabla R(\omega) + \big(\nabla\ell(\omega, Z'_i) - \nabla\ell(\omega, Z_i)\big)/N$.

2.2 Generalisation Error Bound for Gradient Descent with Weakly Convex Losses

In this section we present a bounded on the generalisation error when the empirical risk satisfies a notation of weak convexity. We begin with a collection of Lipschitz (Lip.) assumptions on the loss function and its gradients.

Assumption 1 (Loss Regularity) *There exists $L, \beta, \rho \geq 0$ such that for any $\omega, \omega' \in \mathbb{R}^p$ and $Z \in \mathcal{Z}$:*
L-Lip. loss: $|\ell(\omega, Z) - \ell(\omega', Z)| \leq L\|w - w'\|_2$.
β-Lip. gradient: $\|\nabla\ell(\omega, Z) - \nabla\ell(\omega', Z)\|_2 \leq \beta\|\omega - \omega'\|_2$.
ρ-Lip. Hessian: $\|\nabla^2\ell(\omega, Z) - \nabla^2\ell(\omega', Z)\|_2 \leq \rho\|\omega - \omega'\|_2$.

The first two conditions in Assumption 1 respectively state that the loss is L-Lipschitz and β-smooth, which are common assumptions when considering both the optimisation and generalisation of gradient descent algorithms (Nesterov, 2013; Hardt et al., 2016). The third condition has been considered previously when studying the stability of gradient descent (Kuzborskij and Lampert, 2018) and is satisfied if the third-order derivative exists and is bounded (Ge et al., 2015).

The next assumption is related to the Hessian of the empirical risk, and encodes a notation of weak convexity.

Assumption 2 (Pointwise Weak Convexity) *There exists a non-negative deterministic sequence* $\{\epsilon_s\}_{s\geq 0}$ *such that* almost surely *for any* $s \geq 0$ *and* $i = 1, \dots, N$

$$\nabla^2 R(\widehat{w}_s) \succeq -\epsilon_s I \quad \text{and} \quad \nabla^2 R^{(i)}(\widehat{w}_s^{(i)}) \succeq -\epsilon_s I \, .$$

Assumption 2 states that the smallest negative Eigenvalue of the Hessian is *almost surely* lower bounded at the points evaluated by gradient descent. This is milder than standard weak convexity, which requires a lower bound everywhere, i.e., for there to exist $\epsilon > 0$ such that $\nabla^2 R(\omega) \succeq -\epsilon I$ for any $\omega \in \mathbb{R}^p$. The lower bound also depends upon the time step $s \geq 1$ of gradient descent. This allows us to encode that the magnitude of the most negative Eigenvalue can decrease during training (Sagun et al., 2017; Yuan et al., 2019) (precisely Figure 2 in (Yuan et al., 2019)), which is itself linked to the residuals contributing to the Hessian's negative Eigenvalues (Pennington and Bahri, 2017; Jacot et al., 2019). In Section 4 we demonstrate for the composition of a smooth convex function and a two layer neural network, that Assumption 2 can be satisfied provided the network is sufficiently wide. With these assumptions we now bound the generalisation error of gradient descent.

Theorem 1 (Generalisation Error Bound) *Consider Assumptions 1 and 2,* $1 \geq \alpha \geq 0$ *and* $t \geq 1$. *If* $\max_{t\geq k\geq 0} \eta_k \beta \leq 3/2$, $\max_{t\geq k\geq 0} \eta_k(\epsilon_k + \frac{2\beta}{N}) + \eta_k^{\frac{1}{\alpha}} < 1/2$ *and*

$$N \geq 24\rho L \exp\Big(2\sum_{s=1}^{t} \eta_s\big(\epsilon_s + \frac{4\beta}{N}\big) + \eta_s^{\frac{1}{\alpha}}\Big) \sum_{j=1}^{t} \eta_j^{1-\frac{1}{2\alpha}} \sum_{\ell=0}^{j-1} \eta_\ell,$$

then the generalisation error of gradient descent satisfies

$$\mathbf{E}[R(\widehat{w}_t) - r(\widehat{w}_t)] \leq \frac{4L^2}{N} \sum_{j=0}^{t-1} \exp\Big(2 \sum_{s=j+1}^{t-1} \eta_s\big(\epsilon_s + \frac{4\beta}{N}\big) + \eta_s^{\frac{1}{\alpha}}\Big) \eta_j.$$

We now provide some discussion of Theorem 1. The condition on the sample size ensures that a sufficiently small step size $\{\eta_s\}_{s\geq 0}$ and number of iterations t is taken with respect to the total number of data points N. In particular, if we choose the step size $\eta_s = (s+1)^{-\alpha}/\log(t)$ for $s \geq 0$ and the minimum Eigenvalue is lower bounded so that $\epsilon_s \leq 1/t^{1-\alpha}$ for $s \geq 1$, then the term within the exponential is $O(t^{1-\alpha}/N)$. We would then expect this quantity to be small as it aligns with the Generalistion Error bound in the smooth and convex setting (Hardt et al., 2016). The condition on the sample size N is then dominated by the summation over step sizes, which are $\sum_{j=1}^{t}(j+1)^{-\alpha+1/2}\sum_{\ell=0}^{j-1}(\ell+1)^{-\alpha} = O(t^{2(1-\alpha)+1/2})$, and thus, we require the number of iterations to be upper bounded by $t = O(N^{1/(2(1-\alpha)+1/2)})$. Picking $\alpha = 3/4$, the iterations can grow linearly with the total number of samples, while $\alpha > 3/4$ allows the iterations to grow as $O(N^q)$ for some $1 \leq q \leq 2$. Given the condition on the sample size is satisfied, the resulting generalisation error bound is on the order of $O(t^{1-\alpha}/N)$, aligning with the smooth and convex setting (Hardt et al., 2016).

In comparison to previous generalisation error bounds for gradient descent with non-convex objectives, see for instance (Hardt et al., 2016; Kuzborskij and Lampert, 2018), we highlight that the bound in Theorem 1 allows for a larger step size to be taken. Previous bounds held in the non-convex setting with $\eta_s = O(1/s)$, while here under the additional assumption of point wise weak convexity, we can consider $\eta_s = O(1/s^\alpha)$ for $\alpha \in [0, 1]$. Such a step size will allow us to achieve guarantees on the optimisation error, and thus, the test error (see Section 3).

2.3 Proof Sketch of Theorem 1

Here a proof sketch of Theorem 1 is provided, with the full proof given in Appendix B.1. Recalling that the loss is L-Lipschitz and recalling (1), we see to bound the generalisation error it is sufficient to bound the deviation between the iterates of gradient descent with and without the resampled datapoint, i.e., $\widehat{\omega}_t - \widehat{\omega}_t^{(i)}$. Using the definition of the empirical risk gradient with the sampled data point $\nabla R^{(i)}(\omega)$ and the triangle inequality, we get for $k \geq 1$

$$\|\widehat{\omega}_k - \widehat{\omega}_k^{(i)}\|_2 \leq \frac{2\eta L}{N} + \underbrace{\|\widehat{\omega}_{k-1} - \widehat{\omega}_{k-1}^{(i)} - \eta\big(\nabla R(\widehat{\omega}_{k-1}) - \nabla R(\widehat{\omega}_{k-1}^{(i)})\|_2}_{\textbf{Expansiveness of Gradient Update}},$$

where $2\eta L/N$ arises from using the Lipschitz property to upper bound $\eta\big(\nabla\ell(\omega, Z_i') - \nabla\ell(\omega, Z_i)\big)/N$ for any $\omega \in \mathbb{R}^p$. The remaining term is referred to as the expansiveness of the gradient update (Hardt et al., 2016). To provide context, we now describe some previous approaches for bounding this term.

Following previous work (Hardt et al., 2016), when the loss is convex we have for any $x, y \in \mathbb{R}^p$ that $\|x - y - \eta(\nabla R(x) - \nabla R(y))\|_2 \leq \|x - y\|_2$. This allows the deviation to be simply unravelled and bounded $\|\widehat{\omega}_k - \widehat{\omega}_k^{(i)}\|_2 \leq 2\eta k/N$. Meanwhile, in the non-convex setting the expansiveness was upper bounded as

$\|x-y-\eta(\nabla R(x)-\nabla R(y))\|_2 \leq (1+\eta\beta)\|x-y\|_2$, ultimately yielding a bound on the order of $\|\widehat{\omega}_k - \widehat{\omega}_k^{(i)}\|_2 \leq \exp(\eta\beta t)\eta t/N$. To control this exponential term we then require $\eta = O(1/t)$.

The presence of the smoothness coefficient β within the exponential in the non-convex setting arises to control the loss curvature through an upper bound on the Hessian's spectral norm $\|\nabla^2\ell(\cdot, Z)\|_2$ (see also the data-dependent bound in (Kuzborskij and Lampert, 2018) and Lemma 12 in appendix of (Yuan et al., 2019)). Although, when comparing to the convex case, we intuitively believe that the exponential term may only depend upon negative Eigenvalues of the Hessian, i.e., the least Eigenvalue, since that can differentiate the convex and non-convex cases. For clarity, let us then assume that the loss is ϵ-weakly convex so $\nabla^2 R(\omega) \succeq -\epsilon I$ for any $\omega \in \mathbb{R}^p$. Then the expansiveness of the gradient updated can be upper bounded $\|x-y-\eta(\nabla R(x)-\nabla R(y))\|_2 \leq \|x-y\|_2/\sqrt{1-2\eta\epsilon}$ (see proof of Theorem 5 in Appendix B.3). The multiplicative factor can then be interpreted as $1/(1-2\eta\epsilon) = 1+2\eta\epsilon/(1-2\eta\epsilon) \leq \exp\big(2\eta\epsilon/(1-2\eta\epsilon)\big)$ and therefore directly controlled by the weak convexity parameter ϵ. In contrast, previously the multiplicative factor was $1+\eta\beta$. Unravelling the deviation with this upper bound yields a generalisation error bound on the order $\exp(\epsilon\eta t)\eta t/N$ (see Theorem 5 in Appendix), with the convex case being recovered when $\epsilon = 0$. Now, the condition on the sample size N within Theorem 1 arises from the fact we consider a milder *pointwise* weak convexity assumption which only evaluates the Hessian at the iterates of gradient descent (note standard weak convexity is a global lower bound). Precisely, the milder *pointwise* assumption results in the expansiveness of the gradient update containing higher order terms.

Remark 1 (Stochastic Gradient Descent) *Extending to the stochastic gradient setting, where a subset of randomly chosen data points are evaluated at each iteration, is challenging here as the higher order terms in the gradient expansiveness bound results in higher order moments of the deviation at previous iterations. We thus leave this direction to future work as we may require stronger high probability bounds, see also (Madden et al., 2020).*

3 Test Error Bounds for Gradient Descent with Weakly Convex Losses

In this section we use the generalisation error bounds to achieve guarantees on the population risk for gradient descent with weakly convex losses. The remainder of this section is then as follows. Section 3.1 presents the error decomposition. Section 3.2 presents bounds on the population risk for the iterates of gradient descent.

3.1 Test Error Decomposition

Recall we wish to produce a model that minimised the *population risk* $r(\omega)$. Therefore, denoting a population risk minimiser $\omega^\star \in \operatorname{argmin}_\omega r(\omega)$, we now set to investigate the *Test Error* of gradient descent. Specifically, for an independent uniform random variable $I \sim \text{Uniform}(1, \ldots, t)$ we consider $\mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I)]] - r(\omega^\star)$. We note that the iterate is evaluated at a uniform random variable $\widehat{\omega}_I$ as the loss is non-convex, and thus, the loss at the average of iterates $\frac{1}{t}\sum_{s=1}^t \widehat{\omega}_s$ is not immediately upper bounded by the average of the losses $E_I[r(\widehat{\omega}_I)]$.

To bound the test error we consider the following decomposition

$$\underbrace{\mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I)]] - r(\omega^\star)}_{\textbf{Test Error}} = \underbrace{\mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I) - R(\widehat{\omega}_I)]]}_{\textbf{Gen. Error}} + \underbrace{\mathbf{E}_I[E[R(\widehat{\omega}_I)]] - r(\omega^\star)}_{\textbf{Opt. \& Approx. Error}}. \quad (2)$$

The first term intuitively accounts for the discrepancy between the empirical and population risk and aligns with the generalisation error studied within Section 2. The second term, the **Optimisation and Approximation Error**, will be composed of two components. The first component is an optimisation error that, intuitively, accounts for how well the iterates of gradient descent minimise the empirical risk. The second component is an approximation error arising from the non-convex setting. Specifically, it results from a proof technique of adding regularisation to make the loss convex, and thus, the optimisation error tractable. The approximation error therefore reflects the weak convexity assumption. Specifically, it will be scaled by the weak convexity parameter and more generally (see Proposition 1 in Appendix A) can depend upon the Hessian structure. Section 3.2 then presents the upper bound on the test error utilising this error decomposition.

3.2 Test Error for Weakly Convex Losses

In this section we present a test error bound for weakly convex losses. Let us begin with the following assumption.

Assumption 3 (Weak Convexity) *There exists an $\epsilon > 0$ such that $\nabla^2 R(\omega) \succeq -\epsilon I$ for all $\omega \in \mathbb{R}^p$.*

Given this assumption, let us define the minimiser of the penalised objective $\widehat{\omega}_\epsilon^\star = \operatorname{argmin}_\omega \big\{R(\omega) + \epsilon\|\widehat{\omega}_0 - \omega\|_2^2\big\}$, as well as a minimiser of the unpenalised objective $\widehat{\omega}^\star \in \operatorname{argmin}_\omega R(\omega)$. The test error bound is then presented within the following theorem.

Theorem 2 *Let Assumptions 1 and 3 hold, and consider constant step size $\eta_s = \eta$ for all $s \geq 1$. If $\eta\beta \leq 1/2$, $2\eta\epsilon < 1$ then the test error of gradient descent is bounded*

$$\mathbf{E}_I[\mathbf{E}[r(\widehat{\omega}_I)]] - r(\omega^\star) \leq \underbrace{\frac{2L^2\eta t}{N}\exp\left(\frac{\eta t\epsilon}{1-2\eta\epsilon}\right)}_{\textit{Generalisation Error}} + \underbrace{\frac{\mathbf{E}[\|\widehat{\omega}_0 - \widehat{\omega}_\epsilon\|_2^2]}{2\eta t}}_{\textit{Optimisation Error}} + \underbrace{\epsilon\left(\eta t\mathbf{E}[R(\widehat{\omega}_0) - R(\widehat{\omega}^\star)] + \|\widehat{\omega}_0 - \omega^\star\|_2^2\right)}_{\textit{Approximation Error}}$$

We firstly note that we have now assumed standard weak convexity (Assumption 3), and therefore, there is no condition on the sample size N to control the **Generalisation Error** (recall proof sketch in Section 2.3). Meanwhile, the **Optimisation Error** and **Approximation Error** upper bounds the **Optimisation and Approximation Error** term given in (2). Naturally, the **Optimisation Error** decreases with both the step size and number of iterations ηt. Meanwhile, the **Approximation Error** arises, in short, from the proof technique of using the empirical risk penalised by $\epsilon\|\cdot\|_2^2$ within the analysis. To achieve guarantees on the test error, the product of the smallest Eigenvalue and the norm of the population risk minimiser $\epsilon\|\omega^\star\|_2^2$ must be sufficiently small, which can then be interpreted as a complexity assumption on the learning problem. This interplay is precisely investigated for a two layer neural network within Section 4.3. Given that the **Approximation Error** is sufficiently small, we then see that an $O(1/\sqrt{N})$ test error bound can be achieved by choosing $\eta t = \sqrt{N}$. The test error bound in the convex case (Hardt et al., 2016) is then be recovered by setting $\epsilon = 0$.

Remark 2 (Theorem 1 Iteration Condition) *Let us consider the test error bound if we restrict ourselves to the iteration condition in Theorem 1. In short, if $\eta = (t+1)^{-\alpha}/\log(t)$ we then get an $O(N^{-\frac{1-\alpha}{1/2+2(1-\alpha)}})$ test error bound, with $\alpha = 3/4$ then yielding an $O(N^{-1/4})$ bound in $t = O(N)$ iterations. Interestingly, this aligns with (Davis and Drusvyatskiy, 2019) who showed a stochastic proximal algorithm converges to a stationary point at the rate of $O(t^{-1/4})$ in the weakly convex case. Although, some care should be taken in making a direct comparison between our work and (Davis and Drusvyatskiy, 2019), since they consider: a more general non-smooth objective, a different algorithm and convergence to a stationary point. We therefore leave an investigation into a connection between the two methods to future work.*

4 Two Layer Neural Networks

In this section investigate the case of a two layer neural network. Section 4.1 formally introduces the setting we consider. Section 4.2 investigates the weak convexity parameter for two layer neural networks. Section 4.3 considers the approximation error for two layer neural networks. Section 4.4 presents experimental results.

4.1 Setup

Consider the standard supervised learning setting where data points decompose as $Z = (x, y)$ with covariates $x \in \mathbb{R}^d$ and a response $y \in \mathbb{R}$. Let $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ denote a smooth convex function that quantifies the discrepancy $g(\widehat{y}, y)$ between predicted $\widehat{y}$ and observed y responses. Consider prediction functions parameterised by $\omega \in \mathbb{R}^p$ and denoted $f(\cdot, \omega) : \mathbb{R}^d \to \mathbb{R}$. The loss function at the point $Z = (x, y)$ is then the composition between $g(\cdot, y)$ and $f(x, \omega)$ so that $\ell(\omega, Z) := g(f(x, \omega), y)$. We will be considering the gradient with respect to ω, therefore denote the first and second order gradient of the function g with respect to the first argument as $g'(\cdot, y) : \mathbb{R} \to \mathbb{R}$ and $g''(\cdot, y) : \mathbb{R} \to \mathbb{R}$, respectively. Throughout we will assume that g has uniformly bounded derivative $\max_{\widetilde{y},y} |g'(\widetilde{y}, y)| \leq L_{g'}$, which is satisfied, for instance, if g is the logistic loss.

The class of prediction functions considered in this section will be two layer neural networks of width $M > 0$ and scaling $1 \geq c \geq 1/2$. The network consists of a first layer of weights represented as a matrix $A \in \mathbb{R}^{M\times d}$ with jth row $A_j \in \mathbb{R}^d$, a second layer of weights denoted as a vector $v = (v_1, \ldots, v_M) \in \mathbb{R}^M$ and a smooth activation $\sigma : \mathbb{R} \to \mathbb{R}$. For an input $x \in \mathbb{R}^d$ the neural network then takes the form

$$f(x, \omega) := \frac{1}{M^c}\sum_{j=1}^{M} v_j\sigma(\langle A_j, x\rangle). \tag{3}$$

Typically, both the first and second layer of weights of the network are optimised, in which case, let parameter vector ω be the concatenation $\omega = (A, v) \in \mathbb{R}^{Md+M}$, where A is vectorised in a row-major manner. As we will see later, studying the case of optimising both layers simultaneously is challenging owing to the interactions between the first and second layers. It will therefore be insightful to also consider fixing the second layer of weights and only optimising the first layer. In this case denote $\omega = (A) \in \mathbb{R}^{Md}$.

4.2 Weak Convexity of Two Layer Neural Networks

We begin with the following theorem which considers weak convexity of the empirical risk. With data points

$Z_i = (x_i, y_i)$ for $i = 1, \dots, N$ let us denote the covariates empirical covariance as $\widehat{\Sigma} := \sum_{i=1}^N x_i x_i^\top / N \in \mathbb{R}^{d\times d}$.

Theorem 3 *Consider the loss function in Section 4.1 with Two Layer Neural Network* (3)*. Suppose that the activation σ has first and second derivatives bounded by $L_{\sigma'}$ and $L_{\sigma''}$ respectively. Then with $\omega = (A, v)$,*

$$\nabla^2 R(\omega) \succeq -\Big(L_{g'} L_{\sigma''} \|v\|_\infty \|\widehat{\Sigma}\|_2 + 2L_{\sigma'} L_{g'} \sqrt{\|\widehat{\Sigma}\|_2}\Big) \frac{1}{M^c} I$$

Theorem 3 demonstrates that the weak convexity parameter is on the order of the neural network scaling $O(1/M^c)$, suggesting it is smaller for wider networks with a larger scaling c. Note the bound consists of two terms: the first arising from the Hessian restricted of the first layer; and the second from the first and second layer of weights interacting. We now discuss using this bound in conjunction with Theorem 1 and 2.

Recall the generalisation error bound in Theorem 1 requires pointwise weak convexity Assumption 2. This considers the Hessian evaluated at the iterates of the gradient descent, and thus, if the initialisation and activation are bounded (to ensure the infinity norm of the second layer is $O(\sum_{s=0}^t \eta_s / M^c)$) as well as the covariates co-ordinates so $\max_{i=1,\dots,N} \|x_i\|_\infty \leq L_x$, Theorem 3 then ensures pointwise weak convexity Assumption 2 holds with $\epsilon_s = O(d/M^c)$. As a consequence, it is sufficient to scale the network width $M \geq b(d\sum_{s=0}^t \eta_s)^{1/c}$, for some constant $b \geq 0$, to ensure the exponential terms within the generalisation error bound of Theorem 1 remain bounded.

To bound the test error utilising Theorem 2 recall we require standard weak convexity Assumption 3 to hold. Since the first term in Theorem 4.2 depends upon the infinity norm of the second layer $\|v\|_\infty$, it is not feasible here to satisfy standard weak convexity as we can take $\|v\|_\infty \to \infty$. This (as well as Remark 4) motivates a more refined notation of weak convexity which is discussed in Appendix A. For this reason, let us now consider fixing the second layer and optimising the first layer only. Applying Theorem 2 then requires controlling the **Approximation Error** which itself requires an assumption on the squared Euclidean norm of the population risk minimiser $\|\omega^\star\|_2^2$. Since the scaling and complexity of the network can interplay, this is discussed within Section 4.3.

Remark 3 (Three Layer Neural Networks) *In Appendix D.2 we show that the minimum Eigenvalue can be bounded for a three layer neural network when fixing the middle layer. If the width of each layer is M the magnitude of the minimum negative Eigenvalue is $O(1/M^{2c-1/2})$, and thus, can be smaller than a two layer network when $c > 1/2$.*

4.3 Approximation Error for Two Layer Neural Networks

In this section we study the **Approximation Error** term within the test error bound of Theorem 2 in the case of a two layer neural network. Following the discussion at the end of Section 4.2 (as well as remark 4), we fix the second layer and only optimise the first so $\omega = (A) \in \mathbb{R}^{Md}$, with the case of optimising both layers studied in Appendix A.

Controlling the **Approximation Error** requires placing assumptions on the Euclidean norm of the neural network weights, and thus, the complexity of the network. We therefore compare to another notion of complexity to ensure the network is not made too simple. While a number of different notions of complexity have been investigated in previous work, see for instance (Neyshabur et al., 2015; Bartlett et al., 2017; Arora et al., 2019a), we consider the network's Total Weight (Bartlett, 1998) defined as $\mathrm{TW}(f) := \sum_{j=1}^M |v_j| \|A_j\|_2 / M^c$. The following assumption then introduces the complexity assumption used to control the approximation error in our case.

Assumption 4 *There exists $1 \geq \mu \geq 0$ and population risk minimiser $A^\star$ such that $\|A^\star\|_F \leq M^{1/2-\mu}$.*

Assumption 4 states there exists a set of first layer weights $A^\star$ with Frobenius norm bounded by $M^{1/2-\mu}$ that achieves the minimum population risk. Therefore, larger μ aligns with a stronger assumption, as the minimum risk can be attained by a set of weights with smaller norm. Although, some choices of μ are more natural than others when considering the choice of scaling $1 \geq c \geq 1/2$. Specifically, consider the following upper bound on the Total Weight $\mathrm{TW}(f) = \frac{\|v\|_2}{M^c} \sum_{i=1}^M \frac{|v_i|}{\|v\|_2} \|A_i\|_2 \leq \|v\|_2 \|A^\star\|_F / M^c$. Now, if each of the second layer weights are a constant so $\|v\|_2 = O(\sqrt{M})$ the Total Weight of the population risk minimiser under Assumption 4 is $O(M^{1-\mu-c})$. As such if $\mu > 1 - c$, the Total Weight of the network goes to zero as the network becomes wider (i.e., larger M), yielding effectively constant prediction functions for wide networks. Therefore, a larger scaling c can be seen to encourage networks with large norm weights.

Let us now consider the approximation error in Theorem 2 when gradient descent is initialised at $\widehat{\omega}_0 = 0$ and Assumption 4 holds. The approximation error is then $O(d(\eta t + M^{1-2\mu})/M^c)$, and therefore, to ensure it is decreasing in the width of the network M, we require $\mu > (1-c)/2$. How the network scaling c and the statistical assumption μ then interplay can be clearly seen in Figure 1. Intuitively, increasing the scaling c allows networks with larger norms (larger μ) to be approximated, but at the cost of not being able

to consider networks with smaller norms. Whether encouraging larger norm networks through the scaling c leads to improved generalisation will likely depend upon the problem setting.

Remark 4 (Limitation of Weak Convexity) *Consider optimising both the first and second layers for a Two Layer Neural Network and let $\omega^\star = (A^\star, v^\star)$ denote a population risk minimiser. Then the* ***Approx. Error*** *in Theorem 2 is $O((\|A^\star\|_F^2 + \|v^\star\|_2^2)/M^c)$, while the Total Weight is upper-bounded (from Young's inequality) as $TW(f) \leq (\|A^\star\|_F^2 + \|v^\star\|_2^2)/M^c$. It is therefore not possible in this case for the* ***Approx. Error*** *to decrease without the Total Weight vanishing. To remedy this, we show in Appendix A that a generalised weak convexity assumption can yield an* ***Approx. Error*** *that can alternatively be upper bounded by the Total Weight.*

4.4 Experimental Results

In this section we present experiments supporting the discussion in section 4.3. We consider a classification task on subsets of 3 datasets: HIGGS, SUSY (Baldi et al., 2014) and COVTYPE (Blackard and Dean, 1999), all of which can be found on the UCI Machine Learning Repository (Dua and Graff, 2017). The loss function is as in Section 4.1 with g being the logistic loss, and f a two layer neural network with sigmoid activation σ and both layers being optimised. Full batch gradient descent is performed on the N samples with a fixed step size, and the population risk is estimated every 500 iterations with training stopped once it increased consecutively for more than 5 batches of 500 iterations. Due to performing full batch gradient descent, we considered $c \in [0.5, 0.65]$ as the iterations required increased with the scaling c. The network was initialised with the first layer at 0 and the second layer from a standard Gaussian. Figure 2 then plots the Frobenius norm of the first layer and population risk (Test Error) against the neural network scaling c. Observe across the three data sets that the Frobeinus norm of the first layer is positively correlated with the scaling c, supporting the theoretical results discussed in Section 4.3. Moreover, note the estimated population risk decreases with the width of the neural network M.

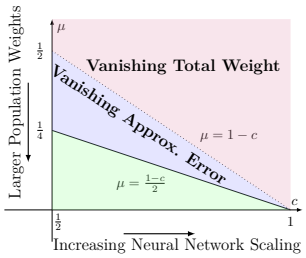

Figure 1: Weight Assumption 4 (μ) versus network scaling (c) *Red Region* (**Vanishing Total Weight**): Total Weight decreasing in M. *Blue region* (**Vanishing Approx. Error**): approximation error decreasing in M. *Solid line and green region*: bounds do not guarantee approximation error decreasing in M.

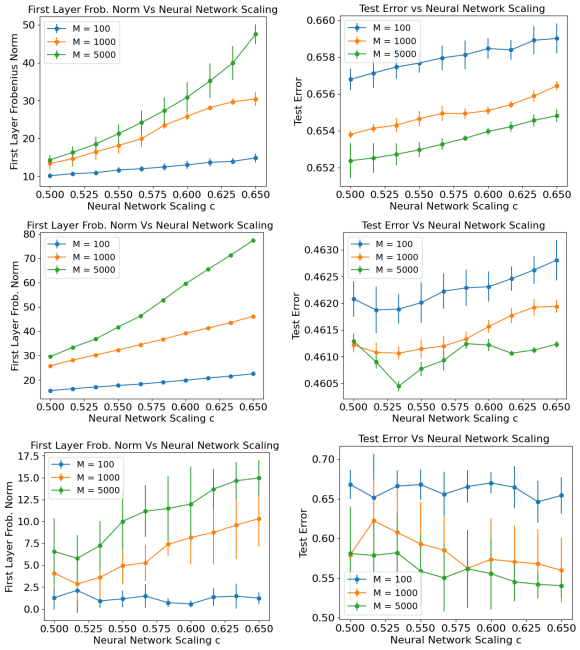

Figure 2: Plot of (*Left*) first layer Frobenius norm $\|A\|_F$ and (*Right*) population risk (labelled test error) versus network scaling c, for datasets: HIGGS (*Top*), SUSY (*Middle*) and COVTYPE (*Bottom*). *Top and Middle*: Step size $\eta = 0.1$, train size $N = 10^3$, test set size 10^5. Error bars from 2 subsets of data, each replicated 10 times. *Bottom*: step size $\eta = 10^{-3}$, train size $N = 10^4$ test set size 4×10^4. Error bars from 10 replications with single subset of data.

5 Conclusion

In this work we have investigated the stability of gradient descent under a notation of weak convexity. Generalisation error bounds were proven that both, hold under milder assumptions on the step size when compared to previous works (Hardt et al., 2016; Kuzborskij and Lampert, 2018), and can be combined with optimisation and approximation error bounds to achieve guarantees on the test error. In the case of a two layer neural networks, we demonstrated that the network scaling can control the weak convexity parameter, and thus, allow test error bounds to be achieved when a complexity assumption is placed on the population risk minimiser. Moving forward it would be natural to extend to stochastic and accelerated gradient methods, as well as deeper neural networks.

6 Acknowledgements

We would like to thank Levent Sagun, Aaron Defazio, Yann Ollivier, Sho Yaida, Patrick Rebeschini, Tomas Vaškevičius and Tyler Farghly for their feedback and suggestions.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8141–8150, 2019b.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

PL Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

J. Bourrely. Parallelization of a neural network learning algorithm on a hypercube. *Hypercube and distributed computers. Elsiever Science Publishing*, 1989.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1675–1685, 2019a. URL `http://proceedings.mlr.press/v97/du19c.html`.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations (ICLR)*, 2019b.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241, 2019.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1225–1234, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Arthur Jacot, Franck Gabriel, and Clement Hongler. The asymptotic spectrum of the hessian of dnn throughout training. In *International Conference on Learning Representations*, 2019.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016a.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *The Journal of Machine Learning Research*, 17(1):2718–2755, 2016b.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

EA Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, 1973.

Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in hilbert spaces. In *Advances in Neural Information Processing Systems*, pages 14481–14491, 2019.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806, 2017.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems 31*, pages 8125–8135. 2018.

René Poliquin and R Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

René A Poliquin and R Tyrrell Rockafellar. Amenable functions in optimization. *Nonsmooth optimization: methods and applications (Erice, 1991)*, pages 338–353, 1992.

R Tyrrell Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. 1981.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the

hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, volume 125, pages 3635–3673. PMLR, 2020.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. In *Advances in Neural Information Processing Systems*, pages 4949–4959, 2018.

Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pages 2608–2618, 2019.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2055–2064, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.