Global Convergence of Sobolev Training for Overparameterized Neural Networks

Jorio Cocola[1(✉)] and Paul Hand[1,2]

[1] Department of Mathemathics, Northeastern University, Boston, MA, USA
[2] Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
{cocola.j,p.hand}@northeastern.edu

Abstract. Sobolev loss is used when training a network to approximate the values and derivatives of a target function at a prescribed set of input points. Recent works have demonstrated its successful applications in various tasks such as distillation or synthetic gradient prediction. In this work we prove that an overparameterized two-layer relu neural network trained on the Sobolev loss with gradient flow from random initialization can fit any given function values and any given directional derivatives, under a separation condition on the input data.

Keywords: Gradient flow · Neural networks · Sobolev training

1 Introduction

Deep neural networks are ubiquitous and have established state of the art performances in a wide variety of applications and fields. These networks often have a large number of parameters which are tuned via gradient descent (or its variants) on an empirical risk minimization task. In particular in supervised learning it is often required that the output of the network fits certain values/labels that can be thought as coming from an unknown target function. In many settings, though, additional prior information on the task or target function might be available, and enforcing them might be of interest. One such example is the case of high order derivatives of the unknown target function, which, as shown in [5], naturally arises in problems such as distillation, in which a large teacher network is used to train a more compact student network, or prediction of synthetic gradients for training deep complex models. Therefore [5] proposed the *"Sobolev training"* which given training inputs $\{x_i\}_{i=1}^n$, attempts to minimizes the following empirical risk:

$$L(W) = \sum_{i=1}^{n} \left[\ell(f(W, x_i), f^*(x_i)) + \sum_{j=1}^{K} \ell_j (D_x^j f(W, x_i), D_x^j f^*(x_i)) \right] \tag{1}$$

where $f(W, x)$ is a neural network with input x and parameters W, f^* denotes the target function, ℓ is a loss penalizing the deviation from the outputs of f,

G. Nicosia et al. (Eds.): LOD 2020, LNCS 12565, pp. 574–586, 2020.
https://doi.org/10.1007/978-3-030-64583-0_51

and ℓ_j are loss functions penalizing the deviations of the j-th derivative $D_x^j f$ of the network f with respect to x from the j-th derivative $D_x^j f^*$ of the target f^*.

The empirical successes of Sobolev training have been demonstrated in a number of works. In [5] it was shown that Sobolev training leads to smaller generalization errors than standard training, in tasks such as distillation and synthetic gradient prediction especially in the low data regime. Similar results were also obtained for transfer learning via Jacobian matching in [14]. Earlier Sobolev training was applied in [13] in order to enforce invariance to translations and small rotations. More recently, instead, Sobolev training has been used in the context of anisotropic hyperelasticity in order to improve the predictions on the stress tensor (derivative of the network with respect to the input deformation tensor) in [16]. Finally, the idea of Sobolev training is also tightly connected to other techniques which have been recently successfully employed, such as attention matching in student distillation [18] and [5], or convex data augmentation for generalization and robustness improvement [19].

On the theoretical side, justification for Sobolev training was given by [5], extending the classical work of Hornik [9] and giving universal approximation properties of neural networks with relu activation function in Sobolev spaces. This result was then further improved for deep networks in [7]. While these works motivated the use of the Sobolev loss (1), conditions under which it can be successfully minimized were not given. In particular even though the network used in Sobolev training are usually shallow, the resulting loss (1) is highly non-convex and therefore the success of first order methods is not a priori guaranteed.

In this paper we study a two-layer relu neural network trained with a Sobolev loss when at each input point the output values and a set of directional derivatives of the target function are given. Leveraging recent results on training with standard losses [1,2,12,20,20] we show that if the network is sufficiently overparameterized, the weights are randomly initialized, and the data satisfy certain natural non-degeneracy assumptions, Gradient Flow achieves a global minimum.

2 Main Result

We study the training of neural networks with *"Directional Sobolev Training"*. In particular we assume we are given training data

$$\{x_i, y_i, V_i, h_i\}_{i=1}^n, \tag{2}$$

where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $h_i \in \mathbb{R}^k$ and $V_i \in \mathbb{R}^{d\times k}$ with orthonormal columns (unit Euclidean norm and pairwise orthogonal). This training data can be thought as being generated by a differentiable function $f^* : \mathbb{R}^d \to \mathbb{R}$ according to

$$y_i = f^*(x_i), \quad \text{and} \quad h_i = V_i^T \nabla f^*(x_i) \qquad \text{for } i = 1, \ldots, n \tag{3}$$

so that each entry of the vector h_i corresponds to a directional derivative of f^* in the direction given by the corresponding column of the matrix V_i. We will denote by $y \in \mathbb{R}^n$ and $h \in \mathbb{R}^{nk}$ the vectors with n entries y_i and n blocks h_i respectively.

In this work we study the training of a two-layer neural network with *width* m:

$$f(W, x) = \sum_{r=1}^{m} a_r \sigma(w_r^T x), \tag{4}$$

where a_r are fixed at initialization, $\sigma(z) = \max(0, z)$ is the relu activation function, and W is the weight matrix with rows $\{w_r^T\}_{r=1}^m$. The network weights $\{w_r\}_{r=1}^m$ are learned by minimizing the *Directional Sobolev Loss*

$$\min_{W \in R^{m \times d}} L(W) := \frac{1}{2}\sum_{i=1}^{n}(f(W, x_i) - y_i)^2 + \frac{1}{2}\sum_{i=1}^{n}\|V_i^T \nabla_x f(W, x_i) - h_i\|_2^2.$$

via the *Gradient Flow*

$$\frac{\mathrm{d}w_r(t)}{\mathrm{d}t} = -\frac{\partial L(W(t))}{\partial w_r}. \tag{5}$$

Note that even though the relu activation function σ is not differentiable, we let $\sigma'(z) = \mathbb{I}\{z > 0\}$ and $\sigma''(z) = 0$. This corresponds to the choice made in most of the deep learning libraries, and the dynamical system (5) can then be seen as the one followed in practice when using Sobolev training. Explicit formulas for the partial derivatives are given in the next section.

In this work we prove that for wide enough networks, gradient flow converges to a global minimizer of $L(W)$. In particular define the vectors of residuals $e(t) \in \mathbb{R}^n$ and $S(t) \in \mathbb{R}^{k \cdot n}$ with coordinates

$$[e(t)]_i = y_i - f(W(t), x_i) \qquad \big[S(t)\big]_i = h_i - V_i^T \nabla_x f(W(t), x_i). \tag{6}$$

We show that $e(t) \to 0$ and $S(t) \to 0$ as $t \to \infty$, under the following assumption of non-degeneracy of the training data.

Assumption 1. *The data are normalized so that* $\|x_i\|_2 = 1$ *and there exist* $0 < \delta_1$ *and* $0 \leq \delta_2 < k^{-1}$ *such that the following hold:*

$$\min_{i \neq j}(\|x_i - x_j\|_2, \|x_i + x_j\|_2) \geq \delta_1, \tag{7}$$

and for every $i = 1, \ldots, n$*:*

$$\max_{1 \leq j \leq k} |v_{i,j}^T x_i| \leq \delta_2, \tag{8}$$

where $v_{i,j}$ *are the columns of* V_i*.*

Given $w \in \mathbb{R}^d$ define the following "feature maps":

$$\phi_w(x_i) := \sigma'(w^T x_i) x_i,$$

$$\psi_w(x_i) := \sigma'(w^T x_i) V_i,$$

and matrix:

$$\Omega(w) = [\phi_w(x_1), \ldots, \phi_w(x_n), \psi_w(x_1), \ldots \psi_w(x_n)] \in \mathbb{R}^{d\times(k+1)n}.$$

The next quantity plays an important role in the proof of convergence of the gradient flow (5).

Definition 1. *Define the matrix $H^\infty \in \mathbb{R}^{n(k+1)\times n(k+1)}$ with entries given by $[H^\infty]_{i,j} = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}[\Omega(w)^T\Omega(w)]_{i,j}$, and let λ_* be its smallest eigenvalue.*

Under the non-degeneracy of the training set we show that H^∞ is strictly positive definite.

Proposition 1. *Under the Assumptions 1 the minimum eigenvalue of H^∞ obeys:*

$$\lambda_* \geq \frac{(1-k\delta_2)\delta_1}{100\,n^2}.$$

We are now ready to state the main result of this work.

Theorem 1. *Assume Assumption 1 is satisfied. Consider a one hidden layer neural network (4), let $\gamma = \|y\|_2 + \|h\|_2$, set the number of hidden nodes to $m = \Omega(n^6k^4\gamma^2/(\lambda_*^4\delta^3))$ and i.i.d. initialize the weights according to:*

$$w_r \sim \mathcal{N}(0, I_d) \quad \text{and} \quad a_r \sim \text{unif}\{-\frac{1}{m^{1/2}}, \frac{1}{m^{1/2}}\} \qquad \text{for } r = 1, \ldots, m. \tag{9}$$

Consider the Gradient Flow (5), then with probability $1-\delta$ over the random initialization of $\{w_r\}_r$ and $\{a_r\}_r$, for every $t \geq 0$:

$$\|e(t)\|_2^2 + \|S(t)\|_2^2 \leq \exp(-\lambda_* t)\left(\|e(0)\|_2^2 + \|S(0)\|_2^2\right)$$

and in particular $L(W(t)) \to 0$ as $t \to \infty$.

The proof of this theorem is given in Sect. 3, below we will show how to extend this result to a network with bias.

2.1 Consequences for a Network with Bias

Given training data (2) generated by a target function f^* according to (3), in this section we demonstrate how the previous theory can be extended to the Sobolev training of a two-layer network with width m and bias term b:

$$g(W, b, x) = \sum_{r=1}^{m} a_r\sigma(\alpha\, w_r^T x + \beta\, b_r) \tag{10}$$

where[1] $\alpha = 1/(2k)$ and $\beta = \sqrt{1-\alpha^2}$.

[1] Notice that the introduction of the constants α and β does not change the expressivity of the network.

Similarly as before, the network weights $\{w_r\}_{r=1}^m$ and biases $\{b_r\}_{r=1}^m$ are learned by minimizing the *Directional Sobolev Loss*

$$\min_{W\in\mathbb{R}^{m\times d}, b\in\mathbb{R}^d} L(W,b) := \frac{1}{2}\sum_{i=1}^n (g(W,b,x_i)-y_i)^2 + \frac{1}{2}\sum_{i=1}^n \| \frac{1}{\alpha} V_i^T \nabla_x g(W,b,x_i) - \frac{h_i}{\alpha}\|_2^2. \tag{11}$$

via the *Gradient Flow*

$$\frac{\mathrm{d}w_r(t)}{\mathrm{d}t} = -\frac{\partial L(W(t))}{\partial w_r} \quad \text{and} \quad \frac{\mathrm{d}b_r(t)}{\mathrm{d}t} = -\frac{\partial L(W(t))}{\partial b_r}. \tag{12}$$

Based on the following separation conditions on the input point $\{x_i\}_i$ we will prove convergence to zero training error of the Sobolev loss.

Assumption 2. *The data are normalized so that* $\|x_i\|_2 = 1$ *and there exists* $\hat{\delta}_1 > 0$ *such that the following holds*

$$\min_{i\neq j}(\|x_i - x_j\|_2) \geq \hat{\delta}_1. \tag{13}$$

Define the vectors of residuals $e(t) \in \mathbb{R}^n$ and $S(t) \in \mathbb{R}^{k\cdot n}$ with coordinates

$$[e(t)]_i = y_i - g(W(t), b(t), x_i), \quad \big[S(t)\big]_i = (h_i - V_i^T \nabla_x g(W(t), b(t), x_i))/\alpha,$$

then the next theorem follows readily from the analysis in the previous section.

Theorem 2. *Assume Assumption 2 is satisfied. Consider a two-layer neural network (10), let* $\gamma = \|y\|_2 + \|h/\alpha\|_2$*, set the number of hidden nodes to* $m = \Omega(n^6 k^4 \gamma^2/(\lambda_0^4 \delta^3))$ *and i.i.d. initialize the weights according to:*

$$w_r \sim \mathcal{N}(0, I_d), \quad b_r \sim \mathcal{N}(0,1) \quad \textit{and} \quad a_r \sim \text{unif}\{-\frac{1}{m^{1/2}}, \frac{1}{m^{1/2}}\}.$$

Consider the Gradient Flow (12), then with probability $1-\delta$ *over the random initialization of* $\{b_r\}_r$*,* $\{w_r\}_r$ *and* $\{a_r\}_r$*, for every* $t \geq 0$

$$\|e(t)\|_2^2 + \|S(t)\|_2^2 \leq \exp(-\lambda_* t)\,(\|e(0)\|_2^2 + \|S(0)\|_2^2)$$

where $\lambda_* \geq \frac{\delta}{200\, n^2}$ *and* $\delta = \min(\alpha\hat{\delta}_1, 2\beta)$*. In particular* $L(W(t), b(t)) \to 0$ *as* $t \to \infty$*.*

2.2 Discussion

Theorem 2 establishes that the gradient flow (12) converges to a global minimum and therefore that a wide enough network, randomly initialized and trained with the Sobolev loss (11) can interpolate any given function values and directional derivatives. We observe that recent works in the analysis of standard training [12,21] have shown that using more refined concentration results and control on the weight dynamics, the polynomial dependence on the number of samples n can be lowered. We believe that by applying similar techniques to the Sobolev training, the dependence of m from the number of samples n and derivatives k can be further improved.

Regarding the assumptions on the input data, we note that [2,6,12] have shown convergence of gradient descent to a global minimum of the standard ℓ_2 loss, when the input points $\{x_i\}_{i=1}^n$ satisfy the separation conditions (7). These conditions ensure that no two input points x_i and x_j are parallel and reduce to (13) for a network with bias. While the separation condition (7) is also required in Sobolev training, the condition (8) is only required in case of a network without bias as a consequence of its homogeneity.

Finally, the analysis of gradient methods for training overparameterized neural networks with standard losses has been used to study their inductive bias and ability to learn certain classes of functions (see for example [2,3]). Similarly, the results of this paper could be used to shed some light on the superior generalization capabilities of networks trained with a Sobolev loss and their use for knowledge distillation.

3 Proof of Theorem 1

We follow the lines of recent works on the optimization of neural networks in the Neural Tangent Kernel regime [4,10,12] in particular the analysis of [2,6,17]. We investigate the dynamics of the residuals error $e(t)$ and $S(t)$, beginning with that of the predictions. Let $\bar{F}(W(t), x_i) = V_i^T \nabla_x f(W(t), x_i)$, then:

$$\begin{aligned}\frac{d}{dt} f(W(t), x_i) &= \sum_{r=1}^{m} \frac{\partial f(W(t), x_i)}{\partial w_r}^T \frac{\mathrm{d}w_r(t)}{\mathrm{d}t} \\ &= \sum_{j=1}^{n} A_{ij}(t)(y_j - f(W(t), x_j)) + \sum_{j=1}^{n} B_{ij}(t)(h_j - \bar{F}(W(t), x_j)),\end{aligned}$$

$$\begin{aligned}\frac{d}{dt} \bar{F}(W(t), x_i) &= \sum_{r=1}^{m} \frac{\partial \bar{F}(W(t), x_i)}{\partial w_r}^T \frac{\mathrm{d}w_r(t)}{\mathrm{d}t} \\ &= \sum_{j=1}^{n} B_{ji}(t)^T (y_j - f(W(t), x_j)) + \sum_{j=1}^{n} C_{ij}(t)(h_j - \bar{F}(W(t), x_j)),\end{aligned}$$

where we defined the matrices $A(t) = \sum_{r=1}^m A_r(t) \in \mathbb{R}^{n\times n}$, $B(t) = \sum_{r=1}^m B_r(t) \in \mathbb{R}^{n \times n\cdot k}$, $C(t) = \sum_{r=1}^m C_r(t) \in \mathbb{R}^{n\cdot k \times n\cdot k}$, with block structure:

$$[A_r(t)]_{ij} := \frac{\partial f(W(t), x_i)}{\partial w_r}^T \frac{\partial f(W(t), x_j)}{\partial w_r} = \frac{1}{m}\sigma'(w_r^T x_i)\sigma'(w_r^T x_j) x_i^T x_j,$$

$$[B_r(t)]_{ij} := \frac{\partial f(W(t), x_i)}{\partial w_r}^T \frac{\partial \bar{F}(W(t), x_j)}{\partial w_r} = \frac{1}{m}\sigma'(w_r^T x_i)\sigma'(w_r^T x_j) x_i^T V_j,$$

$$[C_r(t)]_{ij} := \frac{\partial \bar{F}(W(t), x_i)}{\partial w_r}^T \frac{\partial \bar{F}(W(t), x_j)}{\partial w_r} = \frac{1}{m}\sigma'(w_r^T x_i)\sigma'(w_r^T x_j) V_i^T V_j.$$

The residual errors (6) then follow the dynamical system:

$$\frac{d}{dt}\begin{pmatrix} e \\ S \end{pmatrix} = -H(t)\begin{pmatrix} e \\ S \end{pmatrix} \tag{14}$$

where $H(t) \in \mathbb{R}^{n(k+1)\times n(k+1)}$ is given by:

$$H(t) = \begin{bmatrix} A(t) & B(t) \\ B(t)^T & C(t) \end{bmatrix}.$$

We moreover observe that if we define:

$$\Omega_r(t) := [\frac{\partial f(W(t), x_1)}{\partial w_r}, \ldots, \frac{\partial f(W(t), x_n)}{\partial w_r}, \frac{\partial \bar{F}(W(t), x_1)}{\partial w_r}, \ldots, \frac{\partial \bar{F}(W(t), x_n)}{\partial w_r}]$$

and $H_r(t) := \Omega_r(t)^T \Omega_r(t)$, then direct calculations show that $H(t) = \sum_r H_r(t)$ and $H(t)$ is symmetric positive semidefinite for all $t \geq 0$. In the next section we will show that $H(t)$ is strictly positive definite in a neighborhood of initialization, while in Sect. 3.2 we will show that this holds for large enough time leading to global convergence to zero of the errors.

3.1 Analysis Near Initialization

In this section we analyze the behavior of the matrix $H(t)$ and the dynamics of the errors $e(t), S(t)$ near initialization. We begin by bounding the output and directional derivatives of the network for every t.

Lemma 1. *For all $t \geq 0$ and $1 \leq i \leq n$, it holds:*

$$\begin{aligned} \|\frac{\partial f(W(t), x_i)}{\partial w_r}\|_2 &\leq \frac{1}{\sqrt{m}} \\ \|\frac{\partial \bar{F}(W(t), x_i)}{\partial w_r}\|_2 &\leq \sqrt{\frac{k}{m}} \\ \|H_r(t)\|_2 &\leq \frac{n(k+1)}{m}. \end{aligned}$$

We now lower bound the smallest eigenvalue of $H(0)$.

Lemma 2. *Let $\delta \in (0, 1)$, and $m \geq \frac{32}{\lambda_*} n(k+1)\ln(n(k+1)/\delta)$ then with probability $1 - \delta$ over the random initialization:*

$$\lambda_{min}(H(0)) \geq \frac{3}{4}\lambda_*$$

We now provide a bound on the expected value of the residual errors at initialization.

Lemma 3. *Let $\{w_r\}_r$ and $\{a_r\}$ be randomly initialized as in (9) then the residual errors (6) at time zero satisfy with probability at least $1-\delta$:*

$$\|e(0)\|_2 + \|S(0)\|_2 \leq \frac{2\sqrt{nk}+\gamma}{\delta^{1/2}}$$

where $\gamma = \|y\|_2 + \|h\|_2$.

Next define the neighborhood around initialization:

$$B_0 := \{W : \|H(W) - H(W(0))\|_F \leq \frac{\lambda_*}{4}\}$$

and the escape time

$$\tau_0 := \inf\{t : W(t) \notin B_0\}. \tag{15}$$

We can now prove the main result of this section which characterizes the dynamics of $e(t)$, $S(t)$ and the weights $w_r(t)$ in the vicinity of $t=0$.

Lemma 4. *Let $\delta \in (0,1)$ and $m \geq \frac{32}{\lambda_*} n(k+1)\ln(n(k+1)/\delta)$ then with probability $1-\delta$ over the random initialization, for every $t \in [0, \tau_0]$:*

$$\|e(t)\|_2^2 + \|S(t)\|_2^2 \leq \exp(-\lambda_* t)\left(\|e(0)\|_2^2 + \|S(0)\|_2^2\right)$$

and

$$\|w_r(t) - w_r(0)\|_2 \leq \frac{4}{\lambda_*}\sqrt{\frac{kn}{m}}(\|e(0)\|_2 + \|S(0)\|_2) =: R$$

Proof. Observe that if $t \in [0, \tau_0]$, by Lemma 2 with probability $1-\delta$:

$$\lambda_{\min}(H(t)) \geq \lambda_{\min}(H(0)) - \|H(t) - H(0)\|_F \geq \frac{\lambda_*}{2}.$$

Therefore using (14) it follows that for any $t \in [0, \tau_0]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\left(\|e(t)\|_2^2 + \|S(t)\|_2^2\right) \leq -\frac{\lambda_*}{2}\left(\|e(t)\|_2^2 + \|S(t)\|_2^2\right),$$

which implies the first claim by Gronwall's lemma. Next, using (5), the bounds in Lemma 1 and the above inequality we obtain:

$$\begin{aligned}
\|\frac{\mathrm{d}}{\mathrm{d}t} w_r(t)\|_2 &= \|\frac{\partial}{\partial w_r} L(W(t))\|_2 \\
&= \|\sum_{i=1}^n (y_i - f(W(t), x_i))\frac{\partial f(W(t), x_i)}{\partial w_r} + \sum_{i=1}^n \frac{\partial \bar{F}_i(t)}{\partial w_r}(h_i - \bar{F}_i(t))\| \\
&\leq \frac{1}{\sqrt{m}}\sum_{i=1}^n |y_i - f(W(t), x_i)| + \sqrt{\frac{k}{m}}\sum_{i=1}^n \|h_i - \bar{F}_i(W(t))\| \\
&\leq 2\sqrt{\frac{kn}{m}} e^{-\frac{\lambda_*}{2}t}(\|e(0)\| + \|S(0)\|).
\end{aligned}$$

We can therefore conclude by bounding the distance from initialization as:

$$\|w_r(t) - w_r(0)\| \leq \int_0^t \|\frac{\mathrm{d}}{\mathrm{d}t} w_r(s)\|_2 \mathrm{d}t \leq \frac{4}{\lambda_*}\sqrt{\frac{kn}{m}}(\|e(0)\| + \|S(0)\|)$$

3.2 Proof of Global Convergence

In order to conclude the proof of global convergence, according to Lemma 4, we need only to show that $\tau_0 = \infty$ where τ_0 is defined in (15). Arguing by contradiction, assume this is not the case and $\tau_0 < \infty$. Below we bound $\|H(\tau_0) - H(0)\|_F$.

Let $Q_{ijr} := |\sigma'(w_r^T(\tau_0)x_i)\sigma'(w_r^T(\tau_0)x_j) - \sigma'(w_r^T(0)x_i)\sigma'(w_r^T(0)x_j)|$, then from the formulas for $[A_r]_{ij}$, $[B_r]_{ij}$ and $[C_r]_{ij}$ in the previous sections, we have:

$$|[A(\tau_0) - A(0)]_{ij}| \leq \frac{1}{m}\sum_{r=1}^{m} Q_{ijr},$$

$$\|[B^T(\tau_0) - B^T(0)]_{ij}\|_2 \leq \frac{\sqrt{k}}{m}\sum_{r=1}^{m} Q_{ijr},$$

$$\|[C(\tau_0) - C(0)]_{ij}\|_F \leq \frac{k}{m}\sum_{r=1}^{m} Q_{ijr}.$$

Let $R > 0$ as in in Lemma 4, then with probability at least $1-\delta$ for all $1 \leq r \leq m$ we have $\|w_r(\tau_0) - w_r(0)\| \leq R$. Moreover observe that if $\|w_r(\tau_0) - w_r(0)\| \leq R$ and $\sigma'(w_r(0)^T x_i) \neq \sigma'(w_r(\tau_0)^T x_i)$, then $|w_r(0)^T x_i| \leq R$. Therefore, for any $i \in [n]$ and $r \in [m]$ we can define the event $E_{i,r} = \{|w_r(0)^T x_i| \leq R\}$ and observe that:

$$\mathbb{I}\{\sigma'(w_r(0)^T x_i) \neq \sigma'(w_r(\tau_0)^T x_i)\} \leq \mathbb{I}\{E_{i,r}\} + \mathbb{I}\{\|w_r(\tau_0) - w_r(0)\| > R\}.$$

Next note that $w_r(0)^T x_i \sim \mathcal{N}(0,1)$, so that $\mathbb{P}(E_{i,r}) \leq \frac{2R}{\sqrt{2\pi}}$ and in particular:

$$\frac{1}{m}\sum_{r=1}^{m}\mathbb{E}[Q_{ijr}] \leq \frac{1}{m}\sum_{r=1}^{m}(\mathbb{P}[E_{i,r}]+\mathbb{P}[E_{j,r}])+2\mathbb{P}[\cup_r\{\|w_r(\tau_0)-w_r(0)\| > R\}] \leq \frac{4R}{\sqrt{2\pi}}+2\delta.$$

By Markov inequality we can conclude that with probability at least $1-\delta$:

$$\|A(\tau_0) - A(0)\|_F \leq \sum_{i,j}|[A(\tau_0)]_{ij} - [A(0)]_{ij}| \leq \frac{4n^2}{\sqrt{2\pi}\delta}R + \frac{2n^2}{m},$$

$$\|B^T(\tau_0) - B^T(0)\|_F \leq \sum_{i,j}\|[B^T(\tau_0)]_{ij} - [B^T(0)]_{ij})\|_2 \leq \frac{4n^2\sqrt{k}}{\sqrt{2\pi}\delta}R + \frac{2n^2}{m},$$

$$\|C(\tau_0) - C(0)\|_F \leq \sum_{i,j}\|[C(\tau_0)]_{ij} - [C(0)]_{ij})\|_F \leq \frac{4n^2 k}{\sqrt{2\pi}\delta}R + \frac{2n^2}{m},$$

and using Lemma 3 together with the definition of H and R :

$$\|H(\tau_0) - H(0)\|_F \leq \frac{16n^2\,k}{\sqrt{2\pi}\,\delta}R + \frac{8n^2}{m} = \mathcal{O}\Big(\frac{n^3k^2}{\sqrt{m}\delta^2}\frac{\max(1,\gamma)}{\lambda_*}\Big)$$

Then choosing $m = \Omega(n^6k^4\gamma^2/(\lambda_*^4\delta^3))$ we obtain

$$\|H(\tau_0) - H(0)\|_F < \frac{\lambda_*}{4}$$

which contradicts the definition of τ_0 and therefore $\tau_0 = \infty$.

Acknowledgements. PH is supported in part by NSF CAREER Grant DMS-1848087.

A Supplementary proofs for Sect. 3.1

In this section we provide the remaining proofs of the results in Sect. 3.1. We begin recalling the following matrix Chernoff inequality (see for example [15, Theorem 5.1.1]).

Theorem 3 (Matrix Chernoff). *Consider a finite sequence X_k of $p \times p$ independent, random, Hermitian matrices with $0 \preceq X_k \preceq LI$. Let $X = \sum_k X_k$, then for all $\epsilon \in [0, 1)$*

$$\mathbb{P}\Big[\lambda_{min}(X) \leq \epsilon\lambda_{min}\big(\mathbb{E}[X]\big)\Big] \leq pe^{-(1-\epsilon)^2\lambda_{min}(\mathbb{E}[X])/2L} \tag{16}$$

In order to lower bound the smallest eigenvalue of $H(0)$ we use Lemma 1 together with the previous concentration result.

Proof (Lemma 2). We first note that $\mathbb{E}[H(0)] = \mathbb{E}[\sum_r H_r(0)] = H^\infty$, and moreover $H_r(0)$ is symmetric positive semidefinite with $\lambda_{\max}(H_r) \leq n(k+1)/m$ by Lemma 1. Applying then the concentration bound (16) with the assumption $m \geq \frac{32}{\lambda_*} n(k+1)\ln(n(k+1)/\delta)$ gives the thesis.

We next upper bound the errors at initialization.

Proof (Lemma 3). Note that for any x_i, due the the assumption on the independence of the weights at initialization and the normalization of the data:

$$\mathbb{E}[(f(W, x_i))^2] = \sum_{r=1}^{m} \frac{1}{m}\mathbb{E}[\sigma(w_r^T x_i)^2] \leq 1$$

and similarly for the directional derivatives

$$\mathbb{E}[\|\bar{F}(W, x_i)\|_2^2] = \mathbb{E}_{g\sim\mathcal{N}(0,I)}[\|\sigma'(g^T x_i)V_i^T g\|_2^2] \leq \sum_{j=1}^{k} \mathbb{E}[(v_{i,j}^T g)^2] \leq k.$$

We conclude the proof by using Jensen's and Markov's inequalities.

B Proof of Proposition 1

Consider the $d \times (k+1)$ matrices $\mathbf{X}_i = [x_i, V_i]$, and for $w \in \mathbb{R}^d$ define

$$\hat{\psi}_w(x_i) = \sigma'(w^T x_i)\mathbf{X}_i.$$

and the $d \times (k+1)n$ matrix:

$$\widehat{\Omega}(w) = [\hat{\psi}_w(x_1), \ldots, \hat{\psi}_w(x_n)] \ \in \mathbb{R}^{d\times n(k+1)}$$

which corresponds to a column permutation of $\Omega(w)$. Next observe that the matrix $\widehat{H}^\infty = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}[\widehat{\Omega}(w)^T\widehat{\Omega}(w)]$ is similar to H^∞ and therefore has the same eigenvalues. In this section we lower bound λ_* by analyzing $\widehat{H}^\infty$.

We begin recalling some facts about the spectral properties of the products of matrices.

Definition 2 **([8]).** *Let* $\mathbf{A} = [A_{\alpha\beta}]_{\alpha=1,\dots,n}^{\beta=1,\dots,n}$ *and* $\mathbf{B} = [B_{\alpha\beta}]_{\alpha=1,\dots,n}^{\beta=1,\dots,n}$ *be* $np \times np$ *matrices in which each block is in* $p \times p$*. Then we define the block Hadamard product of* $\mathbf{A}\square\mathbf{B}$ *as the* $np \times np$ *matrix with:*

$$\mathbf{A}\square\mathbf{B} := [A_{\alpha\beta}B_{\alpha\beta}]_{\alpha=1,\dots,n}^{\beta=1,\dots,n}$$

where $A_{\alpha\beta}B_{\alpha\beta}$ *denotes the usual matrix product between* $A_{\alpha\beta}$ *and* $B_{\alpha\beta}$*.*

Generalizing Schur's Lemma one has the following regarding the eigenvalues of the block Hadamard product of two block matrices.

Proposition 2 **([8]).** *Let* $\mathbf{A} = [A_{\alpha\beta}]_{\alpha=1,\dots,n}^{\beta=1,\dots,n}$ *and* $\mathbf{B} = [B_{\alpha\beta}]_{\alpha=1,\dots,n}^{\beta=1,\dots,n}$ *be* $np \times np$ *positive semidefinite matrices. Assume that every* $p \times p$ *block of* $\mathbf{A}$ *commutes with every* $p \times p$ *block of* $\mathbf{B}$*, then:*

$$\lambda_{min}(\mathbf{B}\square\mathbf{A}) = \lambda_{min}(\mathbf{A}\square\mathbf{B}) \geq \lambda_{min}(A) \cdot \min_{\alpha} \lambda_{min}(B_{\alpha\alpha})$$

We finally recall the following on the eigenvalues of Kronecker product of matrices.

Proposition 3 **([11]).** *Let* $A \in \mathbb{R}^{n\times n}$ *with eigenvalues* $\{\lambda_i\}$ *and* $B \in \mathbb{R}^{m\times m}$ *with eigenvalues* $\{\mu_i\}$*, then Kronecker product* $A \otimes B$ *between* A *and* B *has eigenvalues* $\{\lambda_i\mu_j\}$*.*

We next define the following random kernel matrix.

Definition 3. *Let* $w \sim \mathcal{N}(0, I)$ *then define the random matrix* $\mathcal{M}(w) \in \mathbb{R}^{n\times n}$ *with entries* $[\mathcal{M}(w)]_{ij} = \sigma'(w^T x_i)\sigma'(w^T x_j)$*.*

The next result from [12] establishes positive definiteness of this matrix in expectation, under the separation condition (7).

Lemma 5 **([12]).** *Let* $x_1, \dots, x_d$ *in* $\mathbb{R}^d$ *with unit Euclidean norm and assume that* (7) *is satisfied for all* $i = 1, \dots d$*. Then the following holds:*

$$\mathbb{E}_{w\sim\mathcal{N}(0,I)}[\mathcal{M}(w)] \succeq \frac{\delta_1}{100n^2}$$

Finally let $\mathbf{X} \in \mathbb{R}^{d\times n(k+1)}$ block matrix with $d \times (k+1)$ blocks $\mathbf{X}_i$. Thanks to the assumption (8) the following result on the Gram matrices $\mathbf{X}_i^T\mathbf{X}_i$ holds.

Lemma 6 *Assume that the condition* (8) *is satisfied, then for any* $i = 1, \dots, n$ *we have* $\lambda_{min}(\mathbf{X}_i^T\mathbf{X}_i) \geq 1 - k\delta_2 > 0$*.*

Proof. The claim follows by observing that by *Gershgorin's Disk Theorem*:

$$|\lambda_{\min}(\mathbf{X}_i^T \mathbf{X}_i) - 1| \leq \sum_{1 \leq j \leq k} |x_i^T v_{i,j}| \leq k\delta_2.$$

Finally observe that we can write:

$$\widehat{H}^\infty = \mathbb{E}_{w \sim \mathcal{N}(0,I)} \big[(\mathbf{X}^T \mathbf{X}) \square (\mathcal{M}(w) \otimes I)\big].$$

so that Proposition 2, Proposition 3, Lemma 5 and Lemma 6 allow to derive the thesis of Proposition 1.

References

1. Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. arXiv preprint arXiv:1811.03962 (2018)
2. Arora, S., Du, S.S., Hu, W., Li, Z., Wang, R.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv preprint arXiv:1901.08584 (2019)
3. Bietti, A., Mairal, J.: On the inductive bias of neural tangent kernels. In: Advances in Neural Information Processing Systems. pp. 12873–12884 (2019)
4. Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. arXiv preprint arXiv:1812.07956 (2018)
5. Czarnecki, W.M., Osindero, S., Jaderberg, M., Swirszcz, G., Pascanu, R.: Sobolev training for neural networks. In: Advances in Neural Information Processing Systems, pp. 4278–4287 (2017)
6. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:1810.02054 (2018)
7. Gühring, I., Kutyniok, G., Petersen, P.: Error bounds for approximations with deep relu neural networks in $w^{s,p}$ norms. arXiv preprint arXiv:1902.07896 (2019)
8. Günther, M., Klotz, L.: Schur's theorem for a block Hadamard product. Linear Algebra Appl. **437**(3), 948–956 (2012)
9. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. **4**(2), 251–257 (1991)
10. Jacot, A., Gabriel, F., Hongler, C.: Neural tangent Kernel: convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems, pp. 8571–8580 (2018)
11. Laub, A.J.: Matrix Analysis for Scientists and Engineers, vol. 91. SIAM (2005)
12. Oymak, S., Soltanolkotabi, M.: Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. arXiv preprint arXiv:1902.04674 (2019)
13. Simard, P., Victorri, B., LeCun, Y., Denker, J.: Tangent prop-a formalism for specifying selected invariances in an adaptive network. In: Advances in Neural Information Processing Systems, pp. 895–903 (1992)
14. Srinivas, S., Fleuret, F.: Knowledge transfer with Jacobian matching. arXiv preprint arXiv:1803.00443 (2018)
15. Tropp, J.A., et al.: An introduction to matrix concentration inequalities. Found. Trends® Mach. Learn. **8**(1–2), 1–230 (2015)
16. Vlassis, N., Ma, R., Sun, W.: Geometric deep learning for computational mechanics part i: anisotropic hyperelasticity. arXiv preprint arXiv:2001.04292 (2020)

17. Weinan, E., Ma, C., Wu, L.: A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. Sci. China Math. **63**, 1235–1258 (2020). https://doi.org/10.1007/s11425-019-1628-5
18. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
19. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
20. Zou, D., Cao, Y., Zhou, D., Gu, Q.: Stochastic gradient descent optimizes over-parameterized deep ReLUnetworks. arXiv preprint arXiv:1811.08888 (2018)
21. Zou, D., Gu, Q.: An improved analysis of training over-parameterized deep neural networks. In: Advances in Neural Information Processing Systems, pp. 2053–2062 (2019)