A Theoretical Analysis of “Lazy Training” in Deep Learning

Henry Smith

Yale University

May 29, 2022

Abstract

In “On Lazy Training in Differentiable Programming”, Chizat, Oyallon, and Bach categorize a regime of neural network training in which a differentiable model behaves like the linearization around its initialization. Consequently, training a model which is perhaps highly nonconvex in its weights is equivalent to training an affine model. For our report, we present two principal theorems from Chizat and colleagues' paper which comprise the foundation of “lazy training”. Furthermore, we rigorously prove each of these results, expanding upon the arguments made by the authors to give a fuller understanding of how and why lazy training occurs. In addition to the theory, we also provide an understanding of the practical applications of lazy training. Specifically, inspired by the work of Woodworth and colleagues, we introduce the argument that lazy training leads to poor model generalization in sparse problems.

1 Introduction

The problem of optimizing the weights of a neural network is, in general, a highly nonconvex one. Indeed, in even the simplest of models–those with a single hidden layer, for instance–we observe that the network function is highly nonconvex as a function of its parameter space at each fixed input. While the theoretical results for nonconvex optimization problems are considerably less desirable than their convex counterparts, this has not stopped practitioners from applying gradient-based methods to train neural networks (batch gradient descent, stochastic gradient descent, Adam, etc.). What actually occurs during network training, though, is a more nebulous topic.

In particular, we will study "implicit biases" in gradient descent when training the weights of a neural network. Intuitively, an "implicit bias" means that, under certain circumstances, gradient descent behaves in a predictable way and results in a network with certain properties. The implicit bias in which we are interested has been coined "lazy training" by Chizat, Oyallon, and Bach in their 2018 paper "On Lazy Training in Differentiable Programming". In the lazy training regime, a network behaves as a linearization around its initialization, and so training a model which is highly nonconvex in its parameters is simplified to a training an affine model. When the network is identically zero at its initialization, this means that training is equivalent to a kernel method with feature map given by the gradient of the network at its initialization. Of course, it cannot generally be true that networks are trained in the lazy regime, and so we wish to prove some formal results about when lazy training occurs.

We structure our report of lazy training as follows. In Section 2, we introduce mathematical notation that will be helpful for discussing and proving the theoretical results in Section 3. This section is also of particular importance as it defines the "linearized model", which forms the basis of lazy training. In Section 3, we state, prove, and discuss the implications of three main results from "On Lazy Training in Differentiable Programming" by Chizat, Oyallon and Bach. These results constitute the fundamental theory of lazy training, suggesting under what conditions lazy training occurs and how it is realized throughout training. We conclude our discussion of lazy training in Section 4 by suggesting some extensions of the results from [2]. In particular, we mention the properties (i.e. biases) of those models trained with lazy training and suggest some settings in which non-lazy training is preferable.

2 Preliminaries

Having provided some intuition for lazy training, we proceed to formalize it mathematically. For the sake of convenience, the notation we use is the same as that presented in [2].

We will consider $\mathbb{R}^p$ a parameter space, $\mathcal{F}$ a Hilbert space, $h : \mathbb{R}^p \to \mathcal{F}$ a model, and $R : \mathcal{F} \to \mathbb{R}_+$ a loss function. Notice here that h does not map inputs to outputs, but rather a vector of parameters to an element of the Hilbert space $\mathcal{F}$.

In our particular setting of neural networks, we choose $\mathcal{F}$ to be the Hilbert space consisting of all possible network functions. As a familiar example, suppose that we are given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ and let $\mathcal{D}$ be the corresponding empirical distribution (i.e. $\mathbb{P}((\boldsymbol{x}, y) = (\boldsymbol{x}_1, y_1)) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{(\boldsymbol{x}_i, y_i)=(\boldsymbol{x}_1, y_1)})$. Further, let $\mathcal{D}_{\boldsymbol{x}}$ be the $\boldsymbol{x}$ marginal distribution of $\mathcal{D}$. Then we can choose our Hilbert space $\mathcal{F}$ to be $L^2(\mathcal{D}_{\boldsymbol{x}}, \mathbb{R}^d)$, which consists of those functions which are square integrable with respect to $\mathcal{D}_{\boldsymbol{x}}$. More generally, we can choose $\mathcal{F} = L^2(\rho_{\boldsymbol{x}}, \mathbb{R}^d)$, where $\rho_{\boldsymbol{x}}$ is any probability measure on the input space $\mathbb{R}^d$ [2]. In the case that $\mathcal{F}$ is a function space with $f : \mathbb{R}^d \to \mathbb{R}$ for each $f \in \mathcal{F}$, we let $h : \boldsymbol{w} \mapsto f(\boldsymbol{w}, \cdot)$ denote the map from parameter vector $\boldsymbol{w}$ to network function $f(\boldsymbol{w}, \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$. To continue on with our previous example, we could then choose our loss function to be $R(h(\boldsymbol{w})) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[(y - f(\boldsymbol{w}, \boldsymbol{x}))^2\right]$, which is the mean-squared error, or equivalently the empirical risk corresponding to the square loss.

Throughout our paper, we will only be interested in those models h which are differentiable in $\boldsymbol{w} \in \mathbb{R}^p$

as well as those loss functions R which are differentiable in $f \in \mathcal{F}$. This is because we will use gradient-based methods to minimize the scaled objectives (3), which clearly necessitates that each of h and R is differentiable. We formally state our assumption on h and R as it is given by Chizat and colleagues:

Assumption (from [2])**.** *The model $h : \mathbb{R}^p \to \mathcal{F}$ is differentiable with a locally Lipschitz differential Dh. When we specify that Dh is locally Lipschitz, we are referring to the map $\boldsymbol{w} \mapsto Dh(\boldsymbol{w})$, and so the Lipschitz constant is defined with respect to the operator norm. Moreover, $R : \mathcal{F} \to \mathbb{R}_+$ is differentiable with a Lipschitz gradient.*

Now that we have made clear the model h of interest as well as the assumptions on h, we introduce the linearization of h around its initialization. In particular, given a model h as well as some initialization $\boldsymbol{w}_0 \in \mathbb{R}^p$, we define the linearized model to be

$$\bar{h}(\boldsymbol{w}) = h(\boldsymbol{w}_0) + Dh(\boldsymbol{w}_0)(\boldsymbol{w} - \boldsymbol{w}_0), \quad \boldsymbol{w} \in \mathbb{R}^p. \tag{1}$$

Once again, for the particular case of h mapping a parameter vector to a neural network $h : \boldsymbol{w} \mapsto f(\boldsymbol{w}, \cdot)$, we get

$$\bar{f}(\boldsymbol{w}, \boldsymbol{x}) = f(\boldsymbol{w}_0, \boldsymbol{x}) + D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w} - \boldsymbol{w}_0) \quad \boldsymbol{x} \in \mathbb{R}^d, \quad \boldsymbol{w} \in \mathbb{R}^p.$$

In even greater specificity, when the output of the network is one-dimensional $f(\boldsymbol{w}, \cdot) : \mathbb{R}^d \to \mathbb{R}$, then our linearized model is

$$\begin{aligned}\bar{f}(\boldsymbol{w}, \boldsymbol{x}) =& f(\boldsymbol{w}_0, \boldsymbol{x}) + \nabla_w f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w} - \boldsymbol{w}_0) \\ =& f(\boldsymbol{w}_0, \boldsymbol{x}) + \langle \nabla_w f(\boldsymbol{w}_0, \boldsymbol{x}), \boldsymbol{w} - \boldsymbol{w}_0 \rangle \quad \boldsymbol{x} \in \mathbb{R}^d, \quad \boldsymbol{w} \in \mathbb{R}^p.\end{aligned} \tag{2}$$

One will discern that for this case of $f(\boldsymbol{w}, \cdot) : \mathbb{R}^d \to \mathbb{R}$, $\bar{h}$ is no more than a first-order Taylor expansion of the model h around its initialization $\boldsymbol{w}_0$ for each fixed $\boldsymbol{x} \in \mathbb{R}^d$.

So far, we have suggested two mathematical objects of interest, the model h and its corresponding linearized model $\bar{h}$. For each vector $\boldsymbol{w} \in \mathbb{R}^p$ we compute the misfit of $h(\boldsymbol{w})$ and $\bar{h}(\boldsymbol{w})$ according to $R(h(\boldsymbol{w}))$ and $R(\bar{h}(\boldsymbol{w}))$, respectively. However, rather than dealing with $R(h(\boldsymbol{w}))$ and $R(\bar{h}(\boldsymbol{w}))$, Chizat and colleagues consider the objective functions corresponding to the scaled models αh and $\alpha \bar{h}$ for some $\alpha > 0$:

$$F_\alpha(\boldsymbol{w}) = \frac{1}{\alpha^2} R(\alpha h(\boldsymbol{w})) \qquad \bar{F}_\alpha(\boldsymbol{w}) = \frac{1}{\alpha^2} R(\alpha \bar{h}(\boldsymbol{w})). \tag{3}$$

Here, we are doing no more than scaling the output of each of h and $\bar{h}$ by a positive factor $\alpha > 0$. One should notice that the factor of $\frac{1}{\alpha^2}$ which appears in (3) is simply a positive normalization factor and does not affect the minima of the objective functions (that is, $\frac{1}{\alpha^2} R(\alpha h(\boldsymbol{w}))$ and $R(\alpha h(\boldsymbol{w}))$ have the same set of minimizers).

Corresponding to the scaled objective functions $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$ we define the gradient flow dynamics, denoted $(\boldsymbol{w}_\alpha(t))_{t \geq 0}$ and $(\bar{\boldsymbol{w}}_\alpha(t))_{t \geq 0}$, respectively, with $\boldsymbol{w}_\alpha(0) = \bar{\boldsymbol{w}}_\alpha(0) = \boldsymbol{w}_0$. The gradient flow of F_α is a path in the parameter space space $\mathbb{R}^p$ that solves the initial value problem

$$\boldsymbol{w}'_\alpha(t) = -\nabla F_\alpha(\boldsymbol{w}_\alpha(t)), \quad \boldsymbol{w}_\alpha(0) = \boldsymbol{w}_0. \tag{4}$$

The gradient flow of $\bar{F}_\alpha$ is defined analogously. Of key interest to practitioners of machine learning is gradient descent, which can be thought of as a discrete time version of the gradient flow dynamics [4]. Specifically, using the forward Euler discretization of the gradient flow dynamics with stepsize $\eta > 0$, we get $(\boldsymbol{w}_\alpha(t+1) - \boldsymbol{w}_\alpha(t))/\eta = -\nabla F_\alpha(\boldsymbol{w}_\alpha(t)) \Leftrightarrow \boldsymbol{w}_\alpha(t+1) = \boldsymbol{w}_\alpha(t) - \eta \nabla F_\alpha(\boldsymbol{w}_\alpha(t))$ for each $t \in \mathbb{N} \cup \{0\}$, which is exactly equal to the $t+1$ gradient descent update.

We mention that when the model h is m-positive homogeneous, then scaling the model output by α is equivalent to scaling the model weights by $\alpha^{1/m}$. That is, $h(\alpha \boldsymbol{w}) = \alpha^m h(\boldsymbol{w})$ for every $\boldsymbol{w} \in \mathbb{R}^p$ and each

$\alpha > 0$. Therefore, for m-positive homogeneous model h, the gradient flow on $\frac{1}{\alpha^2}R(\alpha h(\boldsymbol{w}))$ with $\boldsymbol{w}_\alpha(0) = \boldsymbol{w}_0$ is equivalent to the gradient flow on $\frac{1}{\alpha^2}R(h(\boldsymbol{w}))$ with $\boldsymbol{w}_\alpha(0) = \alpha^{1/m}\boldsymbol{w}_0$.

Under suitable conditions on the model h and the loss function R, [2] proves that as $\alpha \to \infty$, the gradient flow of $F_\alpha(\boldsymbol{w})$ approaches that of $\bar{F}_\alpha(\boldsymbol{w})$. This implies that for a neural network that is positive homogeneous its weights, by taking the scale with which we initialize the weights to infinity, then training the model h with gradient flow is equivalent to training the linearized model $\bar{h}$. The specific details of these results from [2] are the primary focus of Section 3.

3 Theoretical Results

Now that we have rigorously defined the linearized model $\bar{h}$ as well as the gradient flow on $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$, we are well-equipped to study the key results from [2] regarding lazy training. In particular, we will characterize the relationship between the gradient flow paths $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$ and $(\bar{\boldsymbol{w}}_\alpha(t))_{t\geq 0}$ as well as the predictor functions $\alpha h(\boldsymbol{w})$ and $\alpha\bar{h}(\boldsymbol{w})$ evaluated along their respective gradient flow paths as the scale of the model output $\alpha \to \infty$. By way of discussing and proving these theorems, we will gain a deeper understanding of lazy training, particularly as it pertains to neural network optimization.

3.1 Finite-time Bounds

The first result that we consider relates the gradient flow dynamics of $F_\alpha(\boldsymbol{w})$ and those of $\bar{F}_\alpha(\boldsymbol{w})$ in the limit $\alpha \to \infty$ for a finite time horizon. In particular, the result we will prove from [2] demonstrates that at any time $t \geq 0$, the gradient flow of $F_\alpha(\boldsymbol{w})$ at time t, $\boldsymbol{w}_\alpha(t)$, is equivalent to that of $\bar{F}_\alpha(\boldsymbol{w})$ at time t, $\bar{\boldsymbol{w}}_\alpha(t)$, in the $\alpha \to \infty$ limit. Therefore, the $t \to \infty$ limit reached by the gradient flow of F_α, $\lim_{t\to\infty} \boldsymbol{w}_\alpha(t)$, is the same as that reached by the gradient flow of $\bar{F}_\alpha$, $\lim_{t\to\infty} \bar{\boldsymbol{w}}_\alpha(t)$, in the $\alpha \to \infty$ limit. That is to say, we observe lazy training as the scale of the model output $\alpha > 0$ grows large. This gives us our first explicit characterization of when lazy training occurs. We state the relevant theorem and proceed to prove the result:

Theorem 2.2 (from [2])**.** *Assume that $h(\boldsymbol{w}_0) = 0$. Given a fixed time horizon $T > 0$, it holds that $\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\| = \mathcal{O}(1/\alpha)$,*

$$\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\| = \mathcal{O}(1/\alpha^2) \quad \textit{and} \quad \sup_{t\in[0,T]} \|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\| = \mathcal{O}(1/\alpha).$$

Proof. For both this proof and that of Theorem 2.4 in Section 3.2 it will be useful to define $y(t) = \alpha h(\boldsymbol{w}_\alpha(t))$ and $\bar{y}(t) = \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))$ to be the dynamics in $\mathcal{F}$. That is, $y(t)$ is simply the scaled model $\alpha h(\boldsymbol{w})$ evaluated along the gradient flow path of $F_\alpha(\boldsymbol{w})$, $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$, $\boldsymbol{w}_\alpha(0) = \boldsymbol{w}_0$, that we previously discussed.

To be consistent with the notation from [2], we define $\Sigma(\boldsymbol{w}) := Dh(\boldsymbol{w})Dh(\boldsymbol{w})^T$ to be the neural tangent kernel (NTK) at weight vector $\boldsymbol{w} \in \mathbb{R}^p$ [3]. The neural tangent kernel has gained recent popularity in the field of theoretical deep learning due to the fact that in the limit $\alpha \to \infty$, the gradient flow (4) with appropriate model and loss function is no more than a kernel method with kernel given by the NTK. While we do not have sufficient space to flesh out this result, we suggest [3] and [2] as references regarding the neural tangent kernel. From our definition, it is evident that $\Sigma(\boldsymbol{w})$ defines a quadratic form on $\mathcal{F}$ given by $f \mapsto f\Sigma(\boldsymbol{w})f$ for each $f \in \mathcal{F}$. Using the neural tangent kernel $\Sigma(\boldsymbol{w})$, we can say that $y(t)$ and $\bar{y}(t)$ must solve the differential equations

$$\begin{aligned}\frac{d}{dt}y(t) &= \alpha\frac{d}{dt}h(\boldsymbol{w}_\alpha(t)) = \alpha Dh(\boldsymbol{w}_\alpha(t))\frac{d}{dt}\boldsymbol{w}_\alpha(t) = -\alpha Dh(\boldsymbol{w}_\alpha(t))\nabla F_\alpha(\boldsymbol{w}_\alpha(t)) \\ &= -\alpha Dh(\boldsymbol{w}_\alpha(t))\left(\alpha Dh(\boldsymbol{w}_\alpha(t))^T\right)\left(\frac{1}{\alpha^2}\nabla R(\alpha h(\boldsymbol{w}_\alpha(t)))\right)\end{aligned}$$

$$
\begin{aligned}
&= -\Sigma(\boldsymbol{w}_\alpha(t))\nabla R(\alpha h(\boldsymbol{w}_\alpha(t))) \\
&= -\Sigma(\boldsymbol{w}_\alpha(t))\nabla R(y(t)). \\
\frac{d}{dt}\bar{y}(t) &= \alpha\frac{d}{dt}\bar{h}(\bar{\boldsymbol{w}}_\alpha(t)) = \alpha D\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\frac{d}{dt}\bar{\boldsymbol{w}}_\alpha(t) = \alpha D\bar{h}(\boldsymbol{w}_\alpha(0))\frac{d}{dt}\bar{\boldsymbol{w}}_\alpha(t) = \alpha D\bar{h}(\boldsymbol{w}_\alpha(0))\nabla\bar{F}_\alpha(\bar{\boldsymbol{w}}_\alpha(t)) \\
&= \alpha D\bar{h}(\boldsymbol{w}_\alpha(0))\left(\alpha D\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))^T\right)\left(\frac{1}{\alpha^2}\nabla R(\alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t)))\right) \\
&= -\Sigma(\boldsymbol{w}_\alpha(0))\nabla R(\bar{\boldsymbol{w}}_\alpha(t)) \\
&= -\Sigma(\boldsymbol{w}_\alpha(0))\nabla R(\bar{y}(t))
\end{aligned}
$$

with initial condition $y(0) = \bar{y}(0) = \alpha h(\boldsymbol{w}_0)$. Here, we employ the chain rule since, by our assumptions, $h : \mathbb{R}^p \to \mathcal{F}$ and $R : \mathcal{F} \to \mathbb{R}_+$ are everywhere differentiable on their domains. Besides the chain rule, the main result that we use in these two derivations is that $\boldsymbol{w}(t)$ and $\bar{\boldsymbol{w}}(t)$ evolve according to the gradient flow dynamics (4). Additionally, from the definition of the linearized model $\bar{h}$, we rewrite

$$
\begin{aligned}
D\bar{h}(\bar{\boldsymbol{w}}_\alpha(t)) &= D\bigg(h(\bar{\boldsymbol{w}}_\alpha(0)) + Dh(\bar{\boldsymbol{w}}_\alpha(0))(\bar{\boldsymbol{w}}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(0))\bigg) \\
&= Dh(\bar{\boldsymbol{w}}_\alpha(0)) = Dh(\boldsymbol{w}_\alpha(0)).
\end{aligned}
$$

That is, $\bar{h}$ is an affine model whose derivative at all input vectors $\boldsymbol{w} \in \mathbb{R}^p$ is equal to the derivative of h at its initialization $\bar{\boldsymbol{w}}_\alpha(0) = \boldsymbol{w}_0$. Now that we have described $y(t)$ and $\bar{y}(t)$ as well as the differential equations that they must satisfy, we are prepared to proceed with our proof.

Accordingly, let $T > 0$ be an arbitrary time horizon for the gradient flow of $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$. We will first tackle the statement $\sup_{t\in[0,T]}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$. This result will give us a bound on how far the gradient flow path $\boldsymbol{w}_\alpha(t)$ moves from its initialization $\boldsymbol{w}_\alpha(0)$ on the interval $[0, T]$. In fact, it will tell us that in the limit $\alpha \to \infty$, the gradient flow path on $F_\alpha(\boldsymbol{w})$ at any time $t \geq 0$, $\boldsymbol{w}_\alpha(t)$, remains fixed at the initialization $\boldsymbol{w}_\alpha(0)$. This provides another characterization of lazy training that we have not yet discussed: lazy training is truly “lazy” in the sense that the gradient flow path remains close to its initialization.

First, by the Fundamental Theorem of Calculus and properties of the integral it holds that for each $t \in [0, T]$,

$$
\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_\alpha(0)\|_2 = \left\|\int_0^t \boldsymbol{w}'_\alpha(s)\ ds\right\|_2 \leq \int_0^t \|\boldsymbol{w}'_\alpha(s)\|_2\ ds \leq \int_0^T \|\boldsymbol{w}'_\alpha(s)\|_2\ ds.
$$

Note that $\boldsymbol{w}'_\alpha(\cdot) : \mathbb{R}_+ \to \mathbb{R}^p$, and so the integral is defined component-wise:

$$
\int_0^t \boldsymbol{w}'_\alpha(s)\ ds := \left(\int_0^t (\boldsymbol{w}_\alpha)'_1(s)\ ds, \ldots, \int_0^t (\boldsymbol{w}_\alpha)'_p(s)\ ds\right) \in \mathbb{R}^p.
$$

Thus, in order to determine a bound on $\sup_{t\in[0,T]}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2$, it suffices to bound the right-hand expression. In particular, we have

$$
\begin{aligned}
\sup_{t\in[0,T]}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 &\leq \int_0^T \|\boldsymbol{w}'_\alpha(s)\|_2\ ds \\
&= \int_0^T \|\nabla F_\alpha(\boldsymbol{w}_\alpha(s))\|_2\ ds && \text{definition of gradient flow (4)} \\
&= \int_0^T \mathbb{1}\cdot\|\nabla F_\alpha(\boldsymbol{w}_\alpha(s))\|_2\ ds \\
&\leq \sqrt{T}\left(\int_0^T \|\nabla F_\alpha(\boldsymbol{w}_\alpha(s))\|_2^2\ ds\right)^{1/2}. && \text{Cauchy-Schwarz for } L^2([0,T])
\end{aligned}
$$

In order to invoke Cauchy-Schwarz in the final line, we must have $\|\boldsymbol{w}'_\alpha(t)\|_2 \in L^2([0, T])$. This is true because each of $\boldsymbol{w}'_\alpha(\cdot) : \mathbb{R}_+ \to \mathbb{R}^p$ and $\|\cdot\|_2 : \mathbb{R}^p \to \mathbb{R}_+$ is continuous. Accordingly, is true that $\|\boldsymbol{w}'_\alpha(t)\|_2$ is continuous on the closed interval $[0, T]$, and so it belongs to $L^2([0, T])$.

Now to simplify the integrand, we use the fact that

$$\frac{d}{dt}F_\alpha(\boldsymbol{w}_\alpha(t)) = \nabla F_\alpha(\boldsymbol{w}_\alpha(t))^T \boldsymbol{w}_\alpha'(t) = \nabla F_\alpha(\boldsymbol{w}_\alpha(t))^T(-\nabla F_\alpha(\boldsymbol{w}_\alpha(t))) = -\|\nabla F_\alpha(\boldsymbol{w}_\alpha(t))\|_2^2.$$

This follows from a straightforward application of the chain rule as well as the knowledge that $\boldsymbol{w}_\alpha(t)$ evolves according to the gradient flow dynamics on the scaled objective function $F_\alpha(\boldsymbol{w})$, (4).

Substituting this expression for $\|\nabla F_\alpha(\boldsymbol{w}_\alpha(t))\|_2^2$ back into the integral, we get

$$\begin{aligned}\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2 &\le \sqrt{T}\left(\int_0^T \|\nabla F_\alpha(\boldsymbol{w}_\alpha(s))\|_2^2\, ds\right)^{1/2} \\ &= \sqrt{T}\left(\int_0^T -\frac{d}{ds}F_\alpha(\boldsymbol{w}_\alpha(s))\, ds\right)^{1/2} \\ &= \sqrt{T}\left(F_\alpha(\boldsymbol{w}_\alpha(0)) - F_\alpha(\boldsymbol{w}_\alpha(T))\right)^{1/2}. \qquad \text{Fundamental Theorem of Calculus}\end{aligned}$$

And since the loss function $R : \mathcal{F} \to \mathbb{R}_+$, then we get

$$\begin{aligned}\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2 &\le \sqrt{T}\left(F_\alpha(\boldsymbol{w}_\alpha(0)) - F_\alpha(\boldsymbol{w}_\alpha(T))\right)^{1/2} \\ &= \sqrt{T}\left(\frac{1}{\alpha^2}\Big(R(\alpha h(\boldsymbol{w}_\alpha(0))) - R(\alpha h(\boldsymbol{w}_\alpha(T)))\Big)\right)^{1/2} \\ &\le \sqrt{T}\left(\frac{1}{\alpha^2}\Big(R(\alpha h(\boldsymbol{w}_\alpha(0)))\Big)\right)^{1/2} \\ &= \frac{1}{\alpha}\Big(T\cdot R(\alpha h(\boldsymbol{w}_\alpha(0)))\Big)^{1/2}.\end{aligned}$$

Therefore, we conclude that for each $T > 0$,

$$\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2 \le \frac{1}{\alpha}\Big(T\cdot R(\alpha h(\boldsymbol{w}_\alpha(0)))\Big)^{1/2} = \mathcal{O}(1/\alpha),$$

as we wished to prove. Notice that, although the $\mathcal{O}(1/\alpha)$ hides the dependence on T, our bound on $\sup_{t\in[0,T]}\|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2$ grows sublinearly in T. In order to achieve a bound which does not depend on this finite time horizon T, we will need the stronger assumptions on h and R that appear in Theorem 2.4.

As a consequence of this bound on $\sup_{t\in[0,T]}\|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2$, we get a couple additional results that will be useful throughout the remainder of our proof.

First, for $y(t)$ defined as in Section 2, we know that

$$\sup_{t\in[0,T]}\|y(t)-y(0)\|_{\mathcal{F}} = \sup_{t\in[0,T]}\|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha h(\boldsymbol{w}_\alpha(0))\|_{\mathcal{F}} = \sup_{t\in[0,T]}\alpha\|h(\boldsymbol{w}_\alpha(t))\|_{\mathcal{F}}.$$

But from the result we just proved, we also know that for every $t \in [0,T]$, $\boldsymbol{w}_\alpha(t) \in B_\epsilon(\boldsymbol{w}_0)$, where $\epsilon = \frac{C}{\alpha}$ for some constant $C \ge 0$. Here $B_\epsilon(\boldsymbol{w}_0)$ denotes the closed Euclidean ball of radius ϵ centered at $\boldsymbol{w}_0$. And since $h : \mathbb{R}^p \to \mathcal{F}$ is continuous by assumption, as is $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}_+$, then the composition $\boldsymbol{w} \mapsto \|h(\boldsymbol{w})\|_{\mathcal{F}}$ is continuous on $\mathbb{R}^p$. Altogether, since $\boldsymbol{w} \mapsto \|h(\boldsymbol{w})\|_{\mathcal{F}}$ is continuous on the compact set $B_\epsilon(\boldsymbol{w}_0)$, then by the Weierstrauss Extreme Value Theorem, $\|h(\boldsymbol{w})\|_{\mathcal{F}} \le C$ for every $\boldsymbol{w} \in B_\epsilon(\boldsymbol{w}_0)$, for some fixed $C \ge 0$. In particular, this implies that $\|h(\boldsymbol{w}_\alpha(t))\|_{\mathcal{F}} \le C$ for every $t \in [0,T]$. Thus, we conclude

$$\sup_{t\in[0,T]}\|y(t)-y(0)\|_{\mathcal{F}} = \sup_{t\in[0,T]}\alpha\|h(\boldsymbol{w}_\alpha(t))\|_{\mathcal{F}} \le C\alpha.$$

By a similar argument, we can show that $\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t)-\boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$ implies $\sup_{t\in[0,T]} \|\nabla R(y(t))\|_{\mathcal{F}} \leq C$ for some constant $C \geq 0$. Specifically, we have assumed that that the loss function R has a Lipschitz gradient, meaning that the map $f \mapsto \nabla R(f),\ f \in \mathcal{F}$ is Lipschitz. And $f \mapsto \nabla R(f)$ Lipschitz implies $f \mapsto \nabla R(f)$ continuous. Also, we know that each of $h : \mathbb{R}^p \to \mathcal{F}$ and $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}_+$ is continuous, and so altogether the composition $\|\nabla R(\alpha h(\boldsymbol{w}))\|_{\mathcal{F}} : \mathbb{R}^p \to \mathbb{R}_+$ is continuous. Therefore, we can apply the same Weierstrauss Extreme Value result as in the previous paragraph to say that for every $\boldsymbol{w} \in B_\epsilon(\boldsymbol{w}_0)$ it holds that $\|\nabla R(\alpha h(\boldsymbol{w}))\|_{\mathcal{F}} \leq C$ for some fixed constant $C \geq 0$. $\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$ gives us that $\boldsymbol{w}_\alpha(t)$ is in the closed ball $B_\epsilon(\boldsymbol{w}_0)$ for every $t \in [0,T]$ with appropriate choice of $\epsilon \geq 0$. Therefore, we conclude

$$\sup_{t\in[0,T]} \|\nabla R(y(t))\|_{\mathcal{F}} = \sup_{t\in[0,T]} \|\nabla R(\alpha h(\boldsymbol{w}_\alpha(t)))\|_{\mathcal{F}} \leq C.$$

We continue on by proving the bound $\sup_{t\in[0,T]} \|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\|_{\mathcal{F}} = \mathcal{O}(1/\alpha)$. While the first result established a bound on the distance of between the gradient flow path and its initialization on the interval $[0,T]$, this result will bound the distance between the scaled model αh and its linearized counterpart $\alpha\bar{h}$ evaluated along their respective gradient flow paths $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$ and $(\bar{\boldsymbol{w}}_\alpha(t))_{t\geq 0}$ on $[0,T]$. Consequently, we observe that as $\alpha \to \infty$ the scaled original model αh evaluated at $\boldsymbol{w}_\alpha(t)$ is equivalent to the scaled linearized model $\alpha\bar{h}$ evaluated at $\bar{\boldsymbol{w}}_\alpha(t)$ for any time $t \geq 0$.

To start off our proof, we recall our notation $y(t) = \alpha h(\boldsymbol{w}_\alpha(t))$, $\bar{y}(t) = \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))$ from Section 2. With these functions $y,\ \bar{y}$, we define $\Delta(t) = \|y(t) - \bar{y}(t)\|_{\mathcal{F}},\ \forall t \geq 0$, which is the distance between $y(t)$ and $\bar{y}(t)$ in the Hilbert space $\mathcal{F}$. By the definition of the linearized model $\bar{h}$, we know that Δ satisfies $\Delta(0) = \|y(0) - \bar{y}(0)\|_{\mathcal{F}} = \alpha\|h(\boldsymbol{w}_0) - \bar{h}(\boldsymbol{w}_0)\|_{\mathcal{F}} = \alpha\|h(\boldsymbol{w}_0) - h(\boldsymbol{w}_0)\|_{\mathcal{F}} = 0$. Furthermore, we derive an upper bound on the derivative $\Delta'(t)$. In particular, for each $t > 0$ we have

$$\begin{aligned}
\frac{d}{dt}\left(\Delta(t)^2\right) &= \frac{d}{dt}\|y(t) - \bar{y}(t)\|_{\mathcal{F}}^2 \\
&= \frac{d}{dt}\langle y(t) - \bar{y}(t), y(t) - \bar{y}(t)\rangle_{\mathcal{F}} \\
&= 2\langle y'(t) - \bar{y}'(t), y(t) - \bar{y}(t)\rangle_{\mathcal{F}} \\
&\leq 2\|y'(t) - \bar{y}'(t)\|_{\mathcal{F}}\|y(t) - \bar{y}(t)\|_{\mathcal{F}} && \text{Cauchy–Schwarz in } \mathcal{F} \\
&= 2\Delta(t)\|y'(t) - \bar{y}'(t)\|_{\mathcal{F}}
\end{aligned}$$

But by the chain rule we also know

$$\frac{d}{dt}(\Delta(t)^2) = 2\Delta(t)\Delta'(t),$$

and so the above result implies that $\forall t > 0$,

$$\begin{aligned}
&2\Delta(t)\Delta'(t) \leq 2\Delta(t)\|y'(t) - \bar{y}'(t)\|_{\mathcal{F}} \\
\implies &\Delta'(t) \leq \|y'(t) - \bar{y}'(t)\|_{\mathcal{F}}.
\end{aligned}$$

Recall the explicit expressions for $y'(t)$ and $\bar{y}'(t),\ t > 0$ that we derived at the beginning of our proof. Substituting them into the bound on $\Delta'(t)$, we get

$$\begin{aligned}
\Delta'(t) &\leq \|y'(t) - \bar{y}'(t)\|_{\mathcal{F}} \\
&= \|\Sigma(\boldsymbol{w}_\alpha(t))\nabla R(y(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\nabla R(\bar{y}(t))\|_{\mathcal{F}} \\
&\leq \|\Sigma(\boldsymbol{w}_\alpha(t))\nabla R(y(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\nabla R(y(t))\|_{\mathcal{F}} + \|\Sigma(\boldsymbol{w}_\alpha(0))\nabla R(y(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\nabla R(\bar{y}(t))\|_{\mathcal{F}} \\
&= \|(\Sigma(\boldsymbol{w}_\alpha(t)) - \Sigma(\boldsymbol{w}_\alpha(0)))\nabla R(y(t))\|_{\mathcal{F}} + \|\Sigma(\boldsymbol{w}_\alpha(0))(\nabla R(y(t)) - \nabla R(\bar{y}(t)))\|_{\mathcal{F}}.
\end{aligned}$$

The second inequality is achieved by adding and subtracting a term of $\Sigma(\boldsymbol{w}_\alpha(0))\nabla R(y(t))$ and subsequently applying the triangle inequality for the norm $\mathcal{F}$. Next, we will invoke the properties of the operator norm, where our operator $f \mapsto \Sigma(\boldsymbol{w})f$ maps from normed vector space $\mathcal{F}$ to itself. In particular, since $Dh(\boldsymbol{w}) : \mathbb{R}^p \to \mathcal{F}$ is continuous and linear (for each $\boldsymbol{w} \in \mathbb{R}^p$), then so is $f \mapsto \Sigma(\boldsymbol{w})f$, where $\Sigma(\boldsymbol{w}) = Dh(\boldsymbol{w})Dh(\boldsymbol{w})^T$. As a result, we have get that for each $f \in \mathcal{F}$, $\|\Sigma(\boldsymbol{w})f\| \leq \|\Sigma(\boldsymbol{w})\|\|f\|_{\mathcal{F}}$, where $\|\cdot\|$ denotes the operator norm. Applying this inequality to the above expression, we get

$$\Delta'(t) \leq \|\Sigma(\boldsymbol{w}_\alpha(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t))\|_{\mathcal{F}} + \|\Sigma(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}.$$

From here, we bound each of the two terms separately, starting with the first term. To consider the factor $\|\Sigma(\boldsymbol{w}_\alpha(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\|$, we cite the result from [2] which states that $\text{Lip}(\Sigma) \leq 2\text{Lip}(h)\text{Lip}(Dh)$. Note that $\text{Lip}(\Sigma)$ is defined with respect to the operator norm. From the first result we proved, we know that we are dealing with $\boldsymbol{w}_\alpha(t)$ contained in a closed Euclidean ball $B_\epsilon(\boldsymbol{w}_0)$ with some radius $\epsilon \geq 0$. And so Dh locally Lipschitz (our assumption) on compact set $B_\epsilon(\boldsymbol{w}_0)$ implies Dh Lipschitz on $B_\epsilon(\boldsymbol{w}_0)$. Also, Dh continuous on $B_\epsilon(\boldsymbol{w}_0)$ implies that h is Lipschitz on $B_\epsilon(\boldsymbol{w}_0)$. Letting $\text{Lip}(Dh)$ and $\text{Lip}(h)$ be the Lipschitz constants of Dh and h on $B_\epsilon(\boldsymbol{w}_0)$, respectively, we get the desired bound on $\text{Lip}(\Sigma)$. Invoking the first result that we proved, we get $\|\Sigma(\boldsymbol{w}_\alpha(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\| \leq 2 \cdot \text{Lip}(h) \cdot \text{Lip}(Dh)\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_\alpha(0)\|_2 \leq 2 \cdot \text{Lip}(h) \cdot \text{Lip}(Dh) \cdot C/\alpha$ for some constant $C \geq 0$. As for the factor $\|\nabla R(y(t))\|_{\mathcal{F}}$, we recall the result we previously proved that $\sup_{\tilde{t} \in [0,T]} \|\nabla R(y(\tilde{t}))\|_{\mathcal{F}} \leq \tilde{C}$ for some constant $\tilde{C} \geq 0$. And so, in all, we have proven that $\|\Sigma(\boldsymbol{w}_\alpha(t)) - \Sigma(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t))\|_{\mathcal{F}} \leq C_1/\alpha$ for some constant $C_1 \geq 0$.

As for the second term, we call upon our assumption that the loss function R has a Lipschitz gradient to say that $\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \leq \text{Lip}(\nabla R)\|y(t) - \bar{y}(t)\|_{\mathcal{F}} = \text{Lip}(\nabla R)\Delta(t)$ where $\text{Lip}(\nabla R) \geq 0$ denotes the Lipschitz constant of $f \mapsto \nabla R(f)$. Accordingly, we have $\|\Sigma(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \leq C_2\Delta(t)$ for some constant $C_2 \geq 0$.

Altogether, we have shown that

$$\Delta'(t) \leq C_1/\alpha + C_2\Delta(t), \quad \Delta(0) = 0$$

for suitable constants $C_1, C_2 \geq 0$. We notice, though, that the equation $u'(t) = C_1/\alpha + C_2u(t)$ with initial condition $u(0) = 0$ defines a first-order, linear differential equation. This equation has a unique solution that exists on all of $\mathbb{R}_+$, which we determine using an integrating factor:

$$\begin{aligned}
&u'(t) = C_1/\alpha + C_2u(t) \\
&u'(t) - C_2u(t) = C_1/\alpha \\
&\exp(-C_2t)u(t) = \int^t C_1/\alpha \exp(-C_2s)\ ds + C \\
&u(t) = -C_1/(C_2\alpha) + C\exp(C_2t) \\
&u(t) = \frac{C_1}{C_2\alpha}(\exp(C_2t) - 1). && u(0) = 0
\end{aligned}$$

But since $\Delta'(t) \leq u'(t)$ for all times $t \geq 0$ and $\Delta(0) = u(0) = 0$, then it holds that $\Delta(t) \leq u(t)$ for all times $t \geq 0$. That is, the curve $\Delta(t)$ lies strictly below the solution curve $u(t)$ to our differential equation $u'(t) = C_1/\alpha + C_2u(t)$, $u(0) = 0$.

And so we have proven

$$\|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\|_{\mathcal{F}} = \Delta(t) \leq \frac{C_1}{C_2\alpha}(\exp(C_2t) - 1) \leq \frac{C_1}{C_2\alpha}(\exp(C_2T) - 1). \quad \forall t \in [0, T]$$

Therefore, we conclude

$$\sup_{t \in [0,T]} \|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\|_{\mathcal{F}} \leq \frac{1}{\alpha}\left(\frac{C_1}{C_2}(\exp(C_2T) - 1)\right) = \mathcal{O}(1/\alpha).$$

One will observe that the resulting bound is worse than than that we derived on $\sup_{t \in [0,T]} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2$, as there is an exponential dependence on the finite time horizon T. That is, fixing some initialization scale $\alpha > 0$, our upper bound on $\sup_{t \in [0,T]} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2$ grows exponentially as a function of T.

The final bound we would like to prove is that on the distance between the gradient flow paths of $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$, $\sup_{t\in[0,T]} \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\| = \mathcal{O}(1/\alpha^2)$. This bound tells us that in the limit $\alpha \to \infty$, the gradient flow of $F_\alpha(\boldsymbol{w})$ is equivalent to that of $\bar{F}_\alpha(\boldsymbol{w})$ at any time $t \geq 0$, where $\boldsymbol{w}_\alpha(0) = \bar{\boldsymbol{w}}_\alpha(0) = \boldsymbol{w}_0$.

Analogous to the function $\Delta(t) : \mathbb{R}_+ \to \mathbb{R}_+$ in the previous portion of our proof, we start by defining $\delta(t) := \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2$ for each $t \geq 0$. We approach the problem of bounding $\delta(t)$ on the interval $[0, T]$ by deriving a bound on $\delta'(t)$.

Our first step in finding a bound on $\delta'(t)$ is very similar to that used in the previous portion of our proof. In particular, we know that $\boldsymbol{w}_\alpha(0) = \bar{\boldsymbol{w}}_\alpha(0) = \boldsymbol{w}_0$, and so $\delta(0) = 0$. Moreover, from our computations of $y'(t)$ and $\bar{y}'(t)$ in Section 2, we know that $\boldsymbol{w}'_\alpha(t) = -\nabla F_\alpha(\boldsymbol{w}_\alpha(t)) = -\frac{1}{\alpha} Dh(\boldsymbol{w}_\alpha(t))^T \nabla R(y(t))$ as well as $\bar{\boldsymbol{w}}'_\alpha(t) = -\nabla \bar{F}_\alpha(\bar{\boldsymbol{w}}_\alpha(t)) = -\frac{1}{\alpha} Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(\bar{y}(t))$ for every $t \geq 0$. As a result, we get the following bound on $\delta'(t)$ for each $t \geq 0$:

$$\begin{aligned}
\delta'(t) =& \frac{d}{dt} \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 \\
=& \frac{d}{dt} \left\| \int_0^t \boldsymbol{w}'_\alpha(s) - \bar{\boldsymbol{w}}'_\alpha(s)\ ds \right\|_2 && \text{Fundamental Theorem of Calculus} \\
\leq& \frac{d}{dt} \int_0^t \|\boldsymbol{w}'_\alpha(s) - \bar{\boldsymbol{w}}'_\alpha(s)\|_2\ ds \\
=& \|\boldsymbol{w}'_\alpha(t) - \bar{\boldsymbol{w}}'_\alpha(t)\|_2. && \text{Fundamental Theorem of Calculus}
\end{aligned}$$

Substituting in our particular expressions for $\boldsymbol{w}'_\alpha(t)$ and $\bar{\boldsymbol{w}}'_\alpha(t)$,

$$\begin{aligned}
\delta'(t) \leq& \|\boldsymbol{w}'_\alpha(t) - \bar{\boldsymbol{w}}'_\alpha(t)\|_2 = \frac{1}{\alpha} \|Dh(\boldsymbol{w}_\alpha(t))^T \nabla R(y(t)) - Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(\bar{y}(t))\|_2 \\
\leq& \frac{1}{\alpha} \|Dh(\boldsymbol{w}_\alpha(t))^T \nabla R(y(t)) - Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(y(t))\|_2 \\
+& \|Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(y(t)) - Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(\bar{y}(t))\|_2 \\
=& \frac{1}{\alpha} \Big(\|(Dh(\boldsymbol{w}_\alpha(t))^T - Dh(\boldsymbol{w}_\alpha(0))^T) \nabla R(y(t))\|_2 + \|Dh(\boldsymbol{w}_\alpha(0))^T (\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_2 \Big)
\end{aligned}$$

The inequality on the second line follows from adding and subtracting a term of $Dh(\boldsymbol{w}_\alpha(0))^T \nabla R(y(t))$ and then invoking the triangle inequality for the ℓ^2 norm. In order to further bound $\delta'(t)$, we use the fact that $Dh(\boldsymbol{w}) : \mathbb{R}^p \to \mathcal{F}$ is a continuous, linear operator for each $\boldsymbol{w} \in \mathbb{R}^p$, and thus so is its adjoint $Dh(\boldsymbol{w})^T$, where both $\mathbb{R}^p$ and $\mathcal{F}$ are normed vector spaces. Consequently, we have that for each $f \in \mathcal{F}$, $\|Dh(\boldsymbol{w})^T f\|_2 \leq \|Dh(\boldsymbol{w})^T\| \|f\|_{\mathcal{F}}$, where $\|Dh(\boldsymbol{w})^T\|$ denotes the operator norm of $Dh(\boldsymbol{w})^T$. Also, we will use the fact that for Dh a continuous, linear operator, then $\|Dh(\boldsymbol{w})\| = \|Dh(\boldsymbol{w})^T\|$. Putting all of these pieces together, we have

$$\begin{aligned}
\delta'(t) \leq& \frac{1}{\alpha} \Big(\|Dh(\boldsymbol{w}_\alpha(t))^T - Dh(\boldsymbol{w}_\alpha(0))^T\| \|\nabla R(y(t))\|_{\mathcal{F}} + \|Dh(\boldsymbol{w}_\alpha(0))^T\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \Big) \\
=& \frac{1}{\alpha} \Big(\|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t))\|_{\mathcal{F}} + \|Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \Big).
\end{aligned}$$

From here, we bound each of the two terms separately. Starting with $\|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_\alpha(0))\| \nabla R(y(t))\|_{\mathcal{F}}$, we recall our assumption that the map $\boldsymbol{w} \mapsto Dh(\boldsymbol{w})$ is locally Lipschitz. And from the first result we also know that $\boldsymbol{w}_\alpha(t)$ is contained in some closed Euclidean ball centered at $\boldsymbol{w}_\alpha(0)$ (that is, $\boldsymbol{w}_\alpha(t) \in B_\epsilon(\boldsymbol{w}_0)$, $\forall t \in [0, T]$ for appropriate choice of $\epsilon \geq 0$). Therefore, we have that $\boldsymbol{w} \mapsto Dh(\boldsymbol{w})$ is Lipschitz on the compact set $B_\epsilon(\boldsymbol{w}_0)$, and so $\|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t))\|_{\mathcal{F}} \leq \text{Lip}(Dh) \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_\alpha(0)\|_2 \|\nabla R(y(t))\|_{\mathcal{F}}$, where $\text{Lip}(Dh)$ denotes the Lipschitz constant of Dh on $B_\epsilon(\boldsymbol{w}_0)$. Also from the first result, we know that $\sup_{\tilde{t}\in[0,T]} \|\boldsymbol{w}_\alpha(\tilde{t}) - \boldsymbol{w}_\alpha(0)\| \leq C_1/\alpha$ for some constant $C_1 \in \mathbb{R}_+$. Similarly, we previously showed that $\sup_{\tilde{t}\in[0,T]} \|\nabla R(y(\tilde{t}))\|_{\mathcal{F}} \leq C_2$ for some $C_2 \in \mathbb{R}_+$. Altogether, we have $\|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t))\|_{\mathcal{F}} \leq C_1 \cdot C_2 \cdot \text{Lip}(Dh)/\alpha$.

And for the second term $\|Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}$, we recall our assumption that the gradient of R, $f \mapsto \nabla R(f)$, is Lipschitz. Therefore, we have $\|Dh(\boldsymbol{w}_\alpha(0))\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \leq \text{Lip}(\nabla R) \cdot$

$\|Dh(\boldsymbol{w}_\alpha(0))\|\|y(t) - \bar{y}(t)\| \leq \text{Lip}(\nabla R) \cdot C_1 \|y(t) - \bar{y}(t)\|$ for some constant $C_1 \geq 0$. Additionally, from the second result we proved, we know that $\sup_{\tilde{t} \in [0,T]} \|\alpha h(\boldsymbol{w}_\alpha(\tilde{t})) - \alpha \bar{h}(\bar{\boldsymbol{w}}_\alpha(\tilde{t}))\| = \sup_{\tilde{t} \in [0,T]} \|y(\tilde{t}) - \bar{y}(\tilde{t})\| \leq C_2/\alpha$ for some constant $C_2 \geq 0$. And so we have shown $\|Dh(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \leq C_1 \cdot C_2 \cdot \text{Lip}(\nabla R)/\alpha$.

Combining these two bounds, we have

$$\begin{aligned}
\delta'(t) &\leq \frac{1}{\alpha}\Big(\|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t))\|_{\mathcal{F}} + \|Dh(\boldsymbol{w}_\alpha(0))\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \Big) \\
&\leq C/\alpha^2
\end{aligned}$$

for some constant $C \geq 0$, for each $t \geq 0$. Thus, we conclude that $\sup_{t \in [0,T]} \delta'(t) \leq C/\alpha^2$. And by our previous justification that $\delta(0) = 0$, then it holds that for each $t \in [0, T]$,

$$\begin{aligned}
\delta(t) &= \int_0^t \delta'(s)\ ds && \text{Fundamental Theorem of Calculus} \\
&\leq \int_0^t \sup_{\tilde{t} \in [0,T]} \delta'(\tilde{t})\ ds \\
&= t \sup_{\tilde{t} \in [0,T]} \delta'(\tilde{t}) \\
&\leq T \sup_{\tilde{t} \in [0,T]} \delta'(\tilde{t}) \leq T \cdot C/\alpha^2.
\end{aligned}$$

This gives us our desired result

$$\sup_{t \in [0,T]} \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 = \sup_{t \in [0,T]} \delta(t) \leq T \cdot C/\alpha^2 = \mathcal{O}(1/\alpha^2).$$

We have demonstrated each of the three bounds stated in Theorem 2.2, and so we conclude our proof. □

So far, we have given a general characterization of lazy training that, while beneficial in the theoretical sense, may be futile in practice. To summarize our results from Theorem 2.2, we have shown that at any time $t \geq 0$, the gradient flow path of $F_\alpha(\boldsymbol{w})$ at time t is equivalent to that of $\bar{F}_\alpha(\boldsymbol{w})$ at time t as $\alpha \to \infty$. Likewise, we demonstrated that for each $t \geq 0$, the original scaled model αh, which is not a priori convex in $\boldsymbol{w} \in \mathbb{R}^p$, evaluated at $\boldsymbol{w}_\alpha(t)$ is equivalent to the linearized scaled model $\alpha \bar{h}$ evaluated at $\bar{\boldsymbol{w}}_\alpha(t)$ as $\alpha \to \infty$. Ultimately, these statements tell us that in the $\alpha \to \infty$ limit, the limit reached by gradient flow on F_α and $\bar{F}_\alpha$ are equivalent, $\lim_{t\to\infty} \boldsymbol{w}_\alpha(t) = \lim_{t\to\infty} \bar{\boldsymbol{w}}_\alpha(t)$, as are the models αh and $\alpha \bar{h}$ evaluated at $\lim_{t\to\infty} \boldsymbol{w}_\alpha(t) = \lim_{t\to\infty} \bar{\boldsymbol{w}}_\alpha(t)$. And so we have shown that lazy training as we presented it in Section 1 occurs as the factor by which we are scaling the model output grows to infinity. If the model h is positive homogeneous, we have equivalently shown that lazy training occurs when the scale of the initialization $\boldsymbol{w}_\alpha(0) = \alpha \boldsymbol{w}_0$ grows to infinity.

Why this result is poorly suited for practical applications, though, is due to the dependence of our bounds on the time horizon T. Expressly, for large times t we would need a very large initialization scale $\alpha > 0$ to see the convergence of $\boldsymbol{w}_\alpha(t)$ to $\bar{\boldsymbol{w}}_\alpha(t)$ and $\alpha h(\boldsymbol{w}_\alpha(t))$ to $\alpha \bar{h}(\bar{\boldsymbol{w}}_\alpha(t))$ in the respective norms within some small threshold $\epsilon > 0$. And since to approximate the limit reached by gradient flow one must consider large t, this makes comparing the gradient flow limits of $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$ onerous. We address this problem with Theorem 2.4, which, by making stronger regularity assumptions on the model h and loss R, extends the bounds we proved in Theorem 2.2 to be uniform in time $t \geq 0$.

3.1.1 Model Generalization

Before continuing on to the uniform time case, we return to the example described in Section 2 where the model h maps each weight vector $\boldsymbol{w} \in \mathbb{R}^p$ to a network function $f(\boldsymbol{w}, \cdot) \in \mathcal{F}$. Here, we suppose that $f(\boldsymbol{w},\cdot):\mathbb{R}^d\to\mathbb{R}^k$; notice that this is different from our discussion in Section 2 where the output of $f(\boldsymbol{w},\cdot)$ was one-dimensional. Also, suppose that we are given some training data $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$, where each $\boldsymbol{x}_i\in\mathbb{R}^d$, $\boldsymbol{y}_i\in\mathbb{R}^k$. Then by Theorem 2.2, we would expect that for each $t\geq 0$ in the limit as $\alpha\to\infty$, $\|\alpha f(\boldsymbol{w}_\alpha(t),\boldsymbol{x}_i)-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(t),\boldsymbol{x}_i)\|_2$ is small for $i=1,\ldots,N$. That is, the scaled original model evaluated at the training input points $\boldsymbol{x}_i$ along its gradient flow path should be equal to the scaled linearized model evaluated at the same points along its gradient flow path. However, it is unclear whether or not the scaled model $\alpha f(\boldsymbol{w}_\alpha(t),\cdot)$ generalizes like the scaled linearized model $\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(t),\cdot)$. That is, we would like to know whether $\|\alpha f(\boldsymbol{w}_\alpha(t),\boldsymbol{x}')-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(t),\boldsymbol{x}')\|_2$ is small for $\boldsymbol{x}'\notin\{\boldsymbol{x}_i\}_{i=1}^N$.

In Proposition A.1, Chizat and colleagues address this question and show that on a certain subset of the input space $\mathcal{X}\subset\mathbb{R}^d$, $\alpha f(\boldsymbol{w}_\alpha(t),\cdot)$ indeed generalizes like the linearized model $\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(t),\cdot)$. We state the authors' proposition and then proceed to prove their result.

Proposition A.1. *Assume that the results of Theorem 2.2 hold. In particular, for some constants $C_1,C_2>0$ it holds that $\|\boldsymbol{w}_\alpha(T)-\bar{\boldsymbol{w}}_\alpha(T)\|_2\leq C_1\log(\alpha)/\alpha^2$ as well $\|\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0\|_2\leq C_2\log(\alpha)/\alpha$. Assume moreover that there exists a set $\mathcal{X}\subset\mathbb{R}^d$ such that $M_1:=\sup_{\boldsymbol{x}\in\mathcal{X}}\|D_{\boldsymbol{w}}f(\boldsymbol{w}_0,\boldsymbol{x})\|<\infty$ and $M_2:=\sup_{\boldsymbol{x}\in\mathcal{X}}\mathit{Lip}(\boldsymbol{w}\mapsto D_{\boldsymbol{w}}f(\boldsymbol{w},\boldsymbol{x}))<\infty$. Then it holds*

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})\|_2\leq\frac{\log(\alpha)}{\alpha}\left(C_1\cdot M_1+\frac{1}{2}C_2^2\cdot M_2\cdot\log(\alpha)\right)\longrightarrow 0\quad\text{as }\alpha\longrightarrow\infty.$$

Proof. To start, we clarify that, unlike in Theorem 2.2, the distance between $\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})$ and $\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})$ is measured in the ℓ^2 norm for $\mathbb{R}^k$, not the Hilbert space $\mathcal{F}$ norm, since the functions $f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})$ and $\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})$ are evaluated at a particular input $\boldsymbol{x}\in\mathcal{X}$. For the same reason, $D_{\boldsymbol{w}}f(\boldsymbol{w}_0,\boldsymbol{x})\in\mathbb{R}^{k\times p}$ is a matrix rather than a function $D_{\boldsymbol{w}}h(\boldsymbol{w}_0):\mathbb{R}^p\to\mathcal{F}$, and so $\|D_{\boldsymbol{w}}f(\boldsymbol{w}_0,\boldsymbol{x})\|$ is taken with respect to the matrix norm $\|\cdot\|_{k,p}$.

Now that we cleared up the statement of the proposition, we appeal to the properties of the supremum to split the quantity that we wish to bound into two ancillary quantities:

$$\begin{aligned}&\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})\|_2\\ &\leq\sup_{\boldsymbol{x}\in\mathcal{X}}\left(\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2+\|\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})\|_2\right)\qquad\text{triangle inequality}\\ &\leq\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2+\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\bar{\boldsymbol{w}}_\alpha(T),\boldsymbol{x})\|_2.\end{aligned}$$

And so we see that it suffices to bound each term separately.

Let us start with the first term $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2$. One will recall from Section 2 that $\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})=f(\boldsymbol{w}_0)+D_{\boldsymbol{w}}f(\boldsymbol{w}_0,\boldsymbol{x})(\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0)$ is simply equal to the first-order approximation of $f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})$ about $\boldsymbol{w}=\boldsymbol{w}_0$. Therefore, for each fixed $\boldsymbol{x}\in\mathcal{X}$, writing $f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})=f(\boldsymbol{w}_0)+D_{\boldsymbol{w}}f(\boldsymbol{w}_0,\boldsymbol{x})(\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0)+\mathcal{R}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})$, we have that $\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2=\alpha\|\mathcal{R}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2$, where $\mathcal{R}(\cdot,\boldsymbol{x}):\mathbb{R}^p\to\mathbb{R}^k$ is the Taylor remainder of our approximation $\bar{f}$. And so we see that if we can bound the norm of the remainder term $\mathcal{R}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})$ for $\boldsymbol{x}\in\mathcal{X}$, then we have a bound on $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\alpha f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-\alpha\bar{f}(\boldsymbol{w}_\alpha(T),\boldsymbol{x})\|_2$.

In order to bound this remainder term, let us define the function $g:\mathbb{R}\to\mathbb{R}^k$ such that $g(t)=f(\boldsymbol{w}_0+t(\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0),\boldsymbol{x})$. Note that since $\boldsymbol{w}\mapsto f(\boldsymbol{w},\boldsymbol{x})$ is differentiable, by assumption, then $g(t)$ is differentiable in $t\in\mathbb{R}$. And so by the Fundamental Theorem of Calculus, we have

$$\begin{aligned}f(\boldsymbol{w}_\alpha(T),\boldsymbol{x})-f(\boldsymbol{w}_0,\boldsymbol{x})&=g(1)-g(0)\\ &=\int_0^1 g'(t)\;dt\\ &=\int_0^1 D_{\boldsymbol{w}}f(\boldsymbol{w}_0+t(\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0),\boldsymbol{x})(\boldsymbol{w}_\alpha(T)-\boldsymbol{w}_0)\;dt.\end{aligned}$$

Just as in our proof of Theorem 2.2, we remark that since $g : \mathbb{R} \to \mathbb{R}^k$, then the integral is defined componentwise. Now, by adding and subtracting the term $D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)$ in the integrand, we get

$$\begin{aligned}
&f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - f(\boldsymbol{w}_0, \boldsymbol{x}) \\
&= \int_0^1 \Big(D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}) \Big)(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)\ dt + \int_0^1 D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)\ dt \\
&= D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0) + \int_0^1 \Big(D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}) \Big)(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)\ dt.
\end{aligned}$$

However, by subtracting the term $D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)$ over to the left-hand side of the inequality,

$$f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) \leq \int_0^1 \Big(D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}) \Big)(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)\ dt.$$

And so to place a bound on the norm of the Taylor remainder $\mathcal{R}(\boldsymbol{w}_\alpha(T), \boldsymbol{x})$, we must bound the norm of the right-hand side of the prior inequality. Exploiting the properties of the norm, we have

$$\begin{aligned}
\|f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x})\|_2 &\leq \left\| \int_0^1 \Big(D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}) \Big)(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)\ dt \right\|_2 \\
&\leq \int_0^1 \left\| \Big(D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}) \Big)(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0) \right\|_2\ dt \\
&\leq \int_0^1 \|D_{\boldsymbol{w}} f(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0), \boldsymbol{x}) - D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})\| \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2\ dt \\
&\leq \int_0^1 \mathrm{Lip}(\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})) \|(\boldsymbol{w}_0 + t(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)) - \boldsymbol{w}_0\|_2 \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2\ dt \\
&= \mathrm{Lip}(\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})) \int_0^1 t \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2\ dt \\
&= \frac{1}{2} \mathrm{Lip}(\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})) \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2.
\end{aligned}$$

In particular, for the third inequality we invoke the property of the matrix norm $\|A\boldsymbol{v}\|_k \leq \|A\|_{k,p} \|\boldsymbol{v}\|_p$. And for the fourth inequality, we use the result from Theorem 2.2 that the map $\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})$ is Lipschitz on some closed Euclidean ball containing the gradient flow path $\boldsymbol{w}_\alpha(t)$, $0 \leq t \leq T$. More specifically, by Theorem 2.2 we know that $\sup_{t \in [0,T]} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$, and so $D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})$ is locally Lipschitz, by assumption, on some closed Euclidean ball containing the gradient flow path $\boldsymbol{w}_\alpha(t)$, $0 \leq t \leq T$; this implies $D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})$ is Lipschitz on the Euclidean ball. As a consequence of this bound on the norm of the Taylor remainder,

$$\begin{aligned}
\sup_{\boldsymbol{x} \in \mathcal{X}} \|\alpha f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x})\|_2 &\leq \sup_{\boldsymbol{x} \in \mathcal{X}} \frac{\alpha}{2} \mathrm{Lip}(\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})) \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2 \\
&\leq \frac{\alpha}{2} \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2 \sup_{\boldsymbol{x} \in \mathcal{X}} \mathrm{Lip}(\boldsymbol{w} \mapsto D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x})) \\
&\leq \frac{\alpha}{2} M_2 \|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2.
\end{aligned}$$

Lastly, appealing to our assumption that the bounds we derived in Theorem 2.2 indeed hold, then we have $\|\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0\|_2^2 \leq C_2^2 \log(\alpha)^2/\alpha^2$. Consequently, we attain

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\alpha f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x})\|_2 \leq \frac{M_2 C_2^2 \log(\alpha)^2}{2\alpha}.$$

Notice that for this first bound, we only used information about how far the gradient flow path $(\boldsymbol{w}_\alpha(t))_{t \geq 0}$ is from its initialization $\boldsymbol{w}_\alpha(0) = \boldsymbol{w}_0$ at time $T > 0$ and not how far the two gradient flow paths $(\boldsymbol{w}_\alpha(t))_{t \geq 0}$ and $(\bar{\boldsymbol{w}}_\alpha(t))_{t \geq 0}$ are from one another at time T. The second term we bound will capture the distance between these two gradient flow paths.

Specifically, we wish to derive a bound on $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha \bar{f}(\boldsymbol{w}_\alpha(T), x) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2$. By the definition of the linearized model $\bar{f}(\boldsymbol{w}, \boldsymbol{x})$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2 \\
=&\sup_{\boldsymbol{x}\in\mathcal{X}} \alpha\|(f(\boldsymbol{w}_0, \boldsymbol{x}) + D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \boldsymbol{w}_0)) - (f(\boldsymbol{w}_0, \boldsymbol{x}) + D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\bar{\boldsymbol{w}}_\alpha(T) - \boldsymbol{w}_0))\|_2 \\
=&\sup_{\boldsymbol{x}\in\mathcal{X}} \alpha\|D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \bar{\boldsymbol{w}}_\alpha(T))\|_2.
\end{aligned}$$

And so by the properties of the matrix norm $\|\cdot\|_{k,p}$, we have

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha \bar{f}(\boldsymbol{w}_\alpha(T), x) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2 =& \sup_{\boldsymbol{x}\in\mathcal{X}} \alpha\|D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})(\boldsymbol{w}_\alpha(T) - \bar{\boldsymbol{w}}_\alpha(T))\|_2 \\
\leq& \alpha\|\boldsymbol{w}_\alpha(T) - \bar{\boldsymbol{w}}_\alpha(T)\|_2 \sup_{\boldsymbol{x}\in\mathcal{X}} \|D_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x})\| \\
\leq& \alpha M_1 \|\boldsymbol{w}_\alpha(T) - \bar{\boldsymbol{w}}_\alpha(T)\|_2.
\end{aligned}$$

Lastly, by our bound from Theorem 2.2 on the distance between the gradient flow paths of $F_\alpha(\boldsymbol{w})$, $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$, and $\bar{F}_\alpha(\boldsymbol{w})$, $(\bar{\boldsymbol{w}}_\alpha(t))_{t\geq 0}$, we deduce

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha \bar{f}(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2 \leq \frac{M_1 C_1 \log(\alpha)}{\alpha}.$$

Altogether, we have proven

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2 \leq \frac{M_1 C_1 \log(\alpha)}{\alpha} + \frac{M_2 C_2^2 \log(\alpha)^2}{2\alpha},$$

which implies that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\alpha f(\boldsymbol{w}_\alpha(T), \boldsymbol{x}) - \alpha \bar{f}(\bar{\boldsymbol{w}}_\alpha(T), \boldsymbol{x})\|_2 \longrightarrow 0 \quad \text{as} \quad \alpha \longrightarrow \infty.$$

□

And so we have shown that for a certain subset $\mathcal{X}$ of the input space $\mathbb{R}^d$ on which the derivative of f at $\boldsymbol{w}_0$ has bounded matrix norm and $\text{Lip}(\boldsymbol{w} \to D_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}))$ is bounded, $\alpha f(\boldsymbol{w}, \boldsymbol{x})$ indeed generalizes like the linearized model $\alpha \bar{f}(\boldsymbol{w}, \boldsymbol{x})$ in limit $\alpha \to \infty$.

3.2 Extending to Uniform-time Bounds

In Section 3.1 we delineated the conditions under which lazy training occurs and provided mathematical characterizations for lazy training that build upon our intuitive understanding from Sections 1 and 2. Still, from the practictioner's perspective, the results we have presented are of limited utility. Specifically, each of the bounds in Theorem 2.2 is dependent on a finite time horizon $T > 0$, meaning that the theoretical convergence we proved may be difficult to observe in practice for large time $t > 0$. As we pointed out, this is problematic because in order to approximate the limit of the gradient flow paths $F_\alpha(\boldsymbol{w})$ and $\bar{F}_\alpha(\boldsymbol{w})$, we must observe $\boldsymbol{w}_\alpha(t)$ and $\bar{\boldsymbol{w}}_\alpha(t)$ for large $t \geq 0$.

To partially remedy this drawback of Theorem 2.2, Chizat and colleagues impose additional assumptions on the model h and loss R in order to achieve uniform convergence in time $t \geq 0$. These assumptions are summarized by Theorem 2.4:

Theorem 2.4. *Consider the M-smooth and m-strongly convex loss R with minimizer $y^\star$ and condition number $\kappa := M/m$. Assume that σ_{min}, the smallest singular value of $Dh(\boldsymbol{w}_0)^T$, is positive and that the initialization satisfies $\|h(\boldsymbol{w}_0)\|_{\mathcal{F}} \leq C_0 := \sigma_{min}^3/(32\kappa^{3/2}\|Dh(\boldsymbol{w}_0)\|Lip(Dh))$, where $Lip(Dh)$ is the Lipschitz constant of Dh. If $\alpha > \|y^\star\|_{\mathcal{F}}/C_0$, then for $t \geq 0$, it holds*

$$\|\alpha h(\boldsymbol{w}_\alpha(t)) - y^\star\|_{\mathcal{F}} \leq \sqrt{\kappa}\|\alpha h(\boldsymbol{w}_0) - y^\star\|_{\mathcal{F}} \exp(-m\sigma_{min}^2 t/4).$$

If moreover $h(\boldsymbol{w}_0) = 0$, it holds as $\alpha \to \infty$, $\sup_{t\geq 0}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$,

$$\sup_{t\geq 0}\|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\|_{\mathcal{F}} = \mathcal{O}(1/\alpha) \quad \textit{and} \quad \sup_{t\geq 0}\|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 = \mathcal{O}(\log\alpha/\alpha^2).$$

The first assumption the authors make is that the loss function R is "nice" to the extent that it is both M-smooth and m-strongly convex. Also, they require that $\boldsymbol{w} \mapsto Dh(\boldsymbol{w})$ is globally Lipschitz, whereas we only needed the operator to be locally Lipschitz for Theorem 2.2. The strongest assumption by far, though, is that $Dh(\boldsymbol{w}_0) : \mathbb{R}^p \to \mathcal{F}$ is surjective, which can only be the case if the Hilbert space $\mathcal{F}$ is finite-dimensional [2]. Professedly, these conditions are met only by very benign problems which we do not typically encounter in the contemporary field of deep learning.

By introducing these additional assumptions, we see that Theorem 2.4 indeed relaxes each of the bounds in Theorem 2.2 to be uniform in time $t \geq 0$. That is, Theorem 2.4 tells us that the convergence of the gradient flow path $\boldsymbol{w}_\alpha(t)$ to $\bar{\boldsymbol{w}}_\alpha(t)$ and the corresponding model $\alpha h(\boldsymbol{w}_\alpha(t))$ to $\alpha\bar{h}(\boldsymbol{w}_\alpha(t))$ is uniform in $t \geq 0$ as $\alpha \to \infty$. Just as salient, Theorem 2.4 also tells us that for sufficiently large $\alpha \geq 0$, the model $\alpha h(\boldsymbol{w})$ evaluated along its gradient flow path $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$ converges linearly to the global minimizer $y^\star \in \mathcal{F}$ of the loss R. Note that since R is m-strongly convex, then it must be the case that $y^\star$ is unique. Considering that the objective function $F_\alpha(\boldsymbol{w})$ is not necessarily convex, the fact that we achieve convergence of the gradient flow to the global minimum of the loss is a remarkable result.

Now that we have stated and provided motivation for Theorem 2.4 by Chizat and colleagues, we consider their proof.

Proof. To begin, we define the closed Euclidean ball with radius $r_0 = \sigma_{\min}/(2\mathrm{Lip}(Dh))$ centered at $\boldsymbol{w}_0$, $B_{r_0}(\boldsymbol{w}_0) = \{\boldsymbol{w} \in \mathbb{R}^p, \|\boldsymbol{w} - \boldsymbol{w}_0\|_2 \leq r_0\}$. By our assumption that the map $\boldsymbol{w} \mapsto Dh(\boldsymbol{w})$ is [globally] Lipschitz, we know that for each $f \in \mathcal{F}$,

$$\begin{aligned}
\mathrm{Lip}(Dh)^2\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2\|f\|_2^2 &\geq \|Dh(\boldsymbol{w}) - Dh(\boldsymbol{w}_0)\|^2\|f\|_{\mathcal{F}}^2 \\
&\geq \|(Dh(\boldsymbol{w}) - Dh(\boldsymbol{w}_0))^T f\|_2^2 \\
&= \|Dh(\boldsymbol{w})^T f - Dh(\boldsymbol{w}_0)^T f\|_2^2 \\
&\geq \big|\|Dh(\boldsymbol{w}_0)^T f\|_2^2 - \|Dh(\boldsymbol{w})^T f\|_2^2\big| \\
&\geq \|Dh(\boldsymbol{w}_0)^T f\|_2^2 - \|Dh(\boldsymbol{w})^T f\|_2^2, \quad \forall \boldsymbol{w} \in \mathbb{R}^p.
\end{aligned}$$

By subtracting $\|Dh(\boldsymbol{w}_0)^T f\|_2^2$ over to the left-hand side of the inequality, we have the bound

$$\|Dh(\boldsymbol{w})^T f\|_2^2 \geq \|Dh(\boldsymbol{w}_0)^T f\|_2^2 - \mathrm{Lip}(Dh)^2\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2\|f\|_2^2, \quad \forall \boldsymbol{w} \in \mathbb{R}^p.$$

More specifically, for $\boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0)$ we know that $\|\boldsymbol{w} - \boldsymbol{w}_0\|_2^2 \leq r_0^2 = \sigma_{\min}^2/(4\mathrm{Lip}(Dh)^2)$, and so

$$\|Dh(\boldsymbol{w})^T f\|_2^2 \geq \|Dh(\boldsymbol{w}_0)^T f\|_2^2 - (1/4)\sigma_{\min}^2\|f\|_{\mathcal{F}}^2, \quad \forall \boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0).$$

And by our assumption that the smallest singular value of $Dh(\boldsymbol{w}_0)^T$ is $\sigma_{\min}^2 > 0$, then we have a lower bound on the right-hand side of the inequality:

$$\|Dh(\boldsymbol{w})^T f\|_2^2 \geq \|Dh(\boldsymbol{w}_0)^T f\|_2^2 - (1/4)\sigma_{\min}^2\|f\|_{\mathcal{F}}^2 \geq \sigma_{\min}^2\|f\|_{\mathcal{F}}^2 - (1/4)\sigma_{\min}^2\|f\|_{\mathcal{F}}^2, \quad \forall \boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0).$$

And so we have proven that $\forall \boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0)$, it holds that

$$\begin{aligned} & f\Sigma(\boldsymbol{w})f = \|Dh(\boldsymbol{w})^T f\|_2^2 \geq (3/4)\sigma_{\min}^2 \|f\|_{\mathcal{F}}^2, \quad \forall f \in \mathcal{F} \\ \implies & \Sigma(\boldsymbol{w}) \succeq (3/4)\sigma_{\min}^2 \mathrm{Id}. \end{aligned}$$

That is, we have shown that the smallest eigenvalue of the neural tangent kernel $\Sigma(\boldsymbol{w})$ on the set $B_{r_0}(\boldsymbol{w}_0)$ is strictly positive. Of course, the natural question to ask is how this result relates to the statements we wish to prove. This is clarified by the following lemma proven by Chizat and colleagues:

Lemma B.1 (Strongly-convex gradient flow in a time-dependent metric)**.** *Let $F : \mathcal{F} \to \mathbb{R}$ be a m-strongly convex function with M-Lipschitz continuous gradient and with global minimizer $y^\star$. Let $\Sigma(t) : \mathcal{F} \to \mathcal{F}$ be a time-dependent, continuous, self-adjoint linear operator with eigenvalues lower bounded by $\lambda > 0$ for $0 \leq t \leq T$. Then the solutions on $[0, T]$ to the differential equation*

$$y'(t) = -\Sigma(t)\nabla F(y(t))$$

satisfy, for $0 \leq t \leq T$,

$$\|y(t) - y^\star\|_{\mathcal{F}} \leq (M/m)^{1/2} \|y(0) - y^\star\|_{\mathcal{F}} \exp(-m\lambda t).$$

Lemma B.1 tells us that since the scaled model evaluated along its gradient flow path, $y(t) = \alpha h(\boldsymbol{w}_\alpha(t))$, satisfies the differential equation

$$\frac{d}{dt} y(t) = -\Sigma(\boldsymbol{w}_\alpha(t))\nabla R(y(t)),$$

where the loss R is M-smooth and m-strongly convex as well as $\Sigma(\boldsymbol{w}) \succeq (3/4)\sigma_{\min}^2 \mathrm{Id}$, $\forall \boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0)$, then we have that $y(t)$ converges linearly to the global minimizer of R, $y^\star$, $\forall t \in [0, T]$, where $T = \inf\{t \geq 0 \mid \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 > r_0\}$. More intuitively, our previous result that $\Sigma(\boldsymbol{w}) \succeq (3/4)\sigma_{\min}^2 \mathrm{Id}$, $\forall \boldsymbol{w} \in B_{r_0}(\boldsymbol{w}_0)$ along with Lemma B.1 tell us that $y(t)$ converges linearly to $y^\star$ as long as the gradient flow path $\boldsymbol{w}_\alpha(t)$ remains in the ball $B_{r_0}(\boldsymbol{w}_0)$.

And so in order to prove the global convergence result we desire, we must find sufficient conditions such that $T = +\infty$, meaning that the gradient flow path never leaves the ball $B_{r_0}(\boldsymbol{w}_0)$. In order to do this, we first bound the norm of $\boldsymbol{w}'(t)$ on $t \in [0, T]$ as follows:

$$\begin{aligned} \|\boldsymbol{w}'(t)\|_2 = & \frac{1}{\alpha} \|Dh(\boldsymbol{w}_\alpha(t))\| \|\nabla R(y(t))\|_{\mathcal{F}} && \text{definition of } \boldsymbol{w}'(t) \text{ from Section 2} \\ \leq & \frac{M}{\alpha} \|Dh(\boldsymbol{w}_\alpha(t))\| \|y(t) - y^\star\|_{\mathcal{F}} && R \text{ is } M\text{-smooth} \\ \leq & \frac{2M}{\alpha} \|Dh(\boldsymbol{w}_0)\| \|y(t) - y^\star\|_{\mathcal{F}}. \end{aligned}$$

Therefore, we have that for all $t \in [0, T]$,

$$\begin{aligned} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = & \left\| \int_0^t \boldsymbol{w}'_\alpha(s)\ ds \right\|_2 \leq \int_0^t \|\boldsymbol{w}'_\alpha(s)\|_2\ ds && \text{Fundamental Theorem of Calculus} \\ \leq & \frac{2M}{\alpha} \|Dh(\boldsymbol{w}_0)\| \int_0^t \|y(t) - y^\star\|_{\mathcal{F}}\ ds \\ \leq & \frac{2M^{3/2}}{m^{1/2}\alpha} \|Dh(\boldsymbol{w}_0)\| \|y(0) - y^\star\|_{\mathcal{F}} \\ & \cdot \int_0^t \exp(-(3m\sigma_{\min}^2/4) \cdot s)\ ds && \text{Lemma B.1} \\ \leq & \frac{8\kappa^{3/2}}{\alpha\sigma_{\min}^2} \|Dh(\boldsymbol{w}_0)\| \|y(0) - y^\star\|_{\mathcal{F}}. \end{aligned}$$

Let us consider $\|y(0) - y^\star\|_{\mathcal{F}} \leq 2\alpha C_0$. If this is the case, then the previous inequality implies

$$\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \leq \frac{8\kappa^{3/2}}{\alpha\sigma_{\min}^2}\|Dh(\boldsymbol{w}_0)\|\|y(0) - y^\star\|_{\mathcal{F}} \leq \frac{16C_0\kappa^{3/2}}{\sigma_{\min}^2}\|Dh(\boldsymbol{w}_0)\| \leq \frac{\sigma_{\min}}{2\mathrm{Lip}(Dh)} = r_0,$$

meaning that $T = \inf\{t \geq 0 \mid \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 > r_0\} = +\infty$, as we wished. That is, we have shown that for $y(0)$ satisfying $\|y(0) - y^\star\|_{\mathcal{F}} \leq 2\alpha C_0$, the gradient flow path $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$ remains in $B_{r_0}(\boldsymbol{w}_0)$ for all times $t \geq 0$, and we attain linear convergence of $\alpha h(\boldsymbol{w}_\alpha(t))$ to the global minimum $y^\star$ for all times $t \geq 0$.

Recall from the statement of Theorem 2.4 that we assume $\|h(\boldsymbol{w}_0)\|_{\mathcal{F}} \leq C_0$ as well as $\alpha > \|y^\star\|_{\mathcal{F}}/C_0$. Therefore, we indeed have $\|y(0) - y^\star\|_{\mathcal{F}} = \|\alpha h(\boldsymbol{w}_0) - y^\star\|_{\mathcal{F}} \leq \alpha\|h(\boldsymbol{w}_0)\|_{\mathcal{F}} + \|y^\star\|_{\mathcal{F}} < 2\alpha C_0$. And so by our assumptions in Theorem B.1, we have shown that we are guaranteed linear convergence to the global minimum for all times $t \geq 0$:

$$\|y(t) - y^\star\|_{\mathcal{F}} \leq \sqrt{\kappa}\|\alpha h(\boldsymbol{w}_0) - y^\star\|_{\mathcal{F}}\exp(-3m\sigma_{\min}^2 t/4) \leq \sqrt{\kappa}\|\alpha h(\boldsymbol{w}_0) - y^\star\|_{\mathcal{F}}\exp(-m\sigma_{\min}^2 t/4), \quad t \geq 0.$$

Now to consider the uniform time bounds, let us suppose that the model h is unbiased at its initialization, $h(\boldsymbol{w}_0) = 0$. We claim that we have already proven the first bound $\sup_{t\geq 0}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$. To see that this is the case, recall that we showed for all $t \in [0, T] = [0, +\infty)$,

$$\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \leq \frac{1}{\alpha}\left(\frac{8\kappa^{3/2}}{\sigma_{\min}^2}\|Dh(\boldsymbol{w}_0)\|\|y(0) - y^\star\|_{\mathcal{F}}\right) = \frac{1}{\alpha}\left(\frac{8\kappa^{3/2}}{\sigma_{\min}^2}\|Dh(\boldsymbol{w}_0)\|\|y^\star\|_{\mathcal{F}}\right).$$

However, one will observe that the right-hand side of the inequality is independent of time $t \geq 0$, and so we obtain the desired result

$$\sup_{t\geq 0}\|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \leq \frac{1}{\alpha}\left(\frac{8\kappa^{3/2}}{\sigma_{\min}^2}\|Dh(\boldsymbol{w}_0)\|\|y^\star\|_{\mathcal{F}}\right) = \mathcal{O}(1/\alpha).$$

That is, the gradient flow path of $F_\alpha(\boldsymbol{w})$, $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$, remains asymptotically fixed at its initialization $\boldsymbol{w}_\alpha(0) = \alpha\boldsymbol{w}_0$, and this convergence is uniform in time $t \geq 0$.

For the next result $\sup_{t\geq 0}\|\alpha h(\boldsymbol{w}_\alpha(t)) - \alpha\bar{h}(\bar{\boldsymbol{w}}_\alpha(t))\|_{\mathcal{F}} = \|y(t) - \bar{y}(t)\|_{\mathcal{F}} = \mathcal{O}(1/\alpha)$, we must appeal to a second lemma formulated and proven by Chizat and colleagues:

Lemma B.2 (Stability Lemma)**.** *Let $R : \mathcal{F} \to \mathbb{R}_+$ be a m-strongly convex function and let $\Sigma(t)$ be a time-dependent positive definite operator on $\mathcal{F}$ such that $\Sigma(t) \succeq \lambda Id$ for $t \geq 0$. Consider the paths $y(t)$ and $\bar{y}(t)$ on $\mathcal{F}$ that solve for $t \geq 0$,*

$$y'(t) = -\Sigma(t)\nabla R(y(t)) \qquad \text{and} \qquad \bar{y}'(t) = -\Sigma(0)\nabla R(\bar{y}(t)).$$

Defining $K := \sup_{t\geq 0}\|(\Sigma(t) - \Sigma(0))\nabla R(y(t))\|_{\mathcal{F}}$, it holds for $t \geq 0$,

$$\|y(t) - \bar{y}(t)\|_{\mathcal{F}} \leq \frac{K\|\Sigma(0)\|^{1/2}}{\lambda^{3/2}m}.$$

Once again, we have assumed that the loss function R is m-strongly convex and we know that $\Sigma(t) = Dh(\boldsymbol{w}_\alpha(t))Dh(\boldsymbol{w}_\alpha(t))^T \succeq (3/4)\sigma_{\min}^2\mathrm{Id}$, $\forall t \in [0, T] = [0, +\infty)$ from the previous portion of our proof. Therefore, we can invoke the stability lemma to bound $\sup_{t\geq 0}\|y(t) - \bar{y}(t)\|_{\mathcal{F}}$. In our case, we have that the constant K is upper bounded by

$$\begin{aligned} K =& \sup_{t\geq 0}\|(\Sigma(t) - \Sigma(0))\nabla R(y(t))\|_{\mathcal{F}} \\ \leq& \sup_{t\geq 0}\|\Sigma(t) - \Sigma(0)\|\|\nabla R(y(t))\|_{\mathcal{F}} \\ \leq& M \cdot \sup_{t\geq 0}\|\Sigma(t) - \Sigma(0)\|\|y(t) - y^\star\|_{\mathcal{F}} && R \text{ is } M\text{-smooth} \end{aligned}$$

$$
\begin{aligned}
&\leq (2M\mathrm{Lip}(h)\mathrm{Lip}(Dh)) \cdot \sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \|y(t) - y^\star\|_{\mathcal{F}} && \mathrm{Lip}(\Sigma) \leq 2 \cdot \mathrm{Lip}(h)\mathrm{Lip}(Dh) \\
&\leq \left(2\frac{M^{3/2}}{m^{1/2}} \cdot \|y(0) - y^\star\|_{\mathcal{F}} \cdot \mathrm{Lip}(h)\mathrm{Lip}(Dh)\right) \cdot \sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 && \text{Lemma B.1} \\
&= \left(2\frac{M^{3/2}}{m^{1/2}} \cdot \|y^\star\|_{\mathcal{F}} \cdot \mathrm{Lip}(h)\mathrm{Lip}(Dh)\right) \cdot \sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2.
\end{aligned}
$$

Here, $\mathrm{Lip}(h)$ denotes the Lipschitz constant of h on the closed Euclidean ball $B_{r_0}(\boldsymbol{w}_0)$. More precisely, because Dh is continuous on the compact set $B_{r_0}(\boldsymbol{w}_0)$, then h is Lipschitz on $B_{r_0}(\boldsymbol{w}_0)$. Altogether, we can bound $\sup_{t\geq 0} \|y(t) - \bar{y}(t)\|_{\mathcal{F}}$ by

$$
\sup_{t\geq 0} \|y(t) - \bar{y}(t)\|_{\mathcal{F}} \leq \frac{K\|\Sigma(0)\|^{1/2}}{(3\sigma_{\min}^2/4)^{3/2} m} \leq \left(\frac{16\kappa^{3/2}\|\Sigma(0)\|^{1/2} \cdot \mathrm{Lip}(h)\mathrm{Lip}(Dh) \cdot \|y^\star\|_{\mathcal{F}}}{\sigma_{\min}^3}\right) \cdot \sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2.
$$

But we already know that $\sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 = \mathcal{O}(1/\alpha)$, and so the previous bound implies

$$
\sup_{t\geq 0} \|y(t) - \bar{y}(t)\|_{\mathcal{F}} = \mathcal{O}(1/\alpha).
$$

Finally, it remains to prove the bound on the distance between the gradient flow path of $F_\alpha(\boldsymbol{w})$ and that of $\bar{F}_\alpha(\boldsymbol{w})$ at any time $t \geq 0$, $\sup_{t\geq 0} \|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 = \mathcal{O}(\log(\alpha)/\alpha^2)$. In order to do so, we employ a strategy that we have used many times up to this point: bounding $\|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2$ by $\|\boldsymbol{w}'_\alpha(t) - \bar{\boldsymbol{w}}'_\alpha(t)\|_2$. In particular, we have that $\forall t \geq 0$,

$$
\begin{aligned}
&\|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 \\
\leq &\int_0^t \|\boldsymbol{w}'_\alpha(s) - \bar{\boldsymbol{w}}'_\alpha(s)\|_2 \, ds \\
\leq &\int_0^\infty \|\boldsymbol{w}'_\alpha(t) - \bar{\boldsymbol{w}}'_\alpha(t)\|_2 \, dt \\
= &(1/\alpha) \int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t))^T \nabla R(y(t)) - Dh(\boldsymbol{w}_0)^T \nabla R(\bar{y}(t))\|_2 \, dt \\
\leq &(1/\alpha) \int_0^\infty \|(Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0))^T \nabla R(y(t))\|_2 \, dt + (1/\alpha) \int_0^\infty \|Dh(\boldsymbol{w}_0)(\nabla R(y(t)) - \nabla R(\bar{y}(t)))\|_2 \, dt \\
\leq &(1/\alpha) \int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0)\| \|\nabla R(y(t))\|_{\mathcal{F}} \, dt + (1/\alpha) \int_0^\infty \|Dh(\boldsymbol{w}_0)\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \, dt.
\end{aligned}
$$

And so it suffices to show that each of

$$
\int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0)\| \|\nabla R(y(t))\|_{\mathcal{F}} \, dt, \quad \int_0^\infty \|Dh(\boldsymbol{w}_0)\| \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}} \, dt
$$

is on the order of $\log(\alpha)/\alpha$. Starting with the first integral, we have

$$
\begin{aligned}
&\int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0)\| \|\nabla R(y(t))\|_{\mathcal{F}} \, dt \\
\leq &\mathrm{Lip}(Dh) \cdot \int_0^\infty \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \|\nabla R(y(t))\|_{\mathcal{F}} \, dt && Dh \text{ is globally Lipschitz} \\
\leq &(M \cdot \mathrm{Lip}(Dh)) \cdot \int_0^\infty \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \|y(t) - y^\star\|_{\mathcal{F}} \, dt && R \text{ is } M\text{-smooth} \\
\leq &\left(M\sqrt{\kappa} \cdot \mathrm{Lip}(Dh) \cdot \|y(0) - y^\star\|_{\mathcal{F}}\right) \cdot \int_0^\infty \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2 \exp(-m\sigma_{\min}^2 t/4) \, dt && \text{linear convergence} \\
\leq &\frac{1}{\alpha}\left(\frac{8M\kappa^2}{\sigma_{\min}^2} \cdot \|Dh(\boldsymbol{w}_0)\| \cdot \mathrm{Lip}(Dh) \cdot \|y^\star\|_{\mathcal{F}}^2\right) \cdot \int_0^\infty \exp(-m\sigma_{\min}^2 t/4) \, dt && \text{bound on } \|\boldsymbol{w}_\alpha(t) - \boldsymbol{w}_0\|_2
\end{aligned}
$$

$$\leq \frac{1}{\alpha}\left(\frac{32\kappa^3}{\sigma_{\min}^4}\cdot\|Dh(\boldsymbol{w}_0)\|\cdot\mathrm{Lip}(Dh)\cdot\|y^\star\|_{\mathcal{F}}^2\right).$$

Hence, we deduce

$$\int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0)\|\|\nabla R(y(t))\|_{\mathcal{F}}\;dt = \mathcal{O}(1/\alpha),$$

as we wanted to show.

For the sake of brevity, we do not fully work out the second integral, although we will explain how to bound it. In particular, the integral

$$\int_0^\infty \|Dh(\boldsymbol{w}_0)\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}\;dt = \|Dh(\boldsymbol{w}_0)\|\cdot\int_0^\infty \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}\;dt$$

can be split into an integral over $[0, t_0]$ and integral over $[t_0, +\infty)$, where $t_0 := 4\log(\alpha)/(m\sigma_{\min}^2)$. On the interval $[0, t_0]$, the authors use the fact that the loss function R is M-smooth so that

$$\|Dh(\boldsymbol{w}_0)\|\cdot\int_0^{t_0} \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}\;dt \leq (M\cdot\|Dh(\boldsymbol{w}_0)\|)\cdot\int_0^{t_0}\|y(t)-\bar{y}(t)\|_{\mathcal{F}}\;dt.$$

But from previously, we know that $\sup_{t\geq 0}\|y(t)-\bar{y}(t)\|_{\mathcal{F}} = \mathcal{O}(1/\alpha)$. And so, overall, the integral of $\|y(t) - \bar{y}(t)\|_{\mathcal{F}}$ over $[0, t_0]$ is $\mathcal{O}(\log(\alpha)/\alpha)$.

And for the second integral over $[t_0, +\infty)$, the authors use the "crude" bound (i.e. that which does not exploit the smoothness of ∇R):

$$\|Dh(\boldsymbol{w}_0)\|\cdot\int_{t_0}^\infty \|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}\;dt \leq \|Dh(\boldsymbol{w}_0)\|\cdot\int_{t_0}^\infty \left(\|\nabla R(y(t))\|_{\mathcal{F}} + \|\nabla R(\bar{y}(t))\|_{\mathcal{F}}\right)\;dt.$$

Now, since ∇R decreases exponentially along both $y(t)$ and $\bar{y}(t)$, and by out particular choice of t_0, we get that the integral of $\|\nabla R(y(t))\|_{\mathcal{F}} + \|\nabla R(\bar{y}(t))\|_{\mathcal{F}}$ over $[t_0, +\infty)$ is $\mathcal{O}(\log(\alpha)/\alpha)$. Admittedly, our previous statement is quite loaded, and there are a few details one must work out to verify that this is the case.

In summary, we have proven that $\forall t \geq 0$,

$$\begin{aligned}
&\|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2\\
\leq&(1/\alpha)\cdot\int_0^\infty \|Dh(\boldsymbol{w}_\alpha(t)) - Dh(\boldsymbol{w}_0)\|\|\nabla R(y(t))\|_{\mathcal{F}}\;dt + (1/\alpha)\cdot\int_0^\infty \|Dh(\boldsymbol{w}_0)\|\|\nabla R(y(t)) - \nabla R(\bar{y}(t))\|_{\mathcal{F}}\;dt\\
=&\mathcal{O}(\log(\alpha)/\alpha^2),
\end{aligned}$$

which implies

$$\sup_{t\geq 0}\|\boldsymbol{w}_\alpha(t) - \bar{\boldsymbol{w}}_\alpha(t)\|_2 = \mathcal{O}(\log(\alpha)/\alpha^2).$$

Therefore, we have shown the linear convergence of $y(t)$ to $y^\star$ as well as the uniform convergence in time $t \geq 0$ of $\boldsymbol{w}_\alpha(t)$ to $\boldsymbol{w}_0$, $y(t)$ to $\bar{y}(t)$, and $\boldsymbol{w}_\alpha(t)$ to $\bar{\boldsymbol{w}}_\alpha(t)$. □

4 Applications & Extensions of Lazy Training

Although our presentation of lazy training is foremost theory driven, we wish to provide some intuition as to why lazy training is important from a practical perspective. Additionally, we will point out the limitations in the results of Chizat, Oyallon and Bach and discuss future work to better understand the role of lazy training in contemporary deep learning.

As for the practical implications of lazy training, we will first discuss what it means to train in the limit $\alpha \to \infty$. Specifically, as we briefly mentioned in our proof of Theorem 2.2, it is not difficult to show that under certain conditions on the model h loss function R, the gradient flow $(\boldsymbol{w}_\alpha(t))_{t\geq 0}$ in the limit $\alpha \to \infty$ is equivalent to a kernel method with kernel $\Sigma(\boldsymbol{w}_0)$ the neural tangent kernel [1]. This result differs from that of Jacot and colleagues [3], who show that the gradient flow is equivalent to a kernel method with kernel $\Sigma(\boldsymbol{w}_0)$ in the limit as the width (i.e. number hidden units) in the neural network tends to ∞. What this result tells us is that for the problem $h : \boldsymbol{w} \mapsto f(\boldsymbol{w}, \cdot)$, $f(\boldsymbol{w}, \cdot) : \mathbb{R}^d \to \mathbb{R}$, the gradient flow solution $\boldsymbol{w}_\alpha^\star = \lim_{t\to\infty} \boldsymbol{w}_\alpha(t)$ in the $\alpha \to \infty$ limit satisfies $\alpha h(\boldsymbol{w}_\alpha^\star) = \alpha f(\boldsymbol{w}_\alpha^\star, \boldsymbol{x}) = \sum_{i=1}^N \beta_i K(\boldsymbol{x}_i, \boldsymbol{x})$, where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the neural tangent kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}), \nabla_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}')\rangle$ and $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is our training data [1]. As one may suspect, looking for a predictor in the reproducing kernel Hilbert space determined by K may result in a model $f(\boldsymbol{w}_\alpha^\star, \boldsymbol{x})$ which does not generalize well outside of the training set. This is because the kernel predictor considers only a fixed set of features $\{\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_0, \boldsymbol{x}_i)\}_{i=1}^N$ determined by our set of training data.

To examine a concrete application, Woodworth and colleagues consider the lazy training limit $\alpha \to \infty$ for an alternative parameterization of the linear regression problem [5]. In particular, they consider the case in which the linear system $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ is underdetermined, and so there are many solution vectors $\boldsymbol{\beta}$ which minimize the empirical risk R. For this problem, the neural tangent kernel K is proportional to the ℓ^2 kernel, and so the solution reached by gradient flow in the lazy training limit $\alpha \to \infty$ is the minimum ℓ^2 solution, $\boldsymbol{\beta}^{\ell^2}$, of the system $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$. Conversely, in the limit $\alpha \to 0$ the gradient flow solution is the minimum ℓ^1 solution, $\boldsymbol{\beta}^{\ell^1}$, to the system $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ [5].

From these results, the deficiencies of lazy training are apparent. If we suspect that the data $(\boldsymbol{x}, y) \sim \rho$ is drawn from an underlying distribution ρ with implicit sparsity, then, in general, lazy training will generalize poorly for the linear regression model. An example of such a distribution ρ is $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbb{I}_{d\times d})$, $y = \langle \boldsymbol{x}, \boldsymbol{\beta}^\star\rangle$, where $(\boldsymbol{\beta}^\star)_i = 1/\sqrt{d^\star}$ for $1 \leq i \leq d^\star$, $(\boldsymbol{\beta}^\star)_i = 0$ otherwise [5]. Here, only the first $d^\star \ll d$ coordinates of $\boldsymbol{x}$ determine y, whereas the remaining $d - d^\star$ coordinates provide no information about y. For this example, lazy training $\alpha \to \infty$ would not result in a sparse solution vector $\boldsymbol{\beta}$, whereas training far away from the lazy training limit $\alpha \to 0$ would attain the sparse solution $\boldsymbol{\beta}^\star$. Although this is only one problem in which lazy training performs poorly, there is empirical evidence to suggest that it is more generally true. Expressly, there is experimental data suggesting that gradient flow far away from the lazy training limit, $\alpha \to 0$, corresponds to some form of implicit ℓ^1 regularization, which is not the case for lazy training $\alpha \to \infty$.

Our discussion of the implicit biases present in the gradient flow solution $\boldsymbol{w}_\alpha^\star$ resulting from lazy training motivate further study of this subject area. Although the work of Chizat and colleagues is certainly groundbreaking in its characterization of lazy training in the limit $\alpha \to \infty$, it is also very narrow in its applications. More explicitly, as we stated in Section 2, Theorem 2.2 assumes that both the model h and loss function R are everywhere differentiable, which is most definitely not the case for contemporary deep learning models. For instance, by applying the ReLU activation $\max\{0, x\}$ component-wise to the output of each hidden layer, we have a network function that is not differentiable in its weights. Also, as we pointed out in Section 3.2, the uniform time bounds we derived required exceedingly strong assumptions on both the model h and loss R. It is of interest whether or not we can relax any of these conditions and still attain convergence that is uniform in time $t \geq 0$.

5 Conclusion

Our report has taken a theoretical dive into the "lazy training" phenomenon framed by Chizat and colleagues in their paper "On Lazy Training in Differentiable Programming". We began by presenting an overview of lazy training. In particular, we gave a vague definition of lazy training as the phenomenon in which the gradient flow of some model h with loss R approaches the gradient flow for the linearization of h around initialization $\boldsymbol{w}_0$. We then proceeded to formalize this definition of lazy training in Theorem 2.2 and proved that it occurs when the scale of the model output α grows arbitrarily large. Thereafter, we justified that this result, while theoretically cogent, is limited in its applications due to the dependence on the time horizon of the gradient flow dynamics. To partially reconcile this difficulty, we then presented and proved Theorem 2.4, which gave us convergence that is uniform in time t but requires stronger assumptions on h and R. This theorem is also powerful to the extent that it proves the convergence of the model evaluated along the gradient flow path to the global minimum of the loss R. To complete our analysis of lazy training, we detailed some of its limitations, specifically for problems in which the true data distribution ρ is sparse.

By completing this project, I have gained a deeper understanding of the theory dictating lazy training. While my knowledge of lazy training was previously limited to the statements of the main theorems in [2], I now know how these results are proven and why they work. And beyond the subject matter itself, I have gained exposure to some functional analysis concepts, which will undoubtedly be useful in the future.

References

[1] L. Chizat and F. Bach, *A note on lazy training in supervised differentiable programming*, arXiv preprint arXiv:1812.07956, 8 (2018).

[2] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, arXiv preprint arXiv:1812.07956, (2018).

[3] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, arXiv preprint arXiv:1806.07572, (2018).

[4] A. Wibisono, *Gradient flow and gradient descent.* `http://awibisono.github.io/2016/06/13/gradient-flow-gradient-descent.html`, July 2016.

[5] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, *Kernel and rich regimes in overparametrized models*, in Conference on Learning Theory, PMLR, 2020, pp. 3635–3673.