Over-parameterised Shallow Neural Networks with Asymmetrical Node Scaling: Global Convergence Guarantees and Feature Learning

François Caron
Dept of Statistics, University of Oxford
Oxford, United Kingdom

Fadhel Ayed
Huawei Technologies
Paris, France

Paul Jung
Dept of Mathematics, Fordham University
Bronx, New York, USA

Hoil Lee
Dept of Mathematical Sciences, KAIST
Daejeon, South Korea

Juho Lee
Graduate School of AI, KAIST
Daejeon, South Korea

Hongseok Yang
School of Computing, KAIST
Daejeon, South Korea

Abstract

We consider gradient-based optimisation of wide, shallow neural networks with hidden-node ouputs scaled by positive scale parameters. The scale parameters are non-identical, differing from classical Neural Tangent Kernel (NTK) parameterisation. We prove that, for large networks, with high probability, gradient flow converges to a global minimum AND can learn features, unlike in the NTK regime.

1 Introduction

Training neural networks (NNs) involves minimising a non-convex objective function where optimisation methods, such as gradient descent (GD), often find solutions with low training error. To better understand this phenomenon, one line of research has analysed GD training of over-parameterised NNs with a large number m of hidden nodes. Under a "$\sqrt{1/m}$" scaling of hidden nodes, Jacot et al. (2018) showed that, as $m \to \infty$, the GD solution coincides with that of kernel regression under a limiting *Neural Tangent Kernel* (NTK). Under this so-called *NTK scaling*, theoretical guarantees for global convergence and generalisation properties have been shown for large-width NNs (Du et al., 2019b,a; Oymak and Soltanolkotabi, 2020; Arora et al., 2019; Bartlett et al., 2021). However, a number of articles (Chizat et al., 2019; Yang, 2019; Arora et al., 2019) noted that in this large-width regime, there is no feature learning; thus, for large-width NNs under NTK scaling, GD training is performed in a *lazy training* regime, in contrast to the typical feature-learning regime of deep NNs.

We investigate global convergence properties and feature learning in gradient-type training of feedforward neural networks (FFNNs) under a more general asymmetrical node scaling. In particular, each node $j = 1, \dots, m$ has a fixed node-specific scaling $\sqrt{\lambda_{m,j}}$ with

$$\lambda_{m,j} = \gamma \cdot \frac{1}{m} + (1-\gamma) \cdot \frac{\widetilde{\lambda}_j}{\sum_{k=1}^m \widetilde{\lambda}_k} \tag{1}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

where $\gamma \in [0,1]$ and $1 \geq \widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \ldots \geq 0$ are fixed scalars with $\sum_{j=1}^{\infty} \widetilde{\lambda}_j = 1$. Note that $\gamma = 1$ corresponds to the NTK scaling $\sqrt{1/m}$. If $\gamma < 1$, the node scaling is necessarily asymmetrical for large-width networks. A typical example might be to take, for instance, $\widetilde{\lambda}_j = 6\pi^{-2}j^{-2}$ for all $j \geq 1$.

We consider a shallow FFNN with smooth activation function and without bias, where the first layer weights are trained via gradient flow (GF) using empirical risk minimisation under the ℓ_2 loss. We show that, under similar assumptions as (Du et al., 2019b,a) on the data, activation function, and initialisation, when the number of nodes m is sufficiently large: (i) if $\gamma > 0$, the training error goes to 0 at a linear rate with high probability; and (ii) feature learning arises if and only if $\gamma < 1$. In the Supplementary Material, we also provide numerical experiments which illustrate the theoretical results, and which demonstrate empirically that such node-scaling is also useful for transfer learning.

2 Problem setup

Model. Consider a shallow FFNN with m hidden nodes and scalar output. For simplicity, assume there is no bias term. Let $\mathbf{x} \in \mathbb{R}^d$ be an input vector, where d is the input dimension. The model is

$$f_m(\mathbf{x}; \mathbf{W}) = \sum_{j=1}^{m} \sqrt{\lambda_{m,j}} a_j \sigma(Z_j(\mathbf{x}; \mathbf{W})) \quad \text{with} \quad Z_j(\mathbf{x}; \mathbf{W}) = \frac{1}{\sqrt{d}} \mathbf{w}_j^\top \mathbf{x}, \ \text{ for } j \in [m] \tag{2}$$

where $f_m(\mathbf{x}; \mathbf{W})$ is the scalar output of the FFNN; $Z_j(\mathbf{x}; \mathbf{W})$ is the pre-activation of the j-th hidden node; $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function; $\mathbf{w}_j \in \mathbb{R}^d$ is the column vector of weights between node j of the hidden layer and the input nodes; $a_j \in \mathbb{R}$ is the weight between the hidden node j and the output node; $\lambda_{m,j} \geq 0$ is a scaling parameter for hidden node j; $\mathbf{W} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_m^\top)^\top$ is a column vector of dimension md corresponding to the parameters to be optimised.

Assume σ admits a derivative σ'. For $n \geq 1$, let $\boldsymbol{\sigma} : \mathbb{R}^n \to \mathbb{R}^n$ (resp. $\boldsymbol{\sigma}' : \mathbb{R}^n \to \mathbb{R}^n$) be the vector-valued multivariate function that applies σ (resp. σ') element-wise to each of the n input variables. For simplicity, we henceforth assume that the output weights a_j are randomly initialised and fixed afterwards: $a_j \overset{\text{iid}}{\sim} \text{Uniform}(\{-1, 1\})$, $j \geq 1$. This assumption is common for large shallow networks (see e.g. (Du et al., 2019b; Bartlett et al., 2021)), and typically the analysis extends to models which train both layers. The scaling parameter $\lambda_{m,j}$ is fixed and satisfies Equation (1). By construction, $\lambda_{m,1} > 0$ and $\sum_{j=1}^{m} \lambda_{m,j} = 1$ for all $m \geq 1$. The case $\gamma = 1$ corresponds to NTK scaling. The case $\gamma = 0$ and $\widetilde{\lambda}_j = \frac{1}{K}$ for $j \in [K]$ for some $K \leq m$ and 0 otherwise corresponds to a finite FFNN of width K.

Training. Let $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$ be the training dataset of $n \geq 1$ observations. Let $\mathbf{X}$ be the n-by-d matrix whose ith row is $\mathbf{x}_i^\top$. We aim to minimise the empirical risk under ℓ_2 loss. Let

$$L_m(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f_m(\mathbf{x}_i; \mathbf{W}))^2, \tag{3}$$

be the objective function which is, in general, non-convex. For dataset $\mathcal{D}_n$, width $m \geq 1$, output weights a_j, and scaling parameters $(\lambda_{m,j})_{j \in [m]}$, we aim to estimate the trainable parameters $\mathbf{W}$ by minimising $L_m(\mathbf{W})$ using GF (in the Supplementary Material we discuss an extension to GD). Let $\mathbf{W}_0$ be some initialisation. Under GF, $(\mathbf{W}_t)_{t>0}$ is the solution to the following ordinary differential equation (ODE): $\frac{d\mathbf{W}_t}{dt} = -\nabla_{\mathbf{W}} L_m(\mathbf{W}_t)$ with $\lim_{t \to 0} \mathbf{W}_t = \mathbf{W}_0$. Let $\mathbf{w}_{tj}$ be the value of the parameter $\mathbf{w}_j$ at time t, and define $Z_{tj}(\mathbf{x}) = Z_j(\mathbf{x}; \mathbf{W}_t)$. Note that $\nabla_{\mathbf{w}_j} f_m(\mathbf{x}; \mathbf{W}) = \sqrt{\lambda_{m,j}} a_j \sigma'(Z_j(\mathbf{x}; \mathbf{W})) \cdot \frac{1}{\sqrt{d}} \mathbf{x}$. Thus, under gradient flow, for $j \in [m]$ and $\mathbf{x} \in \mathbb{R}^d$,

$$\frac{d\mathbf{w}_{tj}}{dt} = \sum_{i=1}^{n} (y_i - f_m(\mathbf{x}_i; \mathbf{W}_t)) \nabla_{\mathbf{w}_j} f_m(\mathbf{x}; \mathbf{W}_t) = \frac{\sqrt{\lambda_{m,j}} a_j}{\sqrt{d}} \sum_{i=1}^{n} (y_i - f_m(\mathbf{x}_i; \mathbf{W}_t)) \sigma'(Z_{tj}(\mathbf{x}_i)) \mathbf{x}_i.$$

Note that the derivatives associated with each hidden node j are scaled by $\sqrt{\lambda_{m,j}}$. For an input $\mathbf{x} \in \mathbb{R}^d$, the output of the FFNN therefore satisfies the ODE

$$\frac{df_m(\mathbf{x}; \mathbf{W}_t)}{dt} = \nabla_{\mathbf{W}} f_m(\mathbf{x}; \mathbf{W}_t)^\top \frac{d\mathbf{W}_t}{dt} = \sum_{i=1}^{n} (y_i - f_m(\mathbf{x}_i; \mathbf{W}_t)) \Theta_m(\mathbf{x}, \mathbf{x}_i; \mathbf{W}_t),$$

where $\Theta_m : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the neural tangent kernel, defined by

$$\Theta_m(\mathbf{x}, \mathbf{x}'; \mathbf{W}) = \frac{\mathbf{x}^\top \mathbf{x}'}{d} \sum_{j=1}^{m} \lambda_{m,j} \sigma'(Z_j(\mathbf{x}; \mathbf{W}))\sigma'(Z_j(\mathbf{x}'; \mathbf{W})). \tag{4}$$

The associated neural tangent Gram (NTG) matrix $\widehat{\Theta}_m(\mathbf{X}; \mathbf{W})$ is the n-by-n positive semidefinite matrix whose (i, j)-th entry is $\Theta_m(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W})$. It takes the form

$$\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}) = \frac{1}{d} \sum_{j=1}^{m} \lambda_{m,j} \operatorname{diag}\left(\boldsymbol{\sigma}'\left(\frac{\mathbf{X}\mathbf{w}_j}{\sqrt{d}}\right)\right) \mathbf{X}\mathbf{X}^\top \operatorname{diag}\left(\boldsymbol{\sigma}'\left(\frac{\mathbf{X}\mathbf{w}_j}{\sqrt{d}}\right)\right). \tag{5}$$

where $\operatorname{diag}(\mathbf{v})$ denotes an n-by-n diagonal matrix A with $A_{ii} = v_i$ for $\mathbf{v} = (v_1, \ldots, v_n)$.

Henceforth our main assumptions are:

Assumption 2.1 (Dataset)**.** (a) All inputs are non-zero and have norms at most 1: $0 < \|\mathbf{x}_i\| \leq 1$ for all $i \geq 1$. (b) For all $i \neq i'$ and $c \in \mathbb{R}$, $\mathbf{x}_i \neq c\mathbf{x}_{i'}$. (c) There is $C > 0$ such that $|y_i| \leq C$ for all $i \geq 1$.

Assumption 2.2 (Activation function)**.** The activation function is analytic, with $|\sigma'(x)| \leq 1$ and $|\sigma''(x)| \leq M$ for some $M > 0$, and it is not a polynomial.

Assumption 2.3 (Initialisation)**.** For $j \in [m]$, $\mathbf{w}_{0j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathrm{I}_d)$, where I_d is the d-by-d identity matrix.

3 Neural Tangent Kernel at initialisation and its limit

Mean NTG at initialisation and its minimum eigenvalue. Let $\mathbf{W}_0$ be a random initialisation from Assumption 2.3. Consider the mean NTK at initialisation $\Theta^*(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\Theta_m(\mathbf{x}, \mathbf{x}'; \mathbf{W}_0)\right]$. Then, Θ^* becomes the same as the limiting NTK under $1/\sqrt{m}$ scaling. Let $\widehat{\Theta}^*(\mathbf{X}) = \mathbb{E}\left[\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_0)\right]$ be the associated n-by-n mean NTG matrix at initialisation, whose (i, i')-th entry is $\Theta^*(\mathbf{x}_i, \mathbf{x}_{i'})$. Let $\kappa_n = \operatorname{eig}_{\min}(\widehat{\Theta}^*(\mathbf{X}))$ be the minimum eigenvalue of the mean NTG matrix at initialisation. This minimum eigenvalue plays an important role in the analysis of global convergence properties in the symmetrical NTK regime. Based on arguments from Du et al. (2019b,a), one has under Assumptions 2.1 to 2.3, that $\kappa_n > 0$.

Limiting NTG. To set the stage and give some intuition, we now describe the limiting behaviour of the NTG, for a fixed sample size n, as the width m goes to infinity. The proofs of all results are contained in the Supplementary Material.

Proposition 3.1. *Consider a sequence $(\mathbf{w}_{0j})_{j\geq 1}$ of iid random vectors distributed as in Assumption 2.3. Suppose Assumption 2.2 holds. Then,*

$$\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_0) \to \widehat{\Theta}_\infty(\mathbf{X}; \mathbf{W}_0) \tag{6}$$

almost surely as $m \to \infty$, where $\widehat{\Theta}_\infty(\mathbf{X}; \mathbf{W}_0) = \gamma\widehat{\Theta}^*(\mathbf{X}) + (1-\gamma)\widehat{\Theta}^{(2)}_\infty(\mathbf{X}; \mathbf{W}_0)$, with the following random positive semi-definite matrix $\widehat{\Theta}^{(2)}_\infty(\mathbf{X}; \mathbf{W}_0)$:

$$\widehat{\Theta}^{(2)}_\infty(\mathbf{X}; \mathbf{W}_0) = \frac{1}{d} \sum_{j=1}^{\infty} \widetilde{\lambda}_j \operatorname{diag}\left(\boldsymbol{\sigma}'\left(\frac{\mathbf{X}\mathbf{w}_{0j}}{\sqrt{d}}\right)\right) \mathbf{X}\mathbf{X}^\top \operatorname{diag}\left(\boldsymbol{\sigma}'\left(\frac{\mathbf{X}\mathbf{w}_{0j}}{\sqrt{d}}\right)\right).$$

Also, $\mathbb{E}[\widehat{\Theta}_\infty(\mathbf{X}; \mathbf{W}_0)] = \mathbb{E}[\widehat{\Theta}^{(2)}_\infty(\mathbf{X}; \mathbf{W}_0)] = \widehat{\Theta}^*(\mathbf{X})$, and

$$\mathbb{E}\left[\|\widehat{\Theta}_\infty(\mathbf{X}; \mathbf{W}_0) - \widehat{\Theta}^*(\mathbf{X})\|_F^2\right] = C_0(\mathbf{X})(1-\gamma)^2 \sum_{j\geq 1} \widetilde{\lambda}_j^2 \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $C_0(\mathbf{X}) \geq 0$ is some positive constant equal to

$$\sum_{1\leq i, i' \leq n} \left(\frac{\mathbf{x}_i^\top \mathbf{x}_{i'}}{d}\right)^2 \operatorname{Var}\left(\sigma'\left(\frac{1}{\sqrt{d}}\mathbf{w}_{01}^\top \mathbf{x}_i\right)\sigma'\left(\frac{1}{\sqrt{d}}\mathbf{w}_{01}^\top \mathbf{x}_{i'}\right)\right).$$

When $\gamma = 1$ (symmetric NTK scaling), the NTG converges to a constant matrix, and the solution obtained by GF coincides with that of kernel regression. When $\gamma < 1$, Proposition 3.1 shows that the NTG is random at initialisation, even in the infinite-width limit, suggesting that we are not operating in the kernel regime, asymptotically. As shown in Equation (7), the departure from the kernel regime, as measured by the total variance of the limiting random NTG, can be quantified by the nonnegative constant $(1-\gamma)^2\left(\sum_{j\geq 1}\widetilde{\lambda}_j^2\right) \in [0,1]$. When this constant is close to 0, we approach the kernel regime; increasing this value leads to a departure from the kernel regime, and increases the amount of feature learning (see Theorems 4.3 and 4.4). The quantity $\sum_{j\geq 1}\widetilde{\lambda}_j^2 \in (0,1]$ is always strictly positive. More rapid decrease of the $\widetilde{\lambda}_j$ as j increases will lead to higher values of $\sum_{j\geq 1}\widetilde{\lambda}_j^2$ as is illustrated in the Supplementary Material.

Having described the behaviour of the NTG at initialisation in the infinite-width limit, and provided intuition on the node scaling parameters, we are ready to state our main results on global convergence and feature learning properties of large FFNNs under such asymmetrical scaling.

4 Main results

Global convergence for gradient flow. Our main theorem, which is given below, explains what happens during training via GF. It says that with high probability, (i) the loss decays exponentially fast with respect to κ_n and the training time t, and (ii) the weights $\mathbf{w}_{tj}$ and the NTG matrix change by $O((n\lambda_{m,j}^{1/2})/(\kappa_n d^{1/2}\gamma))$ and $O((n^3\sum_{j=1}^m \lambda_{m,j}^2)/(\kappa_n^2 d^3\gamma^2) + (n^2\sqrt{\sum_{j=1}^m \lambda_{m,j}^2})/(\kappa_n d^2\gamma))$, respectively. Define

$$C_1 = \sup_{c\in(0,1]} \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma((cz)/\sqrt{d})^2]. \tag{8}$$

Theorem 4.1. *(Global convergence) Consider $\delta \in (0,1)$. Assume Assumptions 2.1 to 2.3, $\gamma > 0$, and*

$$m \geq \max\left(\frac{2^3 n\log\frac{2n}{\delta}}{\kappa_n d}, \frac{2^{10}n^3M^2(C^2+C_1)}{\kappa_n^3 d^3\gamma^2\delta}, \frac{2^{15}n^4M^2(C^2+C_1)}{\kappa_n^4 d^4\gamma^2\delta}\right)$$

where C is the bound on the y_i's in Assumption 2.1. Then, with probability at least $1-\delta$, the following properties hold for all $t \geq 0$:

(a) $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t)) \geq \frac{\gamma\kappa_n}{4}$;

(b) $L_m(\mathbf{W}_t) \leq e^{-(\gamma\kappa_n t)/2}L_m(\mathbf{W}_0)$;

(c) $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq \sqrt{\lambda_{m,j}} \times \frac{n}{\kappa_n d^{1/2}}\sqrt{\frac{2^7(C^2+C_1)}{\gamma^2\delta}}$ *for all $j \in [m]$;*

(d) $\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t) - \widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\|_2 \leq \left(\frac{2^7n^3M^2(C^2+C_1)}{\kappa_n^2 d^3\gamma^2\delta}\cdot\sum_{j=1}^m \lambda_{m,j}^2\right) + \left(\frac{2^5n^2M(C^2+C_1)^{1/2}}{\kappa_n d^2\gamma\delta^{1/2}}\cdot\sqrt{\sum_{j=1}^m \lambda_{m,j}^2}\right)$.

The theorem says that if $\gamma > 0$, the training error converges to 0 exponentially fast. Moreover, the weight change is bounded by a factor $\sqrt{\lambda_{m,j}}$ and the NTG change is bounded by a factor $\sqrt{\sum_{j=1}^m \lambda_{m,j}^2}$. Note that the upper bound in (c) vanishes in the limit if and only if $\gamma = 1$ (NTK regime); similarly, the upper bound in (d) vanishes if and only if $\gamma = 1$. Although we were not able to obtain matching lower bounds, we next argue that feature learning arises whenever $\gamma < 1$.

Remark 4.2. A result similar to the above holds for the ReLU activation function. Also, a result analogous to (b), showing global convergence also holds for GD. (See the Supplementary Material.)

Feature learning. Next, we state results about feature learning when $\gamma < 1$. We first show that on average, each individual weight in the network changes on the order of $\lambda_{m,j}$ by an infinitesimal gradient update. For $j \in [m]$, $k \in [d]$, let w_{0jk} be the k-th component of the weight vector $\mathbf{w}_{0j}$ at initialisation and define $g_1(\mathbf{x}) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z\|\mathbf{x}\|/\sqrt{d})\sigma'(z\|\mathbf{x}\|/\sqrt{d})\right]$.

Theorem 4.3. *Assume Assumption 2.2. For all $j \in [m]$ and $k \in [d]$, we have*

$$\mathbb{E}\left[\left.\frac{dw_{tjk}}{dt}\right|_{t=0}\right] = -\frac{\lambda_{m,j}}{\sqrt{d}}\sum_{i=1}^n x_{ik}g_1(\mathbf{x}_i).$$

Recall that $\lambda_{m,j} \to (1-\gamma)\widetilde{\lambda}_j$ as $m \to \infty$. So if $\gamma < 1$ and $\widetilde{\lambda}_j > 0$, the expected change of $\mathbf{w}_j$ for an infinitesimal update is non-zero in the infinite-width limit.

Next, we characterise the expected change of the NTK at time 0, using the neural tangent hierarchy (Huang and Yau, 2020). Define $g_2(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \mathbb{E}_{(z_1,z_2,z_3)}\left[\sigma''(z_1)\sigma'(z_2)\sigma'(z_3)\sigma(z_3)\right]$, where (z_1, z_2, z_3) is a centred Gaussian vector with covariance $\mathbb{E}[z_i z_j] = \mathbf{x}_i^\top \mathbf{x}_j/d$, for $1 \le i, j \le 3$.

Theorem 4.4. *Assume Assumption 2.2. For all* $\mathbf{x}_k, \mathbf{x}_\ell \in (\mathbb{R}^d \setminus \{0\})$, *we have*

$$\mathbb{E}\left[\left.\frac{d\Theta_m(\mathbf{x}_k, \mathbf{x}_\ell; \mathbf{W}_t)}{dt}\right|_{t=0}\right] = -\frac{\mathbf{x}_k^\top \mathbf{x}_\ell}{d^{3/2}}\left[\sum_{i=1}^n g_2(\mathbf{x}_k, \mathbf{x}_\ell, \mathbf{x}_i)\mathbf{x}_k^\top \mathbf{x}_i + g_2(\mathbf{x}_\ell, \mathbf{x}_k, \mathbf{x}_i)\mathbf{x}_\ell^\top \mathbf{x}_i\right]\sum_{j=1}^m \lambda_{m,j}^2.$$

The above theorem shows that the expected change to the NTK at initialisation is scaled by the factor $\sum_{j=1}^m \lambda_{m,j}^2$, which converges to $(1-\gamma)^2 \sum_j \widetilde{\lambda}_j^2$ as $m \to \infty$. The expected change in the NTK at the first GD iteration is therefore bounded away from zero when $\gamma < 1$.

References

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

P. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30: 87–201, 2021.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019a.

S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

J. Huang and H.-T. Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning*, pages 4542–4551. PMLR, 2020.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

S. Oymak and M. Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

P. Wolinski, G. Charpiat, and Y. Ollivier. Asymmetrical scaling layers for stable network pruning. *OpenReview Archive*, 2020.

G. Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 9947–9960, 2019.

This Supplementary Material is organised as follows. Appendix A presents additional convergence results and feature learning properties when the activation function is the (non-smooth) ReLU function. In particular, Theorem A.1 states conditions for the global convergence of gradient flow in the ReLU case, and is similar to Theorem 4.1 (smooth case) in the main paper. As noted in Appendix A.2, some of the propositions on feature learning also apply to the ReLU case. Appendix A.3 discusses some open problems in our framework when dealing with a ReLU activation function. Useful bounds and identities are presented in Appendix B. Appendix C gives a proof of the proposition regarding the structure of the limiting NTG at initialisation while Appendix D provides a secondary proposition regarding the minimum eigenvalue of the NTG at initialisation. Appendix E states and proves secondary lemmas on gradient flow dynamics. Appendix F and G give details of the main proof for global convergence of gradient flow, respectively for the ReLU and smooth case. The proofs are rather short and mostly build on the secondary lemmas and propositions of Appendices D and E. Appendix H gives a detailed proof for global convergence of gradient descent in the smooth case. The proof builds on results of convergence of gradient flow. Appendix I gives proofs of the theorems of Section 4 on feature learning. Appendix J provides experiments to illustrate the results, under a smooth activation, namely the Swish activation function. Finally, Appendix K provides experiments, but with a ReLU activation instead of the Swish activation function.

Contents

List of Figures

A Results for the ReLU activation function

Although we assume a smooth activation function in the main text of the paper (Assumption 2.2), some of the results remain true when we drop this assumption and use the ReLU activation function instead. In this section, we explain these results for ReLU. Throughout the section, we assume a weak derivative $\sigma'(x) = \mathbf{1}_{\{x>0\}}$ of the ReLU activation function σ.

A.1 Global convergence under gradient flow

Our global convergence theorem under gradient flow in the main text (Theorem 4.1) has a counterpart for the ReLU case, which is given below. This counterpart says that when we train the network with the ReLU activation, with high probability, the loss decays exponentially fast with respect to κ_n and the training time t, and the weights $\mathbf{w}_{tj}$ and the NTG matrix change by $O\left(\frac{n\lambda_{m,j}^{1/2}}{\kappa_n d^{1/2}}\right)$ and $O\left(\frac{n^2 \sum_{j=1}^m \lambda_{m,j}^{3/2}}{\kappa_n d^{3/2}} + \frac{n^{3/2}\sqrt{\sum_{j=1}^m \lambda_{m,j}^{3/2}}}{\kappa_n^{1/2} d^{5/4}}\right)$, respectively.

Theorem A.1 (Global convergence, gradient flow, ReLU)**.** *Consider $\delta \in (0,1)$. Let $D_0 = \sqrt{2C^2 + (2/d)}$. Assume Assumptions 2.1 and 2.3, and the use of the ReLU activation function. Also, assume $\gamma > 0$ and*

$$m \geq \max\left(\frac{2^3 n \log \frac{4n}{\delta}}{\kappa_n d}, \frac{2^{25} n^4 D_0^2}{\kappa_n^4 d^3 \gamma^2 \delta^5}, \frac{2^{35} n^6 D_0^2}{\kappa_n^6 d^5 \gamma^2 \delta^5}\right).$$

Then, with probability at least $1-\delta$, the following properties hold for all $t \geq 0$:

(a) $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t)) \geq \frac{\gamma\kappa_n}{4}$*;*
(b) $L_m(\mathbf{W}_t) \leq e^{-(\gamma\kappa_n t)/2} L_m(\mathbf{W}_0)$*;*
(c) $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq \frac{2^3 n D_0}{\kappa_n d^{1/2} \gamma \delta^{1/2}} \sqrt{\lambda_{m,j}}$ *for all $j \in [m]$;*
(d) $\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t) - \widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\|_2 \leq \left(\frac{2^9 n^2 D_0}{\kappa_n d^{3/2}\gamma\delta^{5/2}} \cdot \sum_{j=1}^m \lambda_{m,j}^{3/2}\right) + \left(\frac{2^6 n^{3/2} D_0^{1/2}}{\kappa_n^{1/2} d^{5/4} \gamma^{1/2} \delta^{5/4}} \cdot \sqrt{\sum_{j=1}^m \lambda_{m,j}^{3/2}}\right)$.

The proof of the theorem is given in Appendix F, and uses Lemmas E.1 to E.3 and Proposition D.1.

The theorem guarantees that whenever $\gamma > 0$, the training error converges to 0 exponentially fast. Also, it implies that the weight change is bounded by a factor $\sqrt{\lambda_{m,j}}$, and the NTG change is bounded by a factor $\sqrt{\sum_{j=1}^m \lambda_{m,j}^{3/2}}$. As we show in Appendix B.2, as m tends to ∞,

$$\lambda_{m,j} \to (1-\gamma)\widetilde{\lambda}_j \text{ for every } j \geq 1, \quad \text{and} \quad \sum_{j=1}^m \lambda_{m,j}^{3/2} \to (1-\gamma)^{3/2} \sum_{j=1}^\infty \widetilde{\lambda}_j^{3/2}.$$

Thus, when $\widetilde{\lambda}_j > 0$ (note that we necessarily have $\widetilde{\lambda}_1 > 0$), the upper bound in (c) is vanishing in the infinite-width limit if and only if $\gamma = 1$ (NTK regime); similarly, the upper bound in (d) is vanishing if and only if $\gamma = 1$. In fact, feature learning arises whenever $\gamma < 1$, since Theorem 4.3 in Section 4 holds for ReLU as we will explain in the next subsection.

A.2 Feature learning

Theorem 4.3 holds when we drop Assumption 2.2 and assume the use of the ReLU activation function instead. Furthermore, in the ReLU case, $g_1(\mathbf{x}) = \|\mathbf{x}\|/\sqrt{2\pi d}$. Thus, the expected change in the theorem has the following more specific form: for all $j \in [m]$ and $k \in [d]$,

$$\mathbb{E}\left[\left.\frac{dw_{tjk}}{dt}\right|_{t=0}\right] = -\frac{\lambda_{m,j}}{d\sqrt{2\pi}} \sum_{i=1}^n x_{ik}\|\mathbf{x}_i\|.$$

This ReLU version of Theorem 4.3 can be shown by the very proof of the original theorem given in Appendix I.1; the proof does not depend on whether we use ReLU or a smooth activation function satisfying Assumption 2.2.

A.3 Discussion

Theorem A.1 is the counterpart of Theorem 4.1 for the global convergence of gradient flow with a ReLU activation function. Despite empirical evidence from Appendix K suggesting that similar convergence results could potentially be applicable to GD in the ReLU context, we have yet to substantiate this with a comprehensive proof. The proof of the global convergence of GD with smooth activation provided in Appendix H relies on a Taylor approximation. This necessitates the activation function σ to be twice differentiable. It is worth noting that, in the symmetric NTK case, the global convergence of GD with a ReLU activation has been shown by Du et al. (2019b, Section 4). Their proof, however, critically relies on the fact that the weights remain stationary throughout the iterations of GD, which is not the scenario we are dealing with here when $\gamma > 0$. As such, it remains a compelling open question to determine whether the global convergence of GD can be proven within our specific framework when employing a ReLU activation function.

B Useful bounds and identities

B.1 Matrix Chernoff inequalities

The following matrix bounds can be found in (Tropp, 2012).

Proposition B.1. *Consider a finite sequence $(X_1, X_2, \dots, X_p)$ of independent, random, positive semi-definite $n \times n$ matrices with $\operatorname{eig}_{\max}(X_j) \le R$ almost surely for all $j \in [p]$, for some $R > 0$. Define*

$$\mu_{\min} = \operatorname{eig}_{\min}\left(\sum_{j=1}^{p} \mathbb{E}[X_j]\right) \quad \textit{and} \quad \mu_{\max} = \operatorname{eig}_{\max}\left(\sum_{j=1}^{p} \mathbb{E}[X_j]\right).$$

Then, for all $\delta \in [0, 1)$,

$$\Pr\left(\operatorname{eig}_{\min}\left(\sum_{j=1}^{p} X_j\right) \le (1-\delta)\mu_{\min}\right) \le n\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R} \le ne^{-\delta^2\mu_{\min}/(2R)}.$$

Also, for all $\delta \ge 0$,

$$\Pr\left(\operatorname{eig}_{\max}\left(\sum_{j=1}^{p} X_j\right) \ge (1+\delta)\mu_{\max}\right) \le n\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R} \le ne^{-\delta^2\mu_{\max}/((2+\delta)R)}.$$

B.2 Some identities on $(\lambda_{m,j})_{j\in[m]}$

The following proposition summarises a number of useful properties on the scaling parameters.

Proposition B.2. *For all $m \ge 1$,*

$$\sum_{j=1}^{m} \lambda_{m,j} = 1, \tag{9}$$

$$\sqrt{\gamma m} \le \sum_{j=1}^{m} \sqrt{\lambda_{m,j}} \le \sqrt{m}. \tag{10}$$

For every $r > 1$, as $m \to \infty$,

$$\sum_{j=1}^{m} \lambda_{m,j}^{r} \sim \sum_{j=1}^{m} \left(\lambda_{m,j}^{(2)}\right)^{r} \to (1-\gamma)^r \sum_{j\ge1} \widetilde{\lambda}_j^r. \tag{11}$$

Proof. Equation (9) follows from the definition of $\lambda_{m,j}$ as shown below:

$$\sum_{j=1}^{m} \lambda_{m,j} = \sum_{j=1}^{m}\left(\frac{\gamma}{m} + (1-\gamma)\frac{\widetilde{\lambda}_j}{\sum_{k=1}^{m}\widetilde{\lambda}_k}\right) = \gamma + (1-\gamma)\sum_{j=1}^{m}\frac{\widetilde{\lambda}_j}{\sum_{k=1}^{m}\widetilde{\lambda}_k} = \gamma + (1-\gamma) = 1.$$

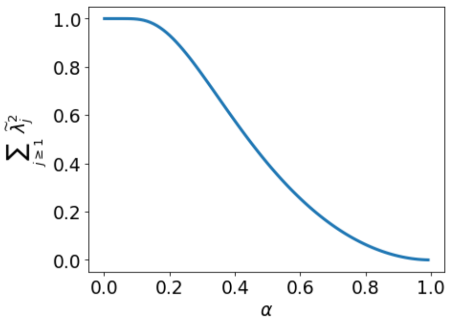

Figure 1: Value of $\sum_{j=1}^{\infty}\widetilde{\lambda}_j^2$ as a function of α, where $(\widetilde{\lambda}_j)_{j\geq 1}$ are defined as in Equation (35), As $\alpha \to 1$, it converges to 0, which corresponds to the kernel regime.

In Equation (10), the upper bound follows from Cauchy-Schwarz and Equation (9), and the lower bound from the definition of $\lambda_{m,j}$:

$$\sqrt{\gamma m} = \sum_{j=1}^{m}\sqrt{\frac{\gamma}{m}} \leq \sum_{j=1}^{m}\sqrt{\lambda_{m,j}} \leq \sqrt{\sum_{j=1}^{m}\lambda_{m,j}}\sqrt{\sum_{j=1}^{m}1} = 1\cdot\sqrt{m}.$$

For Equation (11), we note the following bounds on the sum of the $\lambda_{m,j}^r$ for all $r > 1$:

$$\sum_{j=1}^{m}\left(\lambda_{m,j}^{(2)}\right)^r \leq \sum_{j=1}^{m}\left(\lambda_{m,j}\right)^r \leq \left(\left[\sum_{j=1}^{m}\left(\lambda_{m,j}^{(1)}\right)^r\right]^{1/r} + \left[\sum_{j=1}^{m}\left(\lambda_{m,j}^{(2)}\right)^r\right]^{1/r}\right)^r$$

where the second inequality uses Minkowski inequality. But as $m \to \infty$, the term $\sum_{j=1}^{m}(\lambda_{m,j}^{(1)})^r = \gamma^r m^{-(r-1)} \to 0$. Furthermore, as $m \to \infty$,

$$\sum_{j=1}^{m}\left(\lambda_{m,j}^{(2)}\right)^r = \frac{(1-\gamma)^r}{\left(\sum_{k=1}^{m}\tilde{\lambda}_k\right)^r}\sum_{j=1}^{m}\tilde{\lambda}_j^r \to (1-\gamma)^r\sum_{j\geq 1}\tilde{\lambda}_j^r$$

because $(\sum_{k\geq 1}\tilde{\lambda}_k)^r = 1$. □

Figure 1 shows the value of $\sum_{j\geq 1}\widetilde{\lambda}_j^2 = \frac{\zeta(2/\alpha)}{\zeta(1/\alpha)^2}$ as a function of α, when using Zipf weights Equation (35).

C Proof of Proposition 3.1 on the limiting NTG

This proposition holds also under the ReLU activation case. In what follows, we will give a proof that works for both the smooth activation function and ReLU.

It is sufficient to look at the convergence of individual entries of the NTG matrix; that is, to show that, for each pair $1 \leq i, i' \leq n$,

$$\begin{aligned}\Theta_m(\mathbf{x}_i, \mathbf{x}_{i'}; \mathbf{W}_0) = \frac{\mathbf{x}_i^\top \mathbf{x}_{i'}}{d} \times \Bigg(&\frac{\gamma}{m}\sum_{j=1}^{m}\sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_j(\mathbf{x}_{i'};\mathbf{W}_0)) \\ &+ \frac{(1-\gamma)}{\sum_{k=1}^{m}\widetilde{\lambda}_k}\sum_{j=1}^{m}\widetilde{\lambda}_j\sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_j(\mathbf{x}_{i'};\mathbf{W}_0))\Bigg)\end{aligned} \tag{12}$$

tends to

$$\gamma\Theta^*(\mathbf{x}_i, \mathbf{x}_{i'}) + \frac{(1-\gamma)}{d}\mathbf{x}_i^\top \mathbf{x}_{i'}\sum_{j=1}^{\infty}\widetilde{\lambda}_j\sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_j(\mathbf{x}_{i'};\mathbf{W}_0)) \tag{13}$$

almost surely as $m \to \infty$. Using the fact that $|\sigma'(z)| \leq 1$ and the triangle inequality, the modulus of the difference between the RHS of Equation (12) and Equation (13) is upper bounded by

$$\begin{aligned}
&\left|\frac{\mathbf{x}_i^\top \mathbf{x}_{i'}}{d}\right| \left(\gamma \left| \left(\frac{1}{m}\sum_{j=1}^m \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_j(\mathbf{x}_{i'};\mathbf{W}_0))\right) - \mathbb{E}[\sigma'(Z_1(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_1(\mathbf{x}_{i'};\mathbf{W}_0))]\right|\right. \\
&\qquad\qquad \left. + (1-\gamma)\left[\left(\frac{1}{\sum_{j=1}^m \widetilde{\lambda}_j} - 1\right)\sum_{j=1}^m \widetilde{\lambda}_j + \sum_{j=m+1}^\infty \widetilde{\lambda}_j\right]\right) \\
&= \left|\frac{\mathbf{x}_i^\top \mathbf{x}_{i'}}{d}\right| \left(\gamma \left| \left(\frac{1}{m}\sum_{j=1}^m \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_j(\mathbf{x}_{i'};\mathbf{W}_0))\right) - \mathbb{E}[\sigma'(Z_1(\mathbf{x}_i;\mathbf{W}_0))\sigma'(Z_1(\mathbf{x}_{i'};\mathbf{W}_0))]\right|\right. \\
&\qquad\qquad \left. + 2(1-\gamma)\left[1 - \sum_{j=1}^m \widetilde{\lambda}_j\right]\right)
\end{aligned}$$

which tends to 0 almost surely as m tends to infinity using the law of large numbers and the fact that $\sum_{j=1}^\infty \widetilde{\lambda}_j = 1$.

D Secondary Proposition - NTG at initialisation

The following proposition is a corollary of Lemma 4 in (Oymak and Soltanolkotabi, 2020). It holds under both the ReLU and smooth activation cases. A proof is included for completeness.

Proposition D.1. *Let $\delta \in (0,1)$. Assume Assumptions 2.1 and 2.3, $\gamma > 0$, and $m \geq \frac{2^3 n \log \frac{n}{\delta}}{\kappa_n d}$. Also, assume that the activation function satisfies Assumption 2.2 or it is ReLU. Then, with probability at least $1-\delta$,*

$$\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)) \geq \mathrm{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)) > \frac{\gamma\kappa_n}{2} > 0.$$

Proof. We follow here the proof of Lemma 4 in (Oymak and Soltanolkotabi, 2020).

$$\begin{aligned}
\widehat{\Theta}_m(\mathbf{X};\mathbf{W}) &= \frac{1}{d}\sum_{j=1}^m \lambda_{m,j} A_j \\
&= \frac{1}{d}\sum_{j=1}^m \lambda_{m,j}^{(1)} A_j + \frac{1}{d}\sum_{j=1}^m \lambda_{m,j}^{(2)} A_j
\end{aligned}$$

where

$$A_j = \mathrm{diag}(\boldsymbol{\sigma}'(\mathbf{X}\mathbf{w}_j/\sqrt{d}))\mathbf{X}\mathbf{X}^\top \mathrm{diag}(\boldsymbol{\sigma}'(\mathbf{X}\mathbf{w}_j/\sqrt{d})).$$

Let $\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}) = \frac{1}{d}\sum_{j=1}^m \lambda_{m,j}^{(1)} A_j = \frac{\gamma}{md}\sum_{j=1}^m A_j$. Note that $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) \geq \mathrm{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}))$ a.s., and

$$\mathbb{E}[\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)] = \gamma\widehat{\Theta}^*(\mathbf{X})$$

where $\widehat{\Theta}^*(\mathbf{X})$ is defined in **??**. We have, for all $j \geq 1$,

$$\begin{aligned}
\|A_j\|_2 = \mathrm{eig}_{\max}(A_j) &\leq \mathrm{eig}_{\max}(\mathrm{diag}(\boldsymbol{\sigma}'(\mathbf{X}\mathbf{w}_j/\sqrt{d}))^2)\,\mathrm{eig}_{\max}(\mathbf{X}\mathbf{X}^\top) \leq \mathrm{eig}_{\max}(\mathbf{X}\mathbf{X}^\top) \\
&\leq \mathrm{trace}(\mathbf{X}\mathbf{X}^\top) \leq n.
\end{aligned} \tag{14}$$

At initialisation, $A_1, A_2, \ldots, A_m$ are independent random matrices. Using matrix Chernoff inequalities (see Proposition B.1), we obtain, for any $\epsilon \in [0,1)$,

$$\Pr\left(\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)) \leq (1-\epsilon)\gamma\kappa_n\right) \leq n e^{-\epsilon^2 m\kappa_n d/(2n)}.$$

Let $\delta \in (0,1)$. Taking $\epsilon = 1/2$, we have that, if $\frac{m\kappa_n d}{2^3 n} \geq \log\frac{n}{\delta}$, then

$$\Pr\left(\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)) \leq \frac{\gamma\kappa_n}{2}\right) \leq \delta.$$

□

E Secondary Lemmas on gradient flow dynamics

The proof technique used to prove Theorems 4.1 and A.1 is similar to that of (Du et al., 2019b) (NTK scaling). In particular, we provide in this section Lemmas similar to Lemmas 3.2, 3.3 and 3.4 in (Du et al., 2019b), but adapted to our setting. Lemma E.1 is an adaptation of Lemma 3.3. Lemmas E.2 and E.4 are adaptations of Lemma 3.2, respectively for the ReLU and smooth activation cases. Lemmas E.3 and E.5 are adaptations of Lemma 3.4, respectively for the ReLU and smooth activation cases.

E.1 Lemma on exponential decay of the empirical risk and scaling of the weight changes

The following lemma is an adaptation of Lemma 3.3 of (Du et al., 2019b), and applies to both the ReLU and smooth activation cases. It shows that, if the minimum eigenvalue of the NTG matrix is bounded away from 0, gradient flow converges to a global minimum exponentially fast. Recall that $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$.

Lemma E.1. *Let $t > 0$ and $\zeta > 0$. Assume Assumption 2.1 and $\operatorname{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)) \geq \frac{\zeta}{2}$ for all $0 \leq s \leq t$. Also, assume that the activation function satisfies Assumption 2.2 or it is ReLU. Then,*

$$L_m(\mathbf{W}_t) \leq e^{-\zeta t} L_m(\mathbf{W}_0),$$

and for all $j \in [m]$,

$$\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq \sqrt{\frac{n\lambda_{m,j}}{d}} \|\mathbf{y} - \mathbf{u}_0\| \frac{2}{\zeta}, \tag{15}$$

where $\mathbf{u}_0 = (f_m(\mathbf{x}_1; \mathbf{W}_0), \ldots, f_m(\mathbf{x}_n; \mathbf{W}_0))^\top \in \mathbb{R}^n$.

Proof. For $0 \leq s \leq t$, write $\mathbf{u}_s = (f_m(\mathbf{x}_1; \mathbf{W}_s), \ldots, f_m(\mathbf{x}_n; \mathbf{W}_s))^\top$. We have

$$\frac{d}{ds}\mathbf{u}_s = \widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)(\mathbf{y} - \mathbf{u}_s).$$

It follows that

$$\frac{dL_m(\mathbf{W}_s)}{ds} = -(\mathbf{y} - \mathbf{u}_s)^\top \widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)(\mathbf{y} - \mathbf{u}_s) \leq -\frac{\zeta}{2}(\mathbf{y} - \mathbf{u}_s)^\top(\mathbf{y} - \mathbf{u}_s) = -\zeta L_m(\mathbf{W}_s).$$

Using Grönwall's inequality, we obtain

$$L_m(\mathbf{W}_t) \leq e^{-\zeta t} L_m(\mathbf{W}_0).$$

For $0 \leq s \leq t$, using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\left\|\frac{d\mathbf{w}_{sj}}{ds}\right\|^2 &= \left\|\sqrt{\lambda_{m,j}} \frac{a_j}{\sqrt{d}} \sum_{i=1}^n \sigma'(Z_{sj}(\mathbf{x}_i))\mathbf{x}_i \cdot (y_i - f_m(\mathbf{x}_i; \mathbf{W}_s))\right\|^2 \\
&= \frac{\lambda_{m,j}}{d} \sum_{k=1}^d \left(\sum_{i=1}^n \sigma'(Z_{sj}(\mathbf{x}_i))x_{ik} \cdot (y_i - f_m(\mathbf{x}_i; \mathbf{W}_s))\right)^2 \\
&\leq \frac{\lambda_{m,j}}{d} \sum_{k=1}^d \left(\sum_{i=1}^n x_{ik}^2\right)\left(\sum_{i=1}^n \sigma'(Z_{sj}(\mathbf{x}_i))^2(y_i - f_m(\mathbf{x}_i; \mathbf{W}_s))^2\right) \\
&= \frac{\lambda_{m,j}}{d} \left(\sum_{i=1}^n \sigma'(Z_{sj}(\mathbf{x}_i))^2(y_i - f_m(\mathbf{x}_i; \mathbf{W}_s))^2\right)\left(\sum_{k=1}^d \sum_{i=1}^n x_{ik}^2\right) \\
&\leq \frac{\lambda_{m,j}}{d} \left(\sum_{i=1}^n (y_i - f_m(\mathbf{x}_i; \mathbf{W}_s))^2\right)\left(\sum_{i=1}^n \sum_{k=1}^d x_{ik}^2\right) \\
&\leq \frac{n\lambda_{m,j}}{d} \|\mathbf{y} - \mathbf{u}_s\|^2 \\
&\leq \frac{n\lambda_{m,j}}{d} \|\mathbf{y} - \mathbf{u}_0\|^2 e^{-\zeta s}.
\end{aligned}$$

Integrating and using Minkowski's integral inequality, we obtain

$$\begin{aligned}\|\mathbf{w}_{tj}-\mathbf{w}_{0j}\| = \left\|\int_0^t \frac{d}{ds}\mathbf{w}_{sj}ds\right\| &\le \int_0^t \left\|\frac{d}{ds}\mathbf{w}_{sj}\right\| ds \\ &\le \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y}-\mathbf{u}_0\| \int_0^t e^{-\zeta s/2}ds \\ &\le \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y}-\mathbf{u}_0\| \frac{2}{\zeta}.\end{aligned}$$

□

From now on, the proofs for the ReLU and smooth-activation cases slightly differ.

E.2 Lemma bounding the NTK change and minimum eigenvalue - ReLU case

The next lemma and its proof are similar to Lemma 3.2 in (Du et al., 2019b) and its proof. Recall that $0 < \|\mathbf{x}_i\| \le 1$ for every $i \in [n]$, and the $\mathbf{w}_{0j}$ are iid $\mathcal{N}(0, \mathrm{I}_d)$.

Lemma E.2. *Let $\delta \in (0,1)$, and $c_{m,j} > 0$ for every $j \in [m]$. Assume that Assumptions 2.1 and 2.3 holds and the activation function is ReLU. Then, with probability at least $1-\delta$, the following holds. For every $\mathbf{W} = (\mathbf{w}_1^\top, \dots, \mathbf{w}_m^\top)^\top$, if it satisfies*

$$\|\mathbf{w}_{0j}-\mathbf{w}_j\| \le \frac{\delta^2 c_{m,j}}{4} \quad \textit{for all } j \in [m],$$

we have

$$\left\|\widehat{\Theta}^{(s)}_m(\mathbf{X};\mathbf{W}) - \widehat{\Theta}^{(s)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 \le \frac{n}{d}\sum_{j=1}^m \lambda^{(k)}_{m,j}c_{m,j} + \frac{2n}{d}\sqrt{\sum_{j=1}^m \lambda^{(k)}_{m,j}c_{m,j}} \qquad \textit{for all } k \in [2]$$

and

$$\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) \ge \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0)) - \left(\frac{n\gamma}{dm}\sum_{j=1}^m c_{m,j} + \frac{2n\gamma}{dm^{1/2}}\sqrt{\sum_{j=1}^m c_{m,j}}\right). \tag{16}$$

Proof. For $k \in [2]$, let

$$f^{(k)}_m(-;\mathbf{W}) : \mathbb{R}^d \to \mathbb{R}, \qquad\qquad f^{(k)}_m(\mathbf{x};\mathbf{W}) = \sum_{j=1}^m \sqrt{\lambda^{(k)}_{m,j}}\,a_j\sigma(Z_j(\mathbf{x};\mathbf{W})).$$

Define $\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})$ to be the n-by-(md) matrix whose i-th row is the md-dimensional row vector $(\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{x}_i;\mathbf{W}))^\top$.

Note that for all $k \in [2]$,

$$\begin{aligned}&\left\|\widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}) - \widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 \\ &\quad= \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})^\top - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)^\top\right\|_2 \\ &\quad\le \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \qquad (17)\\ &\qquad + 2\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2.\end{aligned}$$

The justification of the inequality from above is given below (which is an expanded version of the three equations (364-366) in (Bartlett et al., 2021)): for all n-by-(pd) matrices A and B,

$$\begin{aligned}
\left\|AA^\top - BB^\top\right\|_2 &= \left\|\frac{1}{2}(A-B)(A+B)^\top + \frac{1}{2}(A+B)(A-B)^\top\right\|_2 \\
&\le \frac{1}{2}\left(\left\|(A-B)(A+B)^\top\right\|_2 + \left\|(A+B)(A-B)^\top\right\|_2\right) \\
&\le \frac{1}{2}\left(\|A-B\|_2 \times \left\|(A+B)^\top\right\|_2 + \|A+B\|_2 \times \left\|(A-B)^\top\right\|_2\right) \\
&= \|A-B\|_2 \times \|A+B\|_2 \\
&\le \|A-B\|_2 \times (\|A-B+B\|_2 + \|B\|_2) \\
&\le \|A-B\|_2 \times (\|A-B\|_2 + 2\|B\|_2)\,.
\end{aligned}$$

Coming back to the inequality in Equation (17), we next bound the two terms $\left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2$ and $\left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2$ there.

We bound the first term as follows:

$$\begin{aligned}
\left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \le \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_F^2 &= \sum_{i=1}^n \sum_{j=1}^m \left\|\nabla_{\mathbf{w}_j} f_m^{(k)}(\mathbf{x}_i;\mathbf{W}_0)\right\|^2 \\
&= \sum_{i=1}^n \sum_{j=1}^m \lambda_{m,j}^{(k)} \left|\sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right|^2 \frac{\|\mathbf{x}_i\|^2}{d} \\
&\le \frac{n}{d}\sum_{j=1}^m \lambda_{m,j}^{(k)} \le \frac{n}{d}\gamma_k
\end{aligned} \tag{18}$$

where $\gamma_1 = \gamma$ and $\gamma_2 = 1-\gamma$. The second inequality uses the assumption that $|\sigma'(x)| \le 1$ for all $x \in \mathbb{R}$ and $\|\mathbf{x}_i\| \le 1$ for all $i \in [n]$. The third inequality follows from the fact that $\sum_{j=1}^m \lambda_{m,j}^{(k)} \le \sum_{j=1}^m \lambda_{m,j} = 1$.

For the second term, we recall that $Z_j(\mathbf{x}_i;\mathbf{W}) = \frac{1}{\sqrt{d}}\mathbf{w}_j^\top \mathbf{x}_i$. Using this fact, we derive an upper bound for the second term as follows:

$$\begin{aligned}
&\left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \\
&\quad\le \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_F^2 \\
&\quad= \sum_{i=1}^n \sum_{j=1}^m \left\|\nabla_{\mathbf{w}_j} f_m^{(k)}(\mathbf{x}_i;\mathbf{W}) - \nabla_{\mathbf{w}_j} f_m^{(k)}(\mathbf{x}_i;\mathbf{W}_0)\right\|^2 \\
&\quad= \sum_{i=1}^n \sum_{j=1}^m \left\|\sqrt{\lambda_{m,j}^{(k)}}\, a_j \frac{\mathbf{x}_i}{\sqrt{d}}\left[\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right]\right\|^2 \\
&\quad= \frac{1}{d}\sum_{i=1}^n \sum_{j=1}^m \|\mathbf{x}_i\|^2 \lambda_{m,j}^{(k)} \left|\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right|^2.
\end{aligned} \tag{19}$$

In the rest of the proof, we will derive a probabilistic bound on the upper bound just obtained, and show the conclusions claimed in the lemma.

For any $\epsilon > 0$, $i \in [n]$, and $j \in [m]$, we define the event

$$A_{i,j}(\epsilon) = \left\{\exists \mathbf{w}_j \text{ s.t. } \|\mathbf{w}_{0j} - \mathbf{w}_j\| \le \epsilon \text{ and } \sigma'(\mathbf{w}_j^\top \mathbf{x}_i) \ne \sigma'(\mathbf{w}_{0j}^\top \mathbf{x}_i)\right\}.$$

If this event happens, we have $|\mathbf{w}_{0j}^\top \mathbf{x}_i| \le \epsilon$. To see this, assume that $A_{i,j}(\epsilon)$ holds with $\mathbf{w}_j$ as a witness of the existential quantification, and note that since the norm of $\mathbf{x}_i$ is at most 1,

$$\left|\mathbf{w}_{0j}^\top \mathbf{x}_i - \mathbf{w}_j^\top \mathbf{x}_i\right| \le \|\mathbf{w}_{0j} - \mathbf{w}_j\|\,\|\mathbf{x}_i\| \le \epsilon.$$

If $\mathbf{w}_{0j}^\top \mathbf{x}_i > 0$, then $\mathbf{w}_j^\top \mathbf{x}_i \leq 0$ and thus

$$\mathbf{w}_{0j}^\top \mathbf{x}_i \leq \epsilon + \mathbf{w}_j^\top \mathbf{x}_i < \epsilon.$$

Alternatively, if $\mathbf{w}_{0j}^\top \mathbf{x}_i \leq 0$, then $\mathbf{w}_j^\top \mathbf{x}_i > 0$ and thus

$$-\mathbf{w}_{0j}^\top \mathbf{x}_i \leq \epsilon - \mathbf{w}_j^\top \mathbf{x}_i \leq \epsilon.$$

In both cases, we have the desired $|\mathbf{w}_{0j}^\top \mathbf{x}_i| \leq \epsilon$.

Using the observation that we have just explained and the fact that $\mathbf{w}_{0j}^\top \mathbf{x}_i \sim \mathcal{N}(0, \|\mathbf{x}_i\|^2)$, we obtain, for a random variable $N \sim \mathcal{N}(0,1)$,

$$\begin{aligned}
\Pr(A_{i,j}(\epsilon)) \leq \Pr\left(|N| \leq \frac{\epsilon}{\|\mathbf{x}_i\|}\right) &= \operatorname{erf}\left(\frac{\epsilon}{\|\mathbf{x}_i\|\sqrt{2}}\right) \\
&\leq \sqrt{1 - \exp\left(-\left(4\left(\frac{\epsilon}{\|\mathbf{x}_i\|\sqrt{2}}\right)^2\right)/\pi\right)} \\
&\leq \sqrt{\frac{2\epsilon^2}{\|\mathbf{x}_i\|^2\pi}} \leq \frac{\epsilon}{\|\mathbf{x}_i\|},
\end{aligned} \tag{20}$$

where the second inequality uses $\operatorname{erf}(x) \leq \sqrt{1 - \exp(-(4x^2)/\pi)}$. Let $\Psi(\mathbf{W}_0)$ be the constraint on $\mathbf{W} = (\mathbf{w}_1^\top, \dots, \mathbf{w}_m^\top)^\top$ defined by

$$\mathbf{W} \in \Psi(\mathbf{W}_0) \iff \|\mathbf{w}_{0j'} - \mathbf{w}_{j'}\| \leq \frac{\delta^2 c_{m,j'}}{4} \text{ for all } j' \in [m].$$

Then, for all $k = 1, 2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2\right] \\
&\quad\leq \frac{1}{d}\sum_{i=1}^n\sum_{j=1}^m \|\mathbf{x}_i\|^2 \lambda_{m,j}^{(k)} \mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} |\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))|^2\right] \\
&\quad\leq \frac{1}{d}\sum_{i=1}^n\sum_{j=1}^m \|\mathbf{x}_i\|^2 \lambda_{m,j}^{(k)} \Pr\left(\exists \mathbf{W} \in \Psi(\mathbf{W}_0) \text{ s.t. } \sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) \neq \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right) \\
&\quad= \frac{1}{d}\sum_{i=1}^n\sum_{j=1}^m \|\mathbf{x}_i\|^2 \lambda_{m,j}^{(k)} \Pr\left(\exists \mathbf{w}_j \text{ s.t. } \|\mathbf{w}_{0j} - \mathbf{w}_j\| \leq \frac{\delta^2 c_{m,j}}{4} \text{ and } \sigma'(\mathbf{w}_j^\top \mathbf{x}_i) \neq \sigma'(\mathbf{w}_{0j}^\top \mathbf{x}_i)\right) \\
&\quad\leq \frac{1}{d}\sum_{i=1}^n\sum_{j=1}^m \|\mathbf{x}_i\|^2 \lambda_{m,j}^{(k)} \Pr\left(A_{i,j}(\delta^2 c_{m,j}/4)\right) \\
&\quad\leq \frac{(\delta^2/4)}{d}\sum_{i=1}^n\sum_{j=1}^m \|\mathbf{x}_i\| \lambda_{m,j}^{(k)} c_{m,j} \\
&\quad\leq \frac{n(\delta^2/4)}{d}\sum_{j=1}^m \lambda_{m,j}^{(k)} c_{m,j}.
\end{aligned}$$

The first inequality uses the bound in Equation (19), and the fourth inequality uses the inequality derived in Equation (20).

We bring together the bound on the expectation just shown and also the bounds proved in Equations (17) and (18). Recalling that $\gamma_1 = \gamma$ and $\gamma_2 = 1 - \gamma$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}) - \widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2\right] \\
&\qquad + 2\,\mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2 \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2\right] \\
&\qquad + 2\sqrt{\frac{n}{d}\gamma_k}\,\mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2\right] \\
&\qquad + 2\sqrt{\frac{n}{d}\gamma_k}\sqrt{\mathbb{E}\left[\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2\right]} \\
&\leq \frac{n(\delta^2/4)}{d}\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j} + 2\sqrt{\frac{n}{d}\gamma_k}\sqrt{\frac{n(\delta^2/4)}{d}\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j}} \\
&\leq \frac{\delta}{2}\left(\frac{n}{d}\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j} + \frac{2n}{d}\sqrt{\gamma_k\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j}}\right).
\end{aligned}$$

The third inequality uses Jensen's inequality, and the last uses the fact that $\delta/2 \geq (\delta/2)^2$. Hence, for each $k = 1, 2$, by Markov inequality, we have, with probability at least $1 - (\delta/2)$,

$$\sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\|\widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}) - \widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2 \leq \frac{n}{d}\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j} + \frac{2n}{d}\sqrt{\gamma_k}\sqrt{\sum_{j=1}^{m}\lambda_{m,j}^{(k)}c_{m,j}}.$$

By union bound, the conjunction of the above inequalities for the $k = 1$ and $k = 2$ cases holds with probability at least $1 - \delta$.

We prove the last remaining claim using the following lemma.

If A and B are real symmetric matrices, then

$$\mathrm{eig}_{\min}(A) \geq \mathrm{eig}_{\min}(B) - \|A - B\|_2,$$

which holds because

$$\begin{aligned}
\mathrm{eig}_{\min}(A) = \mathrm{eig}_{\min}(B + (A - B)) &\geq \mathrm{eig}_{\min}(B) + \mathrm{eig}_{\min}(A - B) \\
&\geq \mathrm{eig}_{\min}(B) - \mathrm{eig}_{\max}(B - A) \\
&\geq \mathrm{eig}_{\min}(B) - \|B - A\|_2 = \mathrm{eig}_{\min}(B) - \|A - B\|_2.
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\inf_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left(\mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}))\right) \\
&\quad \geq \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0)) - \sup_{\mathbf{W}\in\Psi(\mathbf{W}_0)} \left\| \widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}) - \widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0) \right\|_2 \\
&\quad \geq \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0)) - \left(\frac{n}{d}\sum_{j=1}^m \lambda^{(1)}_{m,j} c_{m,j} + \frac{2n}{d}\sqrt{\gamma \sum_{j=1}^m \lambda^{(1)}_{m,j} c_{m,j}} \right) \\
&\quad = \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0)) - \left(\frac{n\gamma}{dm}\sum_{j=1}^m c_{m,j} + \frac{2n\gamma}{dm^{1/2}}\sqrt{\sum_{j=1}^m c_{m,j}} \right)
\end{aligned}
$$

holds with probability at least $1-\delta$. Equation (16) then follows from the fact that for all $\mathbf{W}$, $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) \geq \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}))$. □

E.3 Lemma on a sufficient condition for Theorem A.1 - ReLU case

We now bring together the results from Proposition D.1 and Lemmas E.1 and E.2, and identify a sufficient condition for Theorem A.1, which corresponds to the condition in Lemma 3.4 in (Du et al., 2019b).

Lemma E.3. *Consider $\delta \in (0,1)$. Assume that Assumptions 2.1 and 2.3 hold, the activation function is ReLU, and $c_{m,j} > 0$ for all $j \in [m]$. Also, assume that $\gamma > 0$ and*

$$
m \geq \max\left(\left(\frac{8n\log\frac{4n}{\delta}}{d\kappa_n}\right), \left(\frac{8n}{d\kappa_n}\sum_{j=1}^m c_{m,j}\right), \left(\frac{16^2 n^2}{d^2\kappa_n^2}\sum_{j=1}^m c_{m,j}\right) \right).
$$

Define

$$
R'_{m,j} = \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y}-\mathbf{u}_0\|\,\frac{4}{\gamma\kappa_n} \quad \textit{and} \quad R_{m,j} = \frac{\delta^2 c_{m,j}}{64}.
$$

If $R'_{m,j} < R_{m,j}$ for all $j \in [m]$ with probability at least $1-\frac{\delta}{2}$, then on an event with probability at least $1-\delta$, we have that for all $j \in [m]$, $R'_{m,j} < R_{m,j}$ and the following properties also hold for all $t \geq 0$:

- *(a)* $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t)) \geq \frac{\gamma\kappa_n}{4}$;
- *(b)* $L_m(\mathbf{W}_t) \leq e^{-(\gamma\kappa_n t)/2} L_m(\mathbf{W}_0)$;
- *(c)* $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq R'_{m,j}$ *for all* $j \in [m]$*; and*
- *(d)* $\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t) - \widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\|_2 \leq \frac{n}{d}\sum_{j=1}^m \lambda_{m,j}c_{m,j} + \frac{2\sqrt{2}\cdot n}{d}\sqrt{\sum_{j=1}^m \lambda_{m,j}c_{m,j}}$.

Proof. Suppose $R'_{m,j} < R_{m,j}$ for all $j \in [m]$ on some event A' having probability at least $1-\frac{\delta}{2}$. Also, we would like to instantiate Proposition D.1 and Lemma E.2 with $\delta/4$, so that each of their claims holds with probability at least $1-\frac{\delta}{4}$. Let A be the intersection of A' with the event that the conjunction of the two claims in Proposition D.1 and Lemma E.2 hold with $\delta/4$. By the union bound, A has probability at least $1-\delta$. We will show that on the event A, the four claimed properties of the lemma hold.

It will be sufficient to show that

$$
\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\| \leq R_{m,j} \quad \text{for all } j \in [m] \text{ and } s \geq 0. \tag{21}
$$

To see why doing so is sufficient, pick an arbitrary $t_0 \geq 0$, and assume the above inequality for all $s \geq 0$. Then, by event A and Lemma E.2, for all $0 \leq s \leq t_0$, we have the following upper bound on the change of the Gram matrix from time 0 to s, and the following lower bound on the smallest eigenvalue of $\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)$:

$$\begin{aligned}\left\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s) - \widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 &\le \sum_{k=1}^{2} \left\|\widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}_s) - \widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2 \\ &\le \sum_{k=1}^{2} \left(\frac{n}{d}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j} + \frac{2n}{d}\sqrt{\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}}\right) \\ &\le \frac{n}{d}\sum_{j=1}^{m} \lambda_{m,j} c_{m,j} + \frac{2\sqrt{2}\cdot n}{d}\sqrt{\sum_{j=1}^{m} \lambda_{m,j} c_{m,j}}\end{aligned}$$

and

$$\begin{aligned}\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)) &\ge \mathrm{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)) - \left(\frac{n\gamma}{dm}\sum_{j=1}^{m} c_{m,j} + \frac{2n\gamma}{dm^{1/2}}\sqrt{\sum_{j=1}^{m} c_{m,j}}\right) \\ &\ge \frac{\gamma\kappa_n}{2} - \frac{\gamma\kappa_n}{4}\cdot\left(\frac{1}{m}\cdot\frac{4n}{d\kappa_n}\sum_{j=1}^{m} c_{m,j} + \frac{1}{m^{1/2}}\cdot\frac{8n}{d\kappa_n}\sqrt{\sum_{j=1}^{m} c_{m,j}}\right) \\ &\ge \frac{\gamma\kappa_n}{2} - \frac{\gamma\kappa_n}{4} = \frac{\gamma\kappa_n}{4}.\end{aligned}$$

We now apply Lemma E.1 with ζ being set to $\frac{\gamma\kappa_n}{2}$, which gives

$$L_m(\mathbf{W}_{t_0}) \le e^{-(\gamma\kappa_n t_0)/2} L_m(\mathbf{W}_0)$$

and

$$\|\mathbf{w}_{t_0 j} - \mathbf{w}_{0j}\| \le \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y} - \mathbf{u}_0\|\,\frac{4}{\gamma\kappa_n} = R'_{m,j} \quad \text{for all } j \in [m].$$

We have just shown that all the four properties in the lemma hold for t_0.

It remains to prove Equation (21) under the event A and the assumption that $R'_{m,j} < R_{m,j}$ for all $j \in [m]$ holds on this event. Suppose that Equation (21) fails for some $j \in [m]$. Let

$$t_1 = \inf\left\{t \mid \|\mathbf{w}_j - \mathbf{w}_{0j}\| > R_{m,j} \text{ for some } j \in [m]\right\}.$$

Then, by the continuity of $\mathbf{w}_{tj}$ on t, we have

$$\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\| \le R_{m,j} \quad \text{for all } j \in [m] \text{ and } 0 \le s \le t_1$$

and for some $j_0 \in [m]$,

$$\|\mathbf{w}_{t_1 j_0} - \mathbf{w}_{0 j_0}\| = R_{m,j_0}. \tag{22}$$

Thus, by the argument that we gave in the previous paragraph, we have

$$\|\mathbf{w}_{t_1 j} - \mathbf{w}_{0j}\| \le R'_{m,j} \quad \text{for all } j \in [m].$$

In particular, $\|\mathbf{w}_{t_1 j_0} - \mathbf{w}_{0 j_0}\| \le R'_{m,j_0}$. But this contradicts our assumption $R'_{m,j_0} < R_{m,j_0}$. □

E.4 Lemma bounding the NTK change and minimum eigenvalue - Smooth activation case

We now give a version of Lemma E.2 for the smooth activation case (that is, under Assumption 2.2). The proof of this version is similar to the one for Lemma 5 in (Oymak and Soltanolkotabi, 2020), and uses the three equations (364-366) in (Bartlett et al., 2021).

Lemma E.4. *Assume that Assumptions 2.1 to 2.3 hold. Let $c_{m,j} > 0$ for every $j \in [m]$. Then, for any fixed $\mathbf{W} = (\mathbf{w}_1^\top, \dots, \mathbf{w}_m^\top)^\top$, if it satisfies*

$$\|\mathbf{w}_{0j} - \mathbf{w}_j\| \le \frac{c_{m,j}}{2} \quad \textit{for all } j \in [m],$$

we have

$$\left\|\widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}) - \widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 \le \frac{nM^2}{4d^2}\sum_{j=1}^m \lambda^{(k)}_{m,j}c^2_{m,j} + \frac{nM}{d^{3/2}}\sqrt{\sum_{j=1}^m \lambda^{(k)}_{m,j}c^2_{m,j}} \qquad \textit{for all } k \in [2]$$

and

$$\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) \ge \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_0)) - \left(\frac{nM^2\gamma}{4d^2m}\sum_{j=1}^m c^2_{m,j} + \frac{nM\gamma}{d^{3/2}m^{1/2}}\sqrt{\sum_{j=1}^m c^2_{m,j}}\right). \tag{23}$$

Note that this lemma has a deterministic conclusion, although its original counterpart (Lemma E.2) has a probabilistic one.

Proof. The beginning part of the proof is essentially an abbreviated version of the beginning part of the proof of Lemma E.2. This repetition is intended to help the reader by not forcing her or him to look at the proof of Lemma E.2 beforehand.

For $k \in [2]$, let

$$f^{(k)}_m(-;\mathbf{W}) : \mathbb{R}^d \to \mathbb{R}, \qquad f^{(k)}_m(\mathbf{x};\mathbf{W}) = \sum_{j=1}^m \sqrt{\lambda^{(k)}_{m,j}}a_j\sigma(Z_j(\mathbf{x};\mathbf{W})),$$

and define $\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})$ to be the n-by-(pd) matrix whose i-th row is the pd-dimensional row vector $(\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{x}_i;\mathbf{W}))^\top$.

For all $k \in [2]$, we have

$$\begin{aligned}
&\left\|\widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}) - \widehat{\Theta}^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2 \\
&\quad = \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W})^\top - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)^\top\right\|_2 \\
&\quad \le \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \qquad (24)\\
&\qquad + 2\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2.
\end{aligned}$$

To see why this inequality holds, see the proof of Lemma E.2. We bound the two terms $\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2$ and $\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2$ in Equation (24). We bound the first term as follows:

$$\begin{aligned}
\left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \le \left\|\nabla_{\mathbf{W}} f^{(k)}_m(\mathbf{X};\mathbf{W}_0)\right\|_F^2 &= \sum_{i=1}^n\sum_{j=1}^m \left\|\nabla_{\mathbf{w}_j} f^{(k)}_m(\mathbf{x}_i;\mathbf{W}_0)\right\|^2 \\
&= \sum_{i=1}^n\sum_{j=1}^m \lambda^{(k)}_{m,j}\left|\sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right|^2\frac{\|\mathbf{x}_i\|^2}{d} \\
&\le \frac{n}{d}\sum_{j=1}^m \lambda^{(k)}_{m,j} \le \frac{n}{d}\gamma_k
\end{aligned}$$

where $\gamma_1 = \gamma$ and $\gamma_2 = 1-\gamma$. The second inequality uses the assumption that $|\sigma'(x)| \le 1$ for all $x \in \mathbb{R}$ and $\|\mathbf{x}_i\| \le 1$ for all $i \in [n]$. The third inequality holds because $\sum_{j=1}^m \lambda^{(k)}_{m,j} \le \sum_{j=1}^m \lambda_{m,j} = 1$. For the second term, we recall that $|\sigma''(x)| \le M$ and so σ' is M-Lipschitz, and also that $Z_j(\mathbf{x}_i;\mathbf{W}) = \frac{1}{\sqrt{d}}\mathbf{w}_j^\top\mathbf{x}_i$. Using these facts, we derive an upper bound for the second term as follows:

$$
\begin{aligned}
&\left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2^2 \\
&\quad\le \left\|\nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}) - \nabla_{\mathbf{W}} f_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_F^2 \\
&\quad= \sum_{i=1}^{n}\sum_{j=1}^{m} \left\|\nabla_{\mathbf{w}_j} f_m^{(k)}(\mathbf{x}_i;\mathbf{W}) - \nabla_{\mathbf{w}_j} f_m^{(k)}(\mathbf{x}_i;\mathbf{W}_0)\right\|^2 \\
&\quad= \sum_{i=1}^{n}\sum_{j=1}^{m} \left\|\sqrt{\lambda_{m,j}^{(k)}}\, a_j \frac{\mathbf{x}_i}{\sqrt{d}} \left[\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right]\right\|^2 \\
&\quad= \frac{1}{d}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \sum_{j=1}^{m} \lambda_{m,j}^{(k)} \left[\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right]^2 \\
&\quad\le \frac{1}{d}\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} \left[\sigma'(Z_j(\mathbf{x}_i;\mathbf{W})) - \sigma'(Z_j(\mathbf{x}_i;\mathbf{W}_0))\right]^2 \\
&\quad\le \frac{M^2}{d^2}\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} \left(\left(\mathbf{w}_j - \mathbf{w}_{0j}\right)^\top \mathbf{x}_i\right)^2 \\
&\quad\le \frac{nM^2}{d^2}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} \|\mathbf{w}_j - \mathbf{w}_{0j}\|^2 \\
&\quad\le \frac{nM^2}{4d^2}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}^2 .
\end{aligned}
$$

The second to last step uses the Cauchy-Schwartz inequality, and the last step uses our assumption that $\|\mathbf{w}_j - \mathbf{w}_{0j}\| \le \frac{c_{m,j}}{2}$ for all $j \in [m]$. From the derived bounds on the first and second terms in the last line of Equation (24), it follows that

$$
\begin{aligned}
\left\|\widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}) - \widehat{\Theta}_m^{(k)}(\mathbf{X};\mathbf{W}_0)\right\|_2 &\le \frac{nM^2}{4d^2}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}^2 + 2\sqrt{\frac{n}{d}\gamma_k}\sqrt{\frac{nM^2}{4d^2}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}^2} \\
&= \frac{nM^2}{4d^2}\sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}^2 + \frac{nM}{d^{3/2}}\sqrt{\gamma_k \sum_{j=1}^{m} \lambda_{m,j}^{(k)} c_{m,j}^2}.
\end{aligned}
$$

Finally, as noted in the proof of Lemma E.2, we have

$$
\begin{aligned}
\operatorname{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) &\ge \operatorname{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W})) \\
&\ge \operatorname{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)) - \left\|\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}) - \widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)\right\|_2 .
\end{aligned}
$$

Thus,

$$
\operatorname{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W})) \ge \operatorname{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X};\mathbf{W}_0)) - \left(\frac{nM^2\gamma}{4d^2 m}\sum_{j=1}^{m} c_{m,j}^2 + \frac{nM\gamma}{d^{3/2}m^{1/2}}\sqrt{\sum_{j=1}^{m} c_{m,j}^2}\right).
$$

□

E.5 Lemma on a sufficient condition for Theorem 4.1 - Smooth activation case

We now give a version of Lemma E.3 for the smooth activation case (i.e., under Assumption 2.2). It brings together the results from Proposition D.1 and Lemmas E.1 and E.4, and identifies a sufficient condition for Theorem A.1, which corresponds to the condition in Lemma 3.4 in (Du et al., 2019b).

Lemma E.5. *Assume that Assumptions 2.1 to 2.3 hold. Let $\delta \in (0,1)$, and $c_{m,j} > 0$ for all $j \in [m]$. Assume that $\gamma > 0$ and*

$$m \geq \max\left(\frac{8n\log\frac{2n}{\delta}}{d\kappa_n},\ \frac{nM^2\delta^2}{8d^2\kappa_n}\sum_{j=1}^m c_{m,j}^2,\ \frac{4n^2M^2\delta^2}{d^3\kappa_n^2}\sum_{j=1}^m c_{m,j}^2\right).$$

For each $j \in [m]$, define

$$R'_{m,j} = \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y} - \mathbf{u}_0\|\,\frac{4}{\gamma\kappa_n} \quad \textit{and} \quad R_{m,j} = \frac{\delta c_{m,j}}{8}.$$

If $R'_{m,j} < R_{m,j}$ for all $j \in [m]$ with probability at least $1 - \frac{\delta}{2}$, then on an event with probability at least $1 - \delta$, we have that for all $j \in [m]$, $R'_{m,j} < R_{m,j}$ and the following properties also hold for all $t \geq 0$:

(a) $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_t)) \geq \frac{\gamma\kappa_n}{4}$*;*

(b) $L_m(\mathbf{W}_t) \leq e^{-(\gamma\kappa_n t)/2} L_m(\mathbf{W}_0)$*;*

(c) $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq R'_{m,j}$ *for all $j \in [m]$; and*

(d) $\|\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_t) - \widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_0)\|_2 \leq \frac{nM^2\delta^2}{8^2d^2}\sum_{j=1}^m \lambda_{m,j}c_{m,j}^2 + \frac{nM\delta}{2^{3/2}d^{3/2}}\sqrt{\sum_{j=1}^m \lambda_{m,j}c_{m,j}^2}$*.*

Proof. The proof is very similar to that of Lemma E.3, although the concrete bounds in these proofs differ due to the differences between Lemma E.2 and Lemma E.4.

Suppose $R'_{m,j} < R_{m,j}$ for all $j \in [m]$ on some event A' having probability at least $1 - \frac{\delta}{2}$. Also, we would like to instantiate Proposition D.1 with $\delta/2$, so that its claim holds with probability at least $1 - \frac{\delta}{2}$. Let A be the intersection of A' with the event that claim in Proposition D.1 holds with $\delta/2$. By the union bound, A has probability at least $1 - \delta$. We will show that on the event A, the four claimed properties of the lemma hold.

It will be sufficient to show that

$$\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\| \leq R_{m,j} \quad \text{for all } s \geq 0. \tag{25}$$

To see why doing so is sufficient, pick an arbitrary $t_0 \geq 0$, and assume the above inequality for all $s \geq 0$. Then, by the event A and Lemma E.4, for all $0 \leq s \leq t_0$, we have the following upper bound on the change of the Gram matrix from time 0 to s, and the following lower bound on the smallest eigenvalue of $\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)$:

$$\begin{aligned}
&\left\|\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s) - \widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_0)\right\|_2 \\
&\quad \leq \left\|\widehat{\Theta}_m^{(1)}(\mathbf{X}; \mathbf{W}_s) - \widehat{\Theta}_m^{(1)}(\mathbf{X}; \mathbf{W}_0)\right\|_2 + \left\|\widehat{\Theta}_m^{(2)}(\mathbf{X}; \mathbf{W}_s) - \widehat{\Theta}_m^{(2)}(\mathbf{X}; \mathbf{W}_0)\right\|_2 \\
&\quad \leq \frac{nM^2\delta^2}{64d^2}\sum_{j=1}^m \lambda_{m,j}c_{m,j}^2 + \frac{nM\delta}{2^{3/2}d^{3/2}}\sqrt{\sum_{j=1}^m \lambda_{m,j}c_{m,j}^2}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)) &\geq \mathrm{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X}; \mathbf{W}_0)) - \left(\frac{nM^2\delta^2\gamma}{64d^2m}\sum_{j=1}^m c_{m,j}^2 + \frac{nM\delta\gamma}{4d^{3/2}m^{1/2}}\sqrt{\sum_{j=1}^m c_{m,j}^2}\right) \\
&> \frac{\gamma\kappa_n}{2} - \frac{\gamma\kappa_n}{4}\left(\frac{1}{m}\cdot\frac{nM^2\delta^2}{16d^2\kappa_n}\sum_{j=1}^m c_{m,j}^2 + \frac{1}{m^{1/2}}\cdot\frac{nM\delta}{d^{3/2}\kappa_n}\sqrt{\sum_{j=1}^m c_{m,j}^2}\right) \\
&\geq \frac{\gamma\kappa_n}{2} - \frac{\gamma\kappa_n}{4}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{\gamma\kappa_n}{4}.
\end{aligned}$$

We now apply the version of Lemma E.1 for the analytic activation σ, with ζ being set to $\frac{\gamma\kappa_n}{2}$. This application gives

$$L_m(\mathbf{W}_{t_0}) \le e^{-(\gamma\kappa_n t_0)/2} L_m(\mathbf{W}_0)$$

and

$$\|\mathbf{w}_{t_0 j} - \mathbf{w}_{0j}\| \le \sqrt{\frac{n\lambda_{m,j}}{d}}\, \|\mathbf{y} - \mathbf{u}_0\| \, \frac{4}{\gamma\kappa_n} = R'_{m,j} \quad \text{for all } j \in [m].$$

We have just shown that all the four properties in the lemma hold for t_0.

It remains to prove Equation (25) under the event A. Suppose that Equation (25) fails for some $j \in [m]$. Let

$$t_1 = \inf\left\{t \mid \|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| > R_{m,j} \text{ for some } j \in [m]\right\}.$$

Then, by the continuity of $\mathbf{w}_{tj}$ on t, we have

$$\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\| \le R_{m,j} \quad \text{for all } j \in [m] \text{ and } 0 \le s \le t_1$$

and for some $j_0 \in [m]$,

$$\|\mathbf{w}_{t_1 j_0} - \mathbf{w}_{0 j_0}\| = R_{m,j_0}. \tag{26}$$

Thus, by the argument that we gave in the previous paragraph, we have

$$\|\mathbf{w}_{t_1 j} - \mathbf{w}_{0j}\| \le R'_{m,j} \quad \text{for all } j \in [m].$$

In particular, $\|\mathbf{w}_{t_1 j_0} - \mathbf{w}_{0j_0}\| \le R'_{m,j_0}$. But this contradicts our assumption $R'_{m,j_0} < R_{m,j_0}$. □

F Proof of Theorem A.1 on the global convergence of gradient flow (ReLU case)

The proof of Theorem A.1 essentially follows Lemma E.3, which itself follows from the secondary Proposition D.1 and Lemmas E.1 and E.2, derived in Appendices D and E. Pick $\delta \in (0,1)$. Let

$$D = \sqrt{n^2\left(C^2 + \frac{1}{d}\right)\frac{2\cdot 512^2}{\gamma^2\delta^5\kappa_n^2 d}}$$

where C is the assumed upper bound on the $|y_i|$'s. Assume $\gamma > 0$ and

$$m \ge \max\left(\left(\frac{8n\log\frac{4n}{\delta}}{\kappa_n d}\right), \left(\frac{8nD}{d\kappa_n}\right)^2, \left(\frac{16^2 n^2 D}{d^2\kappa_n^2}\right)^2\right)$$

and set $c_{m,j}$ as follows:

$$c_{m,j} = \sqrt{\lambda_{m,j}} \cdot \sqrt{n^2\left(C^2 + \frac{1}{d}\right)\frac{2\cdot 512^2}{\gamma^2\delta^5\kappa_n^2 d}} = \sqrt{\lambda_{m,j}} \cdot D.$$

Note that

$$\begin{aligned}\left(\frac{8n}{d\kappa_n}\sum_{j=1}^m c_{m,j}\right)^2 = \left(\frac{8nD}{d\kappa_n}\right)^2 \cdot \left(\sum_{j=1}^m \sqrt{\lambda_{m,j}}\right)^2 &\le \left(\frac{8nD}{d\kappa_n}\right)^2 \cdot \left(\sum_{j=1}^m \lambda_{m,j}\right) \cdot m \\ &= \left(\frac{8nD}{d\kappa_n}\right)^2 \cdot m \le m^2,\end{aligned}$$

and also that

$$\begin{aligned}\left(\frac{16^2 n^2}{d^2\kappa_n^2}\sum_{j=1}^m c_{m,j}\right)^2 = \left(\frac{16^2 n^2 D}{d^2\kappa_n^2}\right)^2 \cdot \left(\sum_{j=1}^m \sqrt{\lambda_{m,j}}\right)^2 &\le \left(\frac{16^2 n^2 D}{d^2\kappa_n^2}\right)^2 \cdot \left(\sum_{j=1}^m \lambda_{m,j}\right) \cdot m \\ &= \left(\frac{16^2 n^2 D}{d^2\kappa_n^2}\right)^2 \cdot m \le m^2.\end{aligned}$$

Thus,

$$m \geq \max \left(\left(\frac{8n \log \frac{4n}{\delta}}{d\kappa_n} \right), \left(\frac{8n}{d\kappa_n} \sum_{j=1}^{m} c_{m,j} \right), \left(\frac{16^2 n^2}{d^2 \kappa_n^2} \sum_{j=1}^{m} c_{m,j} \right) \right).$$

As a result, we can now employ Lemma E.3. Thus, if we find an event A' such that the probability of A' is at least $1 - (\delta/2)$ and under A', we have $R'_{m,j} < R_{m,j}$, then the conclusion of Lemma E.3 holds. In particular, we may further calculate conclusions (c) and (d) of Lemma E.3 as

$$\begin{aligned} \|\mathbf{w}_{tj} - \mathbf{w}_{0j}\| \leq R'_{m,j} < R_{m,j} = \frac{\delta^2 c_{m,j}}{64} &= \frac{\delta^2}{64} \cdot \sqrt{\lambda_{m,j}} \cdot \sqrt{n^2 \left(C^2 + \frac{1}{d} \right) \frac{2 \cdot 512^2}{\gamma^2 \delta^5 \kappa_n^2 d}} \\ &= \frac{8n}{\kappa_n d^{1/2}} \cdot \sqrt{\left(C^2 + \frac{1}{d} \right) \frac{2}{\gamma^2 \delta}} \cdot \sqrt{\lambda_{m,j}}, \end{aligned}$$

and

$$\begin{aligned} \|\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_t) - \widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_0)\|_2 &\leq \frac{n}{d} \sum_{j=1}^{m} \lambda_{m,j} c_{m,j} + \frac{2\sqrt{2} \cdot n}{d} \sqrt{\sum_{j=1}^{m} \lambda_{m,j} c_{m,j}} \\ &= \frac{n}{d} \cdot D \cdot \sum_{j=1}^{m} \lambda_{m,j}^{3/2} + \frac{2\sqrt{2} \cdot n}{d} \cdot \sqrt{D} \cdot \sqrt{\sum_{j=1}^{m} \lambda_{m,j}^{3/2}} \\ &= \frac{n}{d} \cdot \sqrt{n^2 \left(C^2 + \frac{1}{d} \right) \frac{2 \cdot 512^2}{\gamma^2 \delta^5 \kappa_n^2 d}} \cdot \sum_{j=1}^{m} \lambda_{m,j}^{3/2} \\ &\quad + \frac{2\sqrt{2} \cdot n}{d} \cdot \left(n^2 \left(C^2 + \frac{1}{d} \right) \frac{2 \cdot 512^2}{\gamma^2 \delta^5 \kappa_n^2 d} \right)^{1/4} \cdot \sqrt{\sum_{j=1}^{m} \lambda_{m,j}^{3/2}} \\ &= \frac{512 n^2}{\kappa_n d^{3/2}} \cdot \sqrt{\left(C^2 + \frac{1}{d} \right) \frac{2}{\gamma^2 \delta^5}} \cdot \sum_{j=1}^{m} \lambda_{m,j}^{3/2} \\ &\quad + \frac{64 n^{3/2}}{\kappa_n^{1/2} d^{5/4}} \cdot \left(\left(C^2 + \frac{1}{d} \right) \frac{2}{\gamma^2 \delta^5} \right)^{1/4} \cdot \sqrt{\sum_{j=1}^{m} \lambda_{m,j}^{3/2}}. \end{aligned}$$

It remains to find such an event A'. Start by noting that

$$\begin{aligned} \mathbb{E}[\|\mathbf{y} - \mathbf{u}_0\|^2] &= \sum_{i=1}^{n} \left(y_i^2 - 2y_i \mathbb{E}[f_m(\mathbf{x}_i; \mathbf{W}_0)] + \mathbb{E}[f_m(\mathbf{x}_i; \mathbf{W}_0)^2] \right) \\ &= \sum_{i=1}^{n} \left(y_i^2 - 2y_i \cdot 0 + \mathbb{E} \left[\frac{1}{d} \sum_{j=1}^{m} \lambda_{m,j} (\mathbf{w}_j^\top \mathbf{x}_i)^2 \mathbf{1}_{\{\mathbf{w}_j^\top \mathbf{x}_i \geq 0\}} \right] \right) \\ &= \sum_{i=1}^{n} \left(y_i^2 + \frac{1}{d} \sum_{j=1}^{m} \lambda_{m,j} \mathbb{E} \left[(\mathbf{w}_j^\top \mathbf{x}_i)^2 \mathbf{1}_{\{\mathbf{w}_j^\top \mathbf{x}_i \geq 0\}} \right] \right) \\ &\leq n \left(C^2 + \frac{1}{d} \right). \end{aligned}$$

Thus, by Markov inequality, with probability at least $1 - (\delta/2)$,

$$\|\mathbf{y} - \mathbf{u}_0\|^2 < n \left(C^2 + \frac{1}{d} \right) \frac{2}{\delta}.$$

Let A' be the corresponding event for the above inequality. Then, under A', we have

$$\begin{aligned} R'_{m,j} &= \sqrt{\frac{n\lambda_{m,j}}{d}} \left\| \mathbf{y} - \mathbf{u}_0 \right\| \frac{4}{\gamma\kappa_n} \\ &< \sqrt{\frac{n\lambda_{m,j}}{d}} \cdot \sqrt{n\left(C^2 + \frac{1}{d}\right)\frac{2}{\delta}} \cdot \frac{4}{\gamma\kappa_n} \\ &= \sqrt{\lambda_{m,j}} \cdot \sqrt{n^2\left(C^2 + \frac{1}{d}\right)\frac{2\cdot 4^2}{\gamma^2\delta\kappa_n^2 d}} \\ &= \frac{\delta^2 c_{m,j}}{128} < \frac{\delta^2 c_{m,j}}{64} = R_{m,j}. \end{aligned}$$

Thus, A' is the desired event.

G Proof of Theorem 4.1 on the global convergence of gradient flow (smooth case)

The proof of the theorem is similar to that of Theorem A.1. It derives from Lemma E.5, which itself follows from the secondary Proposition D.1 and Lemmas E.1 and E.4, derived in Appendices D and E. Recall that

$$C_1 = \sup_{c\in(0,1]} \mathbb{E}[\sigma(cz)^2]$$

where the expectation is taken over the real-valued random variable z with the distribution $\mathcal{N}(0, 1/d)$. To see that C_1 is finite, note that since $|\sigma'(x)| \leq 1$ for all $x \in \mathbb{R}$, we have

$$|\sigma(cz) - \sigma(0)| \leq |cz| \text{ for all } c \in (0,1].$$

Thus, for every $c \in (0,1]$,

$$\sigma(0) - |cz| \leq \sigma(cz) \leq \sigma(0) + |cz|,$$

which implies that

$$\begin{aligned} \mathbb{E}[\sigma(cz)^2] &\leq \sigma(0)^2 + 2|\sigma(0)| \cdot |c| \cdot \mathbb{E}[|z|] + c^2\mathbb{E}[z^2] \\ &\leq \sigma(0)^2 + 2|\sigma(0)| \cdot \mathbb{E}[|z|] + \mathbb{E}[z^2]. \end{aligned}$$

As a result, $\mathbb{E}[\sigma(cz)^2]$ is bounded, so C_1 is finite.

Pick $\delta \in (0,1)$. Assume $\gamma > 0$ and

$$m \geq \max\left(\left(\frac{8n}{\kappa_n d} \cdot \log\frac{2n}{\delta}\right), \left(\frac{2^{10}n^3M^2}{\kappa_n^3 d^3} \cdot \frac{C^2 + C_1}{\gamma^2\delta}\right), \left(\frac{2^{15}n^4M^2}{\kappa_n^4 d^4} \cdot \frac{C^2 + C_1}{\gamma^2\delta}\right)\right)$$

and instantiate Lemma E.5 using the below $c_{m,j}$:

$$c_{m,j} = \sqrt{\lambda_{m,j}} \cdot \sqrt{n^2\left(C^2 + C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}}$$

where C is the assumed upper bound on the $|y_i|$'s. Note that

$$\begin{aligned} \frac{nM^2\delta^2}{8d^2\kappa_n}\sum_{j=1}^{m} c_{m,j}^2 &= \frac{nM^2\delta^2}{8d^2\kappa_n}\sum_{j=1}^{m}\left(\lambda_{m,j} \cdot n^2\left(C^2 + C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}\right) \\ &= \frac{nM^2\delta^2}{8d^2\kappa_n} \cdot n^2\left(C^2 + C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d} \cdot \sum_{j=1}^{m}\lambda_{m,j} \\ &= \frac{2^{10}n^3M^2}{\kappa_n^3 d^3} \times \frac{C^2 + C_1}{\gamma^2\delta} \end{aligned}$$

and

$$\frac{4n^2M^2\delta^2}{d^3\kappa_n^2}\sum_{j=1}^m c_{m,j}^2 = \frac{4n^2M^2\delta^2}{d^3\kappa_n^2}\sum_{j=1}^m \left(\lambda_{m,j}\cdot n^2\left(C^2+C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}\right)$$
$$= \frac{4n^2M^2\delta^2}{d^3\kappa_n^2}\cdot n^2\left(C^2+C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}\cdot\sum_{j=1}^m\lambda_{m,j}$$
$$= \frac{2^{15}n^4M^2}{\kappa_n^4 d^4}\times\frac{C^2+C_1}{\gamma^2\delta}.$$

Thus,

$$m \geq \max\left(\frac{8n\log\frac{2n}{\delta}}{d\kappa_n},\ \frac{nM^2\delta^2}{8d^2\kappa_n}\sum_{j=1}^m c_{m,j}^2,\ \frac{4n^2M^2\delta^2}{d^3\kappa_n^2}\sum_{j=1}^m c_{m,j}^2\right).$$

This allows us to employ Lemma E.5. Hence, it is sufficient to find an event A' such that the probability of A' is at least $1-(\delta/2)$ and under A', we have $R'_{m,j} < R_{m,j}$. The desired conclusion then follows from the conclusion of Lemma E.5, and the below calculations: if $\|\mathbf{w}_{tj}-\mathbf{w}_{0j}\| \leq R'_{m,j}$ and $R'_{m,j} < R_{m,j}$, then

$$\|\mathbf{w}_{tj}-\mathbf{w}_{0j}\| < R_{m,j} = \frac{\delta c_{m,j}}{8}$$
$$= \frac{\delta}{8}\cdot\sqrt{\lambda_{m,j}}\cdot\sqrt{n^2\left(C^2+C_1\right)\frac{2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}}$$
$$= \sqrt{\lambda_{m,j}}\times\frac{n}{\kappa_n d^{1/2}}\sqrt{\frac{128(C^2+C_1)}{\gamma^2\delta}},$$

and the upper bound on $\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t)-\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\|_2$ in the conclusion of Lemma E.5 can be rewritten to

$$\|\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t)-\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_0)\|_2$$
$$\leq \frac{nM^2\delta^2}{8^2d^2}\sum_{j=1}^m\lambda_{m,j}c_{m,j}^2 + \frac{nM\delta}{2^{3/2}d^{3/2}}\sqrt{\sum_{j=1}^m\lambda_{m,j}c_{m,j}^2}$$
$$= \frac{nM^2\delta^2}{4^3d^2}\sum_{j=1}^m\lambda_{m,j}\left(\lambda_{m,j}n^2\frac{(C^2+C_1)2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}\right)$$
$$+ \frac{nM\delta}{2^{3/2}d^{3/2}}\sqrt{\sum_{j=1}^m\lambda_{m,j}\left(\lambda_{m,j}n^2\frac{(C^2+C_1)2\cdot 64^2}{\gamma^2\delta^3\kappa_n^2 d}\right)}$$
$$= \left(\frac{n^3M^2}{\kappa_n^2d^3}\sum_{j=1}^m\lambda_{m,j}^2\frac{2^7(C^2+C_1)}{\gamma^2\delta}\right) + \frac{n^2M}{\kappa_n d^2}\sqrt{\sum_{j=1}^m\lambda_{m,j}^2\frac{2^{10}(C^2+C_1)}{\gamma^2\delta}}.$$

Note that

$$\mathbb{E}[\|\mathbf{y}-\mathbf{u}_0\|^2] = \sum_{i=1}^n\left(y_i^2 - 2y_i\mathbb{E}[f_m(\mathbf{x}_i;\mathbf{W}_0)] + \mathbb{E}[f_m(\mathbf{x}_i;\mathbf{W}_0)^2]\right)$$
$$= \sum_{i=1}^n\left(y_i^2 - 2y_i\cdot 0 + \mathbb{E}\left[\sum_{j=1}^m\lambda_{m,j}\sigma(Z_j(\mathbf{x}_i;\mathbf{W}_0))^2\right]\right)$$
$$= \sum_{i=1}^n\left(y_i^2 + \sum_{j=1}^m\lambda_{m,j}\mathbb{E}\left[\sigma(Z_j(\mathbf{x}_i;\mathbf{W}_0))^2\right]\right)$$
$$\leq n\left(C^2+C_1\right).$$

Thus, by Markov inequality, with probability at least $1 - (\delta/2)$,

$$\|\mathbf{y} - \mathbf{u}_0\|^2 < n\left(C^2 + C_1\right)\frac{2}{\delta}.$$

Let A' be the corresponding event for the above inequality. Then, under A', we have

$$\begin{aligned}
R'_{m,j} &= \sqrt{\frac{n\lambda_{m,j}}{d}}\,\|\mathbf{y} - \mathbf{u}_0\|\,\frac{4}{\gamma\kappa_n} \\
&< \sqrt{\frac{n\lambda_{m,j}}{d}} \cdot \sqrt{n\left(C^2 + C_1\right)\frac{2}{\delta}} \cdot \frac{4}{\gamma\kappa_n} \\
&= \sqrt{\lambda_{m,j}} \cdot \sqrt{n^2\left(C^2 + C_1\right)\frac{2 \cdot 4^2}{\gamma^2\delta\kappa_n^2 d}} \\
&= \frac{\delta c_{m,j}}{16} < \frac{\delta c_{m,j}}{8} = R_{m,j}.
\end{aligned}$$

Thus, A' is the desired event.

H Global convergence of gradient descent (smooth activation)

Let $\mathbf{y} \in \mathbb{R}^n$ be the vector $(y_1, \ldots, y_n)^\top$ of the outputs in the training dataset $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i \in [n]}$. For each gradient-descent step s, let $\mathbf{u}_s \in \mathbb{R}^n$ be the outputs at step s based on the inputs in $\mathcal{D}_n$, that is, $\mathbf{u}_s = (f_m(\mathbf{x}_1; \mathbf{W}_s), \ldots, f_m(\mathbf{x}_n; \mathbf{W}_s))^\top$. The following convergence theorem intuitively says that if the learning rate η of gradient descent is sufficiently small and the width of the network is large enough, then with high probability, the training error of the network decays exponentially fast to 0.

Theorem H.1. *Consider $\delta \in (0, 1)$. Assume Assumptions 2.1 to 2.3, $\gamma > 0$, and*

$$0 < \eta < \min\left(\frac{2}{\gamma\kappa_n}, \frac{\gamma\kappa_n d^2}{8n^2}, \frac{\gamma\kappa_n d^2 \delta^{1/2}}{2^{9/2} n^2 M (C^2 + C_1)^{1/2}}\right),$$

where C and C_1 are from Assumption 2.1 and Equation (8)*. Let $\alpha = \eta\gamma\kappa_n/2$ and $\beta = (1-\alpha)^{1/2}$. If*

$$m \geq \max\left(\frac{2^3 n \log\frac{2n}{\delta}}{\kappa_n d}, \frac{2^5\eta^2 n^3 M^2 (C^2 + C_1)}{\kappa_n d^3 (1-\beta)^2 \delta}, \frac{2^{11}\eta^2 n^4 M^2 (C^2 + C_1)}{\kappa_n^2 d^4 (1-\beta)^2 \delta}\right),$$

then with probability at least $1 - \delta$,

$$\|\mathbf{y} - \mathbf{u}_s\|^2 \leq (1-\alpha)^s \|\mathbf{y} - \mathbf{u}_0\|^2 \quad \textit{for all } s \in \mathbb{N} \cup \{0\}. \tag{27}$$

Note that the condition on the learning rate requires $\eta = O(\gamma\kappa_n/n^2)$. Thus, the best possible convergence rate from the theorem is $(1 - (\eta\gamma\kappa_n/2)) = (1 - (C_0\gamma^2\kappa_n^2/n^2))$ for some constant C_0.

The proof by induction on the gradient-descent step s and follows the structure of the convergence proof of (Du et al., 2019a, Theorem 5.1) with the necessary modifications, which in particular account for the changing weights and Gram matrices in our setup. That said, te two proofs differ significantly because, as in the case of gradient flow, the weights $\mathbf{w}_{sj}$ and the Gram matrix $\widehat{\Theta}_m(\mathbf{X}; \mathbf{W}_s)$ change during gradient descent in our case, while they remain almost constant in the case of (Du et al., 2019a).

H.1 Sketch of the proof

The proof is by induction on the number of gradient-update steps s. Here is a sketch of the proof for the inductive case. We start by decomposing the error at step $s+1$:

$$\begin{aligned}
\|\mathbf{y} - \mathbf{u}_{s+1}\|^2 &= \|(\mathbf{y} - \mathbf{u}_s) - (\mathbf{u}_{s+1} - \mathbf{u}_s)\|^2 \\
&= \|\mathbf{y} - \mathbf{u}_s\|^2 - 2(\mathbf{y} - \mathbf{u}_s)^\top(\mathbf{u}_{s+1} - \mathbf{u}_s) + \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2 \\
&= \|\mathbf{y} - \mathbf{u}_s\|^2 - 2(\mathbf{y} - \mathbf{u}_s)^\top\mathbf{I}_1 - 2(\mathbf{y} - \mathbf{u}_s)^\top\mathbf{I}_2 + \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2,
\end{aligned} \tag{28}$$

where $\mathbf{I}_1 = \eta\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)(\mathbf{y}-\mathbf{u}_s)$ and $\mathbf{I}_2 = (\mathbf{u}_{s+1} - \mathbf{u}_s - \mathbf{I}_1)$. We can then show that with high probability, both the third and the fourth terms in Equation (28) are $O(\eta^2)\|\mathbf{y}-\mathbf{u}_s\|^2$, so that the sum of these terms can be bounded from above by $(\eta\gamma\kappa_n/4)\|\mathbf{y}-\mathbf{u}_s\|^2$ if η is sufficiently small. On the other hand, the second term can be bounded using the minimum eigenvalue of the positive definite Gram matrix:

$$\begin{aligned}-2(\mathbf{y}-\mathbf{u}_s)^\top\mathbf{I}_1 &= \Big(-2\eta(\mathbf{y}-\mathbf{u}_s)^\top\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)(\mathbf{y}-\mathbf{u}_s)\Big)\\ &\le -2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s))\|\mathbf{y}-\mathbf{u}_s\|^2.\end{aligned}$$

We will show that if the network is large enough, with high probability, $-2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s))$ in the above upper bound is at most $-3\eta\gamma\kappa_n/4$. Putting all these together gives the required bound: with high probability,

$$\begin{aligned}\|\mathbf{y}-\mathbf{u}_{s+1}\|^2 &\le \|\mathbf{y}-\mathbf{u}_s\|^2 - 2(\mathbf{y}-\mathbf{u}_s)^\top\mathbf{I}_1 - 2(\mathbf{y}-\mathbf{u}_s)^\top\mathbf{I}_2 + \|\mathbf{u}_{s+1}-\mathbf{u}_s\|^2\\ &\le \|\mathbf{y}-\mathbf{u}_s\|^2 - \frac{3\eta\gamma\kappa_n}{4}\|\mathbf{y}-\mathbf{u}_s\|^2 + \frac{\eta\gamma\kappa_n}{4}\|\mathbf{y}-\mathbf{u}_s\|^2\\ &\le \left(1-\frac{\eta\gamma\kappa_n}{2}\right)\|\mathbf{y}-\mathbf{u}_s\|^2\\ &\le \left(1-\frac{\eta\gamma\kappa_n}{2}\right)^{s+1}\|\mathbf{y}-\mathbf{u}_0\|^2.\end{aligned}$$

The step of upper-bounding $-2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s))$ by $-3\eta\gamma\kappa_n/4$ is where we have to account for the changing weights and Gram matrix, and this is where the difference between our proof and that of (Du et al., 2019a) lies.

As mentioned already, the Gram matrix $\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)$ changes during gradient descent even when the network is very wide, but we will show that despite these changes, its minimum eigenvalue remains lower-bounded by $3\gamma\kappa_n/8$ with high probability. This can be done using the decomposition

$$\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_t) = \widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_t) + \widehat{\Theta}^{(2)}_m(\mathbf{X};\mathbf{W}_t). \tag{29}$$

where $\lambda^{(1)}_{m,j} = \gamma/m$ and $\lambda^{(2)}_{m,j} = ((1-\gamma)\widetilde{\lambda}_j)/\sum_{k=1}^m \widetilde{\lambda}_k$, and $\lambda^{(1)}_{m,j} + \lambda^{(2)}_{m,j} = \lambda_{m,j}$. At a high level, the reasoning goes like this. The induction hypothesis implies that the weight change $\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\|$ is $O(\sqrt{\lambda_{m,j}})$, which is small enough to guarantee that $\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_{s'})$ remains almost constant during training for a large network. This, in turn, implies that the minimum eigenvalue of $\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_s)$ is lower-bounded by $3\gamma\kappa_n/8$ with high probability. Since $\mathrm{eig}_{\min}(\widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)) \ge \mathrm{eig}_{\min}(\widehat{\Theta}^{(1)}_m(\mathbf{X};\mathbf{W}_s))$, we get the desired upper bound.

H.2 Two key lemmas

Before proving the theorem, we show two useful facts. Let $\mathbf{u}(\mathbf{W})$ be the n-dimensional vector

$$(f_m(\mathbf{x}_1;\mathbf{W}),\ldots,f_m(\mathbf{x}_n;\mathbf{W}))^\top$$

which consists of the network outputs on the training inputs under the parameters $\mathbf{W}$. Note that for each gradient-update step $s \in \mathbb{N}\cup\{0\}$, the vector $\mathbf{u}(\mathbf{W}_s)$ is equal to $\mathbf{u}_s$, the notation that we have been using in the main text of the paper. We also define $\mathbf{u}'(\mathbf{W})$ to be the following n-by-m matrix:

$$\mathbf{u}'(\mathbf{W}) = \frac{\partial\mathbf{u}}{\partial\mathbf{W}}.$$

For each $s \in \mathbb{N}\cup\{0\}$, let $\widehat{\Theta}_m(s) = \widehat{\Theta}_m(\mathbf{X};\mathbf{W}_s)$ and

$$\widetilde{\mathbf{u}}_{s+1} = \left(\mathbf{u}_s - \eta\frac{d\mathbf{u}_t}{dt}\Big|_{\mathbf{W}_t=\mathbf{W}_s}\right) = \Big(\mathbf{u}_s - \eta\widehat{\Theta}_m(s)(\mathbf{u}_s-\mathbf{y})\Big)$$

be the Euler discretisation of the gradient flow of the output. Here $\eta > 0$ is the learning rate.

Lemma H.2. *For all $\mathbf{W}$ and $j \in [m]$,*

$$\left\|\frac{\partial L_m(\mathbf{W})}{\partial\mathbf{w}_j}\right\| \le \frac{\sqrt{\lambda_{m,j}n}}{\sqrt{d}}\|\mathbf{y}-\mathbf{u}(\mathbf{W})\|.$$

Proof.

$$
\begin{aligned}
\left\|\frac{\partial L_m(\mathbf{W})}{\partial \mathbf{w}_j}\right\| &= \left\|\sum_{i=1}^{n}(u(\mathbf{W})_i - y_i) \times \sqrt{\lambda_{m,j}}a_j \times \sigma'\left(\frac{\mathbf{w}_j^\top \mathbf{x}_i}{\sqrt{d}}\right) \times \frac{\mathbf{x}_i}{\sqrt{d}}\right\| \\
&\le \sum_{i=1}^{n}\left\|(u(\mathbf{W})_i - y_i) \times \sqrt{\lambda_{m,j}}a_j \times \sigma'\left(\frac{\mathbf{w}_j^\top \mathbf{x}_i}{\sqrt{d}}\right) \times \frac{\mathbf{x}_i}{\sqrt{d}}\right\| \\
&\le \frac{\sqrt{\lambda_{m,j}}}{\sqrt{d}} \times \sum_{i=1}^{n}|u(\mathbf{W})_i - y_i| \\
&\le \frac{\sqrt{\lambda_{m,j} n}}{\sqrt{d}}\|\mathbf{y} - \mathbf{u}(\mathbf{W})\|.
\end{aligned}
$$

□

The next lemma gives an upper bound on $\|\mathbf{y} - \mathbf{u}_{s+1}\|$. As we will show shortly, this upper bound will play a crucial role in the proof of Theorem H.1.

Lemma H.3. *Assume Assumptions 2.1 to 2.3. Then, for all $s \in \mathbb{N} \cup \{0\}$, we have*

$$
\|\mathbf{y} - \mathbf{u}_{s+1}\|^2 \le \left(1 - 2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s)) + \frac{2\eta^2 M n^{3/2}}{d^2}\|\mathbf{y} - \mathbf{u}_s\| + \frac{\eta^2 n^2}{d^2}\right) \times \|\mathbf{y} - \mathbf{u}_s\|^2. \quad (30)
$$

Proof. Write

$$
\mathbf{u}_{s+1} - \mathbf{u}_s = \underbrace{\widetilde{\mathbf{u}}_{s+1} - \mathbf{u}_s}_{\mathbf{I}_1} + \underbrace{\mathbf{u}_{s+1} - \widetilde{\mathbf{u}}_{s+1}}_{\mathbf{I}_2}.
$$

Then, we have

$$
\begin{aligned}
\|\mathbf{y} - \mathbf{u}_{s+1}\|^2 &= \|(\mathbf{y} - \mathbf{u}_s) - (\mathbf{u}_{s+1} - \mathbf{u}_s)\|^2 \\
&= \|\mathbf{y} - \mathbf{u}_s\|^2 - 2(\mathbf{y} - \mathbf{u}_s)^\top(\mathbf{u}_{s+1} - \mathbf{u}_s) + \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2 \\
&= \|\mathbf{y} - \mathbf{u}_s\|^2 - 2(\mathbf{y} - \mathbf{u}_s)^\top\mathbf{I}_1 - 2(\mathbf{y} - \mathbf{u}_s)^\top\mathbf{I}_2 + \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2.
\end{aligned}
$$

Since the Gram matrix $\widehat{\Theta}_m(s)$ is positive definite and $\eta > 0$, we have

$$
\begin{aligned}
(\mathbf{y} - \mathbf{u}_s)^\top\mathbf{I}_1 = (\mathbf{y} - \mathbf{u}_s)^\top(\widetilde{\mathbf{u}}_{s+1} - \mathbf{u}_s) &= \eta(\mathbf{y} - \mathbf{u}_s)^\top\widehat{\Theta}_m(s)(\mathbf{y} - \mathbf{u}_s) \\
&\ge \eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))\,\|\mathbf{y} - \mathbf{u}_s\|^2.
\end{aligned}
$$

We now get a bound on $\mathbf{I}_2$. Note that $\widehat{\Theta}_m(s) = \mathbf{u}'_s(\mathbf{u}'_s)^\top$ where $\mathbf{u}'_s = \mathbf{u}'(\mathbf{W}_s) = \frac{\partial \mathbf{u}}{\partial \mathbf{W}}\big|_{\mathbf{W}=\mathbf{W}_s}$. Let

$$
L'_m(\mathbf{W}) = \frac{\partial L_m(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{n}(u(\mathbf{W})_i - y_i)u'(\mathbf{W})_i = \mathbf{u}'(\mathbf{W})^\top(\mathbf{u}(\mathbf{W}) - \mathbf{y})
$$

and

$$
L'_m(s) = L'_m(\mathbf{W}_s).
$$

Then,

$$
\begin{aligned}
\mathbf{I}_2 &= \mathbf{u}_{s+1} - \mathbf{u}_s + \eta\mathbf{u}'_s(\mathbf{u}'_s)^\top(\mathbf{u}_s - \mathbf{y}) \\
&= \left(-\int_{r=0}^{\eta}\Big(\mathbf{u}'\big(\mathbf{W}_s - rL'_m(s)\big)\Big)L'_m(s)\,dr\right) + \eta\mathbf{u}'_s(\mathbf{u}'_s)^\top(\mathbf{u}_s - \mathbf{y}) \\
&= \int_{r=0}^{\eta}\Big(\mathbf{u}'_s - \mathbf{u}'(\mathbf{W}_s - rL'_m(s))\Big)L'_m(s)\,dr.
\end{aligned}
$$

Also,

$$
\|L'_m(s)\| = \left\|\sum_{i=1}^{n}(y_i - u_{si})u'_{si}\right\| \le \sum_{i=1}^{n}|y_i - u_{si}|\,\|u'_{si}\|
$$

and

$$\|u'_{si}\|^2 = \sum_{j=1}^{m} \lambda_{m,j} a_j^2 \left(\sigma' \left(\frac{\mathbf{w}_{sj}^\top \mathbf{x}_i}{\sqrt{d}} \right) \right)^2 \frac{\|\mathbf{x}_i\|^2}{d} \leq \frac{1}{d},$$

since $\sum_j \lambda_{m,j} = 1$, $a_j \in \{-1, +1\}$, σ' is 1-Lipschitz, and $\|\mathbf{x}_i\| \leq 1$. Hence, by Cauchy-Schwarz,

$$\|L'_m(s)\| \leq \frac{1}{\sqrt{d}} \sum_{i=1}^{n} |y_i - u_{si}| \leq \frac{\sqrt{n}}{\sqrt{d}} \|\mathbf{y} - \mathbf{u}_s\|.$$

Let $\mathbf{W}_{(s,r)} = \mathbf{W}_s - rL'_m(s)$. For $j \in [m]$, write $\mathbf{w}_{(s,r)j}$ for the part of $\mathbf{W}_{(s,r)}$ going to the j-th node. Then, for all $i \in [n]$,

$$\begin{aligned}
\left\|u'_{si} - u'(\mathbf{W}_{(s,r)})_i\right\|^2 &= \sum_{j=1}^{m} \lambda_{m,j} a_j^2 \left(\sigma' \left(\frac{\mathbf{w}_{sj}^\top \mathbf{x}_i}{\sqrt{d}} \right) - \sigma' \left(\frac{\mathbf{w}_{(s,r)j}^\top \mathbf{x}_i}{\sqrt{d}} \right) \right)^2 \frac{\|\mathbf{x}_i\|^2}{d} \\
&\leq M^2 \sum_{j=1}^{m} \lambda_{m,j} a_j^2 \left(\left(\mathbf{w}_{sj} - \mathbf{w}_{(s,r)j} \right)^\top \mathbf{x}_i \right)^2 \frac{\|\mathbf{x}_i\|^2}{d^2} \\
&\leq \frac{M^2}{d^2} \sum_{j=1}^{m} \lambda_{m,j} \left\|\mathbf{w}_{sj} - \mathbf{w}_{(s,r)j}\right\|^2 \\
&\leq \frac{M^2}{d^2} \left\|\mathbf{W}_s - \mathbf{W}_{(s,r)}\right\|^2.
\end{aligned}$$

The first inequality follows from the M-Lipschitz continuity of σ', and the next inequality from $a_j \in \{-1, 1\}$, $\|\mathbf{x}_i\| \leq 1$, and Cauchy-Schwartz. The last inequality uses the fact that $\sum_j \lambda_{m,j} = 1$. Finally, for all $0 \leq r \leq \eta$,

$$\left\|\mathbf{W}_s - \mathbf{W}_{(s,r)}\right\| = r \|L'_m(s)\| \leq \eta \frac{\sqrt{n}}{\sqrt{d}} \|\mathbf{y} - \mathbf{u}_s\|.$$

Thus,

$$\begin{aligned}
\|\mathbf{I}_2\|^2 &= \sum_{i=1}^{n} \left(\int_{r=0}^{\eta} \left(u'_{si} - u'(\mathbf{W}_{(s,r)})_i \right)^\top L'_m(s)\, dr \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\int_{r=0}^{\eta} \left| \left(u'_{si} - u'(\mathbf{W}_{(s,r)})_i \right)^\top L'_m(s) \right| dr \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\int_{r=0}^{\eta} \|u'_{si} - u'(\mathbf{W}_{(s,r)})_i\| \times \|L'_m(s)\|\, dr \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\int_{r=0}^{\eta} \frac{\eta M \sqrt{n}}{d^{3/2}} \|\mathbf{y} - \mathbf{u}_s\| \times \frac{\sqrt{n}}{\sqrt{d}} \|\mathbf{y} - \mathbf{u}_s\|\, dr \right)^2 \\
&= \frac{\eta^4 M^2 n^3}{d^4} \|\mathbf{y} - \mathbf{u}_s\|^4 \\
&= \left(\frac{\eta^2 M n^{3/2}}{d^2} \|\mathbf{y} - \mathbf{u}_s\|^2 \right)^2.
\end{aligned}$$

As the upper bound depends quadratically on η, we can choose it small enough for gradient descent to converge, as we will show in the proof of Theorem H.1 in the next subsection.

Recall that $\|\mathbf{y} - \mathbf{u}_{s+1}\|^2$ can be expressed as the sum of four terms:

$$\|\mathbf{y} - \mathbf{u}_{s+1}\|^2 = \|\mathbf{y} - \mathbf{u}_s\|^2 - 2(\mathbf{y} - \mathbf{u}_s)^\top \mathbf{I}_1 - 2(\mathbf{y} - \mathbf{u}_s)^\top \mathbf{I}_2 + \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2. \tag{31}$$

Thus far, we have bounded the second and third terms on the RHS of Equation (31):

$$\begin{aligned}
-2(\mathbf{y} - \mathbf{u}_s)^\top \mathbf{I}_1 &\leq -2\eta \operatorname{eig}_{\min}(\widehat{\Theta}_m(s)) \|\mathbf{y} - \mathbf{u}_s\|^2, \\
-2(\mathbf{y} - \mathbf{u}_s)^\top \mathbf{I}_2 &\leq 2\|\mathbf{y} - \mathbf{u}_s\| \|\mathbf{I}_2\| \leq \left(\frac{2\eta^2 M n^{3/2}}{d^2} \|\mathbf{y} - \mathbf{u}_s\|^3 \right).
\end{aligned}$$

These bounds lead to the first three terms in the claimed upper bound of Equation (30). It remains to get an appropriate upper bound of the fourth term on the RHS of Equation (31).

Using the bound on the derivative of the loss in Lemma H.2, we complete the proof:

$$
\begin{aligned}
\|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2 &= \sum_{i=1}^{n} (u_{(s+1)i} - u_{si})^2 \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \sqrt{\lambda_{m,j}} a_j \left(\sigma\left(\frac{\mathbf{w}_{(s+1)j}^\top \mathbf{x}_i}{\sqrt{d}} \right) - \sigma\left(\frac{\mathbf{w}_{sj}^\top \mathbf{x}_i}{\sqrt{d}} \right) \right) \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \sqrt{\lambda_{m,j}} a_j \left| \sigma\left(\frac{\mathbf{w}_{(s+1)j}^\top \mathbf{x}_i}{\sqrt{d}} \right) - \sigma\left(\frac{\mathbf{w}_{sj}^\top \mathbf{x}_i}{\sqrt{d}} \right) \right| \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \sqrt{\lambda_{m,j}} a_j \left| \frac{\mathbf{w}_{(s+1)j}^\top \mathbf{x}_i}{\sqrt{d}} - \frac{\mathbf{w}_{sj}^\top \mathbf{x}_i}{\sqrt{d}} \right| \right)^2 \\
&\leq \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \frac{\sqrt{\lambda_{m,j}} a_j}{\sqrt{d}} \|\mathbf{w}_{(s+1)j} - \mathbf{w}_{sj}\| \|\mathbf{x}_i\| \right)^2 \\
&\leq \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \right) \times \left(\sum_{j=1}^{m} \frac{\sqrt{\lambda_{m,j}} a_j}{\sqrt{d}} \times \|\mathbf{w}_{(s+1)j} - \mathbf{w}_{sj}\| \right)^2 \\
&\leq n \times \left(\sum_{j=1}^{m} \frac{\sqrt{\lambda_{m,j}} a_j}{\sqrt{d}} \times \left\| \eta \frac{\partial L_m(\mathbf{W}_s)}{\partial \mathbf{w}_{sj}} \right\| \right)^2 \\
&\leq n \times \left(\sum_{j=1}^{m} \frac{\sqrt{\lambda_{m,j}} a_j}{\sqrt{d}} \times \frac{\eta \sqrt{\lambda_{m,j}} n}{\sqrt{d}} \|\mathbf{y} - \mathbf{u}_s\| \right)^2 \\
&\leq \frac{\eta^2 n^2}{d^2} \|\mathbf{y} - \mathbf{u}_s\|^2 \left(\sum_{j=1}^{m} \lambda_{m,j} \right)^2 \\
&= \frac{\eta^2 n^2}{d^2} \|\mathbf{y} - \mathbf{u}_s\|^2.
\end{aligned}
$$

□

H.3 Proof of Theorem H.1

Using the lemmas we have just shown, we will prove global convergence of gradient descent. Recall the assumed bound C on $|y_i|$ for every $i \geq 1$ in Assumption 2.1, and also

$$
C_1 = \sup_{c \in (0,1]} \mathbb{E}[\sigma(cz)^2]
$$

where the expectation is taken over the real-valued random variable z distributed as $\mathcal{N}(0, 1/d)$. As shown in Appendix G, C_1 is finite.

By the argument in Appendix G again, there exists an event E_1 such that E_1 happens with probability at least $1 - (\delta/2)$ and conditioned on E_1, we have

$$
\|\mathbf{y} - \mathbf{u}_0\| < \sqrt{n(C^2 + C_1)\frac{2}{\delta}}. \tag{32}
$$

Meanwhile, by Proposition D.1, there is an event E_2 such that E_2 happens with probability at least $1 - (\delta/2)$ and conditioned on E_2, we have

$$
\mathrm{eig}_{\min}(\widehat{\Theta}_m(0)) > \frac{\gamma \kappa_n}{2}. \tag{33}
$$

Let E_3 be the event that is the conjunction of E_1 and E_2. This event happens with probability at least $1 - \delta$, and under this event, Equations (32) and (33) both hold.

Condition on E_3. We prove the inequality in Equation (27) by induction on s. The base case of $s = 0$ is immediate. To prove the inductive case, assume that $s \geq 1$, and that the inequality in Equation (27) holds for all $s' = 0, 1, \ldots, s-1$.

Let

$$c_{m,j} = \frac{\eta n}{1-\beta}\sqrt{\frac{8\lambda_{m,j}(C^2 + C_1)}{\delta d}}.$$

Then,

$$\sum_{j=1}^{m} c_{m,j}^2 = \left(\frac{\eta^2 n^2}{(1-\beta)^2}\frac{8(C^2+C_1)}{\delta d}\sum_{j=1}^{m}\lambda_{m,j}\right) = \left(\frac{\eta^2 n^2}{(1-\beta)^2}\frac{8(C^2+C_1)}{\delta d}\right).$$

Note that for all $j \in [m]$,

$$\begin{aligned}
\|\mathbf{w}_{sj} - \mathbf{w}_{0j}\| &\leq \sum_{s'=0}^{s-1} \|\mathbf{w}_{(s'+1)j} - \mathbf{w}_{s'j}\| \\
&\leq \sum_{s'=0}^{s-1} \eta \left\| \frac{\partial L_m(\mathbf{W}_{s'})}{\partial \mathbf{w}_{s'j}} \right\| \\
&\leq \sum_{s'=0}^{s-1} \eta \sqrt{\frac{\lambda_{m,j} n}{d}} \|\mathbf{y} - \mathbf{u}_{s'}\| \\
&\leq \eta \sqrt{\frac{\lambda_{m,j} n}{d}} \sum_{s'=0}^{s-1} (1-\alpha)^{s'/2} \|\mathbf{y} - \mathbf{u}_0\| \\
&\leq \frac{\eta}{1-\beta} \sqrt{\frac{\lambda_{m,j} n}{d}} \|\mathbf{y} - \mathbf{u}_0\| \\
&\leq \frac{\eta}{1-\beta} \sqrt{\frac{\lambda_{m,j} n}{d}} \sqrt{n(C^2 + C_1)\frac{2}{\delta}} \\
&= \frac{1}{2} \times \frac{\eta n}{1-\beta} \sqrt{\frac{8\lambda_{m,j}(C^2+C_1)}{\delta d}} = \frac{c_{m,j}}{2}
\end{aligned}$$

where the third inequality uses the bound shown in Lemma H.2, the fourth inequality follows from the induction hypothesis, and the sixth inequality uses the bound in (32). Thus, by Lemma E.4 with $c_{m,j}$ from above and the lower bound on the minimum eigenvalue in Equation (33), we have

$$\begin{aligned}
&\mathrm{eig}_{\min}(\widehat{\Theta}_m(s)) \\
&\quad \geq \mathrm{eig}_{\min}(\widehat{\Theta}_m^{(1)}(\mathbf{X}; \mathbf{W}_0)) - \left(\frac{nM^2\gamma}{4d^2m}\sum_{j=1}^{m} c_{m,j}^2 + \frac{nM\gamma}{d^{3/2}m^{1/2}}\sqrt{\sum_{j=1}^{m} c_{m,j}^2}\right) \\
&\quad = \frac{\gamma\kappa_n}{2} - \left(\frac{nM^2\gamma}{4d^2m}\left(\frac{\eta^2 n^2}{(1-\beta)^2}\frac{8(C^2+C_1)}{\delta d}\right) + \frac{nM\gamma}{d^{3/2}m^{1/2}}\sqrt{\frac{\eta^2 n^2}{(1-\beta)^2}\frac{8(C^2+C_1)}{\delta d}}\right) \\
&\quad = \frac{\gamma\kappa_n}{2} - \left(\frac{2\eta^2 n^3 M^2 \gamma (C^2+C_1)}{d^3 m (1-\beta)^2 \delta} + \frac{\sqrt{8}\eta n^2 M \gamma (C^2+C_1)^{1/2}}{d^2 m^{1/2}(1-\beta)\delta^{1/2}}\right).
\end{aligned}$$

Meanwhile, by Lemma H.3, the induction hypothesis, and Equation (32),

$$\begin{aligned}
&\|\mathbf{y}-\mathbf{u}_{s+1}\|^2\\
&\leq \left(1-2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))+\frac{2\eta^2 M n^{3/2}}{d^2}\|\mathbf{y}-\mathbf{u}_s\|+\frac{\eta^2 n^2}{d^2}\right)\|\mathbf{y}-\mathbf{u}_s\|^2\\
&\leq \left(1-2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))+\frac{2\eta^2 M n^{3/2}}{d^2}(1-\alpha)^{s/2}\|\mathbf{y}-\mathbf{u}_0\|+\frac{\eta^2 n^2}{d^2}\right)\|\mathbf{y}-\mathbf{u}_s\|^2\\
&\leq \left(1-2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))+\frac{2\eta^2 M n^{3/2}}{d^2}(1-\alpha)^{s/2}\sqrt{n(C^2+C_1)\frac{2}{\delta}}+\frac{\eta^2 n^2}{d^2}\right)\|\mathbf{y}-\mathbf{u}_s\|^2.
\end{aligned}$$

Thus, we can complete the proof of this inductive case if we show that

$$\left(2\eta\,\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))-\frac{2\eta^2 M n^{3/2}}{d^2}(1-\alpha)^{s/2}\sqrt{n(C^2+C_1)\frac{2}{\delta}}-\frac{\eta^2 n^2}{d^2}\right)\geq\frac{\eta\gamma\kappa_n}{2}$$

which is equivalent to

$$\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))\geq\left(\frac{\eta M n^{3/2}}{d^2}(1-\alpha)^{s/2}\sqrt{n(C^2+C_1)\frac{2}{\delta}}+\frac{\eta n^2}{2d^2}+\frac{\gamma\kappa_n}{4}\right).$$

We will show this sufficient condition by proving the following stronger inequality (stronger because of the lower bound on $\mathrm{eig}_{\min}(\widehat{\Theta}_m(s))$ that we have derived above):

$$\begin{aligned}
\frac{\gamma\kappa_n}{2}-\left(\frac{2\eta^2 n^3 M^2\gamma(C^2+C_1)}{d^3 m(1-\beta)^2\delta}+\frac{\sqrt{8}\eta n^2 M\gamma(C^2+C_1)^{1/2}}{d^2 m^{1/2}(1-\beta)\delta^{1/2}}\right)&\\
\geq\left(\frac{\eta M n^{3/2}}{d^2}(1-\alpha)^{s/2}\sqrt{n(C^2+C_1)\frac{2}{\delta}}+\frac{\eta n^2}{2d^2}+\frac{\gamma\kappa_n}{4}\right)&,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\frac{\gamma\kappa_n}{4}\geq\Bigg(&\frac{2\eta^2 n^3 M^2\gamma(C^2+C_1)}{d^3 m(1-\beta)^2\delta}+\frac{\sqrt{8}\eta n^2 M\gamma(C^2+C_1)^{1/2}}{d^2 m^{1/2}(1-\beta)\delta^{1/2}}\\
&+\frac{\eta M n^{3/2}}{d^2}(1-\alpha)^{s/2}\sqrt{n(C^2+C_1)\frac{2}{\delta}}+\frac{\eta n^2}{2d^2}\Bigg).
\end{aligned}$$

But the four summands on the RHS of the above inequality are at most $\gamma\kappa_n/16$ by the assumed upper bound on η, the assumed lower bound on m, and the fact that $(1-\alpha)\leq 1$. Thus, the inequality from above holds, as desired.

I Proof of the results of Section 4 on feature learning

I.1 Proof of Theorem 4.3 (smooth and ReLU)

We have:

$$
\begin{aligned}
&\mathbb{E}\left[\left.\frac{dw_{sjk}}{ds}\right|_{s=0}\right] \\
&\quad= \mathbb{E}\left[\sum_{i=1}^{n}(y_i - f_m(\mathbf{x}_i;\mathbf{W}_0)) \cdot \left.\left(\frac{df_m(\mathbf{x}_i;\mathbf{W}_t)}{dw_{tjk}}\right)\right|_{t=0}\right] \\
&\quad= \sum_{i=1}^{n}\mathbb{E}\left[(y_i - f_m(\mathbf{x}_i;\mathbf{W}_0)) \cdot \sqrt{\lambda_{m,j}}a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)\frac{x_{ik}}{\sqrt{d}}\right] \\
&\quad= \sqrt{\frac{\lambda_{m,j}}{d}}\left(\sum_{i=1}^{n}\mathbb{E}\left[y_i a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)x_{ik}\right] - \sum_{i=1}^{n}\mathbb{E}\left[f_m(\mathbf{x}_i;\mathbf{W}_0)a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)x_{ik}\right]\right) \\
&\quad= -\sqrt{\frac{\lambda_{m,j}}{d}}\sum_{i=1}^{n}\mathbb{E}\left[f_m(\mathbf{x}_i;\mathbf{W}_0)a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)x_{ik}\right] \\
&\quad= -\sqrt{\frac{\lambda_{m,j}}{d}}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sum_{j'=1}^{m}\sqrt{\lambda_{m,j'}}a_{j'}\sigma\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)\right)a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)x_{ik}\right] \\
&\quad= -\sqrt{\frac{\lambda_{m,j}}{d}}\sum_{i=1}^{n}\mathbb{E}\left[\left(\sqrt{\lambda_{m,j}}a_j\sigma\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)\right)a_j\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)x_{ik}\right] \\
&\quad= -\frac{\lambda_{m,j}}{\sqrt{d}}\sum_{i=1}^{n}x_{ik}\mathbb{E}\left[\sigma\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top\mathbf{x}_i}{\sqrt{d}}\right)\right] \\
&\quad= -\frac{\lambda_{m,j}}{\sqrt{d}}\sum_{i=1}^{n}x_{ik}g_1(\mathbf{x}_i)
\end{aligned}
$$

I.2 Proof of Theorem 4.4 (smooth case)

We can write

$$
\frac{d\Theta_m(\mathbf{x}_\alpha,\mathbf{x}_\beta;\mathbf{W}_t)}{dt} = -\sum_{i=1}^{n}(f_m\mathbf{x}_i;\mathbf{W}_t) - y_i)\Theta_m^{(3)}(\mathbf{x}_\alpha,\mathbf{x}_\beta,\mathbf{x}_i;\mathbf{W}_t)
$$

where

$$
\Theta_m^{(3)}(\mathbf{x}_\alpha,\mathbf{x}_\beta,\mathbf{x}_i;\mathbf{W}_t) = \Big\langle \nabla_{\mathbf{W}}\Theta_m(\mathbf{x}_\alpha,\mathbf{x}_\beta;\mathbf{W}_t),\, \nabla_{\mathbf{W}}f(\mathbf{x}_i;\mathbf{W}_t)\Big\rangle. \tag{34}
$$

We have

$$
\begin{aligned}
&\nabla_{\mathbf{W}}\Theta_m(\mathbf{x}_\alpha,\mathbf{x}_\beta;\mathbf{W}_t) = \\
&\qquad\Big(\nabla^2_{\mathbf{W}}f_m(\mathbf{x}_\alpha;\mathbf{W}_t)\Big)\nabla_{\mathbf{W}}f_m(\mathbf{x}_\beta;\mathbf{W}_t) + \Big(\nabla^2_{\mathbf{W}}f_m(\mathbf{x}_\beta;\mathbf{W}_t)\Big)\nabla_{\mathbf{W}}f_m(\mathbf{x}_\alpha;\mathbf{W}_t).
\end{aligned}
$$

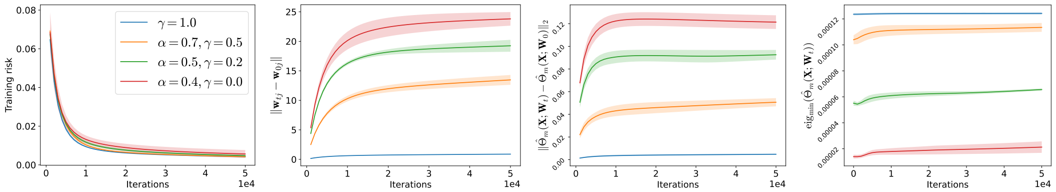

Figure 2: Results on simulated data. From left to right, 1) training risks, 2) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with the j's being those neurons which have maximal differences at the end of the training, 3) differences in NTG matrices, and 4) minimum eigenvalues of NTG matrices.

Thus,

$$
\begin{aligned}
&\Theta_m^{(3)}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_i; \mathbf{W}_t) \\
&= \Big(\nabla_{\mathbf{W}} f_m(\mathbf{x}_i; \mathbf{W}_t)\Big)^\top \Big(\nabla^2_{\mathbf{W}} f(\mathbf{x}_\alpha; \mathbf{W}_t)\Big)\Big(\nabla_{\mathbf{W}} f_m(\mathbf{x}_\beta; \mathbf{W}_t)\Big) \\
&\qquad\qquad + \Big(\nabla_{\mathbf{W}} f_m(\mathbf{x}_i; \mathbf{W}_t)\Big)^\top \Big(\nabla^2_{\mathbf{W}} f(\mathbf{x}_\beta; \mathbf{W}_t)\Big)\Big(\nabla_{\mathbf{W}} f_m(\mathbf{x}_\alpha; \mathbf{W}_t)\Big) \\
&= \sum_{j=1}^m a_j \frac{\lambda_{m,j}^{3/2}}{d^{3/2}} \sigma'\left(\frac{\mathbf{w}_{tj}^\top \mathbf{x}_i}{\sqrt{d}}\right) \mathbf{x}_\alpha^\top \mathbf{x}_\beta \\
&\qquad\qquad \times \left(\mathbf{x}_\alpha^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{tj}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{tj}^\top \mathbf{x}_\beta}{\sqrt{d}}\right) + \mathbf{x}_\beta^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{tj}^\top \mathbf{x}_\beta}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{tj}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\right).
\end{aligned}
$$

We therefore calculate

$$
\begin{aligned}
&\mathbb{E}\left[\left.\frac{d\Theta_m(\mathbf{x}_\alpha, \mathbf{x}_\beta; \mathbf{W}_t)}{dt}\right|_{t=0}\right] \\
&= \mathbb{E}\left[-\sum_{i=1}^n (f_m(\mathbf{x}_i; \mathbf{W}_0) - y_i)\Theta_m^{(3)}(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_i; \mathbf{W}_0)\right] \\
&= -\sum_{j=1}^m \frac{\lambda_{m,j}^{3/2}}{d^{3/2}} \mathbf{x}_\alpha^\top \mathbf{x}_\beta \sum_{i=1}^n \mathbb{E}\Bigg[a_j \left(\sum_{j'=1}^m \sqrt{\lambda_{m,j'}} a_{j'} \sigma\left(\frac{\mathbf{w}_{0j'}^\top \mathbf{x}_i}{\sqrt{d}}\right)\right) \\
&\qquad\qquad \times \Bigg(\mathbf{x}_\alpha^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\beta}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_i}{\sqrt{d}}\right) \\
&\qquad\qquad\qquad + \mathbf{x}_\beta^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\beta}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_i}{\sqrt{d}}\right)\Bigg)\Bigg] \\
&= \sum_{j=1}^m \frac{\lambda_{m,j}^2}{d^{3/2}} \mathbf{x}_\alpha^\top \mathbf{x}_\beta \sum_{i=1}^n \mathbb{E}\Bigg[\sigma\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_i}{\sqrt{d}}\right)\Bigg(\mathbf{x}_\alpha^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\beta}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_i}{\sqrt{d}}\right) \\
&\qquad\qquad\qquad + \mathbf{x}_\beta^\top \mathbf{x}_i \sigma''\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\beta}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_\alpha}{\sqrt{d}}\right)\sigma'\left(\frac{\mathbf{w}_{0j}^\top \mathbf{x}_i}{\sqrt{d}}\right)\Bigg)\Bigg] \\
&= -\frac{\mathbf{x}_\alpha^\top \mathbf{x}_\beta}{d^{3/2}}\left[\mathbf{x}_\alpha^\top\left(\sum_{i=1}^n \mathbf{x}_i g_2(\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_i)\right) + \mathbf{x}_\beta^\top\left(\sum_{i=1}^n \mathbf{x}_i g_2(\mathbf{x}_\beta, \mathbf{x}_\alpha, \mathbf{x}_i)\right)\right]\sum_{j=1}^m \lambda_{m,j}^2.
\end{aligned}
$$

J Experimental results (smooth activation)

We use here a (smooth) swish activation function $\sigma(z) = z/(1 + e^{-z})$. We obtained quantitatively similar results with a ReLU activation function; see Appendix K.

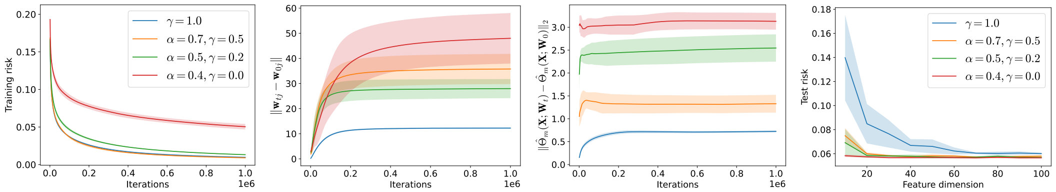

Figure 3: A subset of results for the regression experiments. From left to right, 1) training risks for `concrete` dataset, 2) the differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training for `energy` dataset, 3) the differences in NTG matrices for `airfoil` dataset, 4) test risks of transferred models for `plant` dataset.

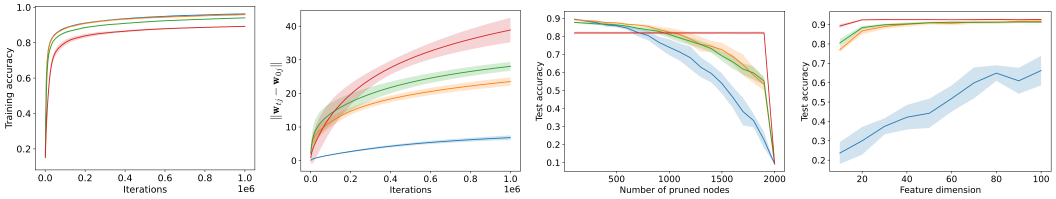

Figure 4: A subset of results for MNIST dataset. From left to right, 1) training risks, 2) difference in weight norms, 3) the test accuracies of the pruned models, 4) test accuracies of transferred models.

Simulated data. We first illustrate our theory on simulated data.[1] We generate $n = 100$ observations where for $i = 1, \ldots, n$, $\mathbf{x}_i$ is $d = 50$ dimensional and sampled uniformly on the unit sphere and $y_i = \frac{5}{d} \sum_{j=1}^{d} \sin(\pi x_{i,j}) + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We use the FFNN of Section 2, with the swish activation function, $m = 2000$ hidden nodes and $\lambda_{m,j}$ as in Equation (1) and

$$\widetilde{\lambda}_j = \frac{1}{\zeta(1/\alpha)} \frac{1}{j^{1/\alpha}}, \quad j \geq 1 \tag{35}$$

where $\gamma \in [0, 1]$ and $\alpha \in (0, 1)$. We consider the four values $(\gamma, \alpha) \in \{(1, -), (0.5, 0.7), (0.5, 0.5), (0, 0.4)\}$. For each setting, we run GD with a learning rate of 1.0 for 50 000 steps, which is repeated five times to get average results. We summarise the results in Figure 2, which shows the training error and the evolution of the weights, NTG, and minimum eigenvalue of the NTG as a function of the GD iterations. We see a clear correspondence between the theory and the empirical results. For $\gamma > 0$, GD achieves near-zero training error. The minimum eigenvalue and the training rates increase with the value of γ. For $\gamma = 1$, we have the highest minimum eigenvalue and the fastest training rate; however, there is no/very little feature learning: the weights and the NTG do not change significantly over the GD iterations. When $\gamma < 1$, there is clear evidence of feature learning: both the weights and the NTG change significantly over time; the smaller γ and α, the more feature learning arises.

Regression. We also validate our model on four real-world regression datasets from the UCI repository[2]: `concrete` ($(n, d) = (1030, 9)$), `energy` ($(n, d) = (768, 8)$), `airfoil` ($(n, d) = (1503, 6)$), and `plant` $(n, d) = (9568, 4)$). We split each dataset into training (40%) , test (20%), and validation sets (40%), and the validation set is used to test transfer learning. We use the same

[1] The code can found at `https://github.com/AnomDoubleBlind/asymmetrical_scaling/`

[2] `https://archive.ics.uci.edu/ml/datasets.php`

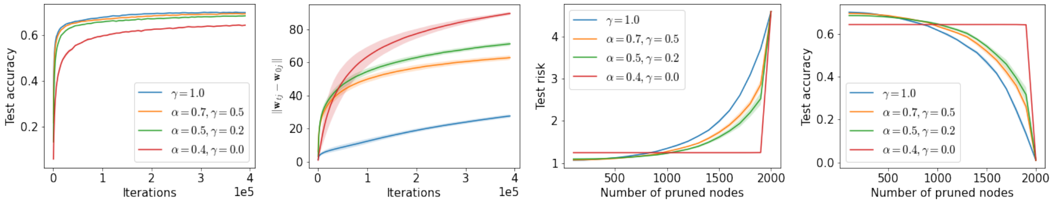

Figure 5: Results for `CIFAR-100`. From left to right, 1) test risk through training, 2) the differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 3) the test risks of the pruned models, and 4) test accuracies of the pruned models.

parameters as in the above paragraph, and we now train our FFNNs for 100 000 steps in each run. To further highlight the presence of feature learning in our model, we test the transferability of features learnt from our networks as follows. We first split the validation set into a held-out training set (50%) and a test set (50%), and extract features of the held-out training set using the FFNNs trained on the original training set. Features are taken to be the outputs of the hidden layers, so each data point in the validation set is represented with a $m = 2000$ dimensional vector. Then, we sort feature dimensions with respect to feature importance measured as $(\lambda_{m,j}\|\mathbf{w}_{t,j}\|^2)_{j\in[m]}$ and use the top-k of these to train an external model. The chosen external model is a FFNN with a single hidden layer having 64 neurons and ReLU activation, and it is trained for 5000 steps of GD with a learning rate of 1.0. Our theory suggests that smaller γ and α values likely lead to better transfer learning. A subset of our results is summarised in Figure 3; additional results are found in Appendix J. In line with the simulated data experiments, we observe a stronger presence of feature learning, in terms of weight-norm changes and NTG changes, for smaller values of γ and α. Also, we observe that models with smaller values of γ have lower risks when a small number of features are used for the transfer. The interpretation is that those models are able to learn a sufficient number of representative features using relatively fewer neurons.

Classification. We apply our model on two image classification tasks. The first is small-scale using the setting assumed in our theory, while the second is larger-scale using a more realistic setting. In addition to the transferability experiment described in the previous pragraph, we also test the prunability of the FFNNs. We gradually prune hidden nodes which have small feature importance and measure risks after pruning. Feature importance is measured as above. Our theory suggests that models with smaller γ and/or α values are likely more robust with respect to pruning, as long as $\gamma < 1$. Wolinski et al. (2020) had similar empirical findings on the benefits of asymmetrical scaling for network pruning when $\gamma = 0$.

MNIST. We take a subset of size 5000 from the MNIST dataset and train the same models used in the previous experiments. We also test pruning and transfer learning, where we use an additional subset of size 5000 to train an external FFNN having a single hidden layer with 128 nodes. To match our theory, instead of using cross-entropy loss, we use the MSE loss by treating one-hot class labels as continuous-valued targets. The outputs of the models are 10 dimensional, so we compute the NTG matrices using only the first dimension of the outputs. In general, we get similar results in line with our previous experiments. The pruning and transfer learning results are displayed in Figure 4. Other results can be found in Figure 10 in the Supplementary Material.

CIFAR. We consider a more challenging image classification task of CIFAR–10 and CIFAR–100. The datasets have 60 000 images of which 50 000 are used for training and the rest for testing. There are, respectively, ten and a hundred different classes. We illustrate the benefits of asymmetrical node scaling and show they hold for this more challenging problem. In many applications, one uses a large model pre-trained on a general task and then performs fine-tuning or transfer learning to adapt it to the task at hand. We implement this approach on a Resnet-18 model, pre-trained on ImageNet data. With this model, we transform each original image to a vector of dimension 512. We then train shallow FFNNs as described in **??**, with $m = 2000$, and output dimension 10 (resp. 100). This experiment differs from previous results as 1) we use stochastic GD with a mini-batch size of 64 instead of full batch GD; 2) we use cross-entropy loss instead of MSE; and 3) both layers are trained. All experiments are run five times, and the learning rate is 5.0. In Figure 5, we report the pruning results for the same four values of pairs (γ, α) as above, for CIFAR–100. Similar results are obtained for CIFAR–10 (see Appendix J). Similar conclusions as before hold here, even though the theory does not apply directly.

J.1 Regression

In Figures 6, 7, 8 and 9 we respectively provide the detailed results for the datasets `concrete`, `energy`, `airfoil` and `plant`.

J.2 Classification

We provide in Figure 10 detailed results for the MNIST dataset, and in Figure 11 results for the CIFAR–10 dataset. In Figure 12, we provide further details on the individual impact of the parameter $\gamma \in [0, 1]$. Recall that the smaller the value of γ, the more asymmetry is introduced, where $\gamma = 1$

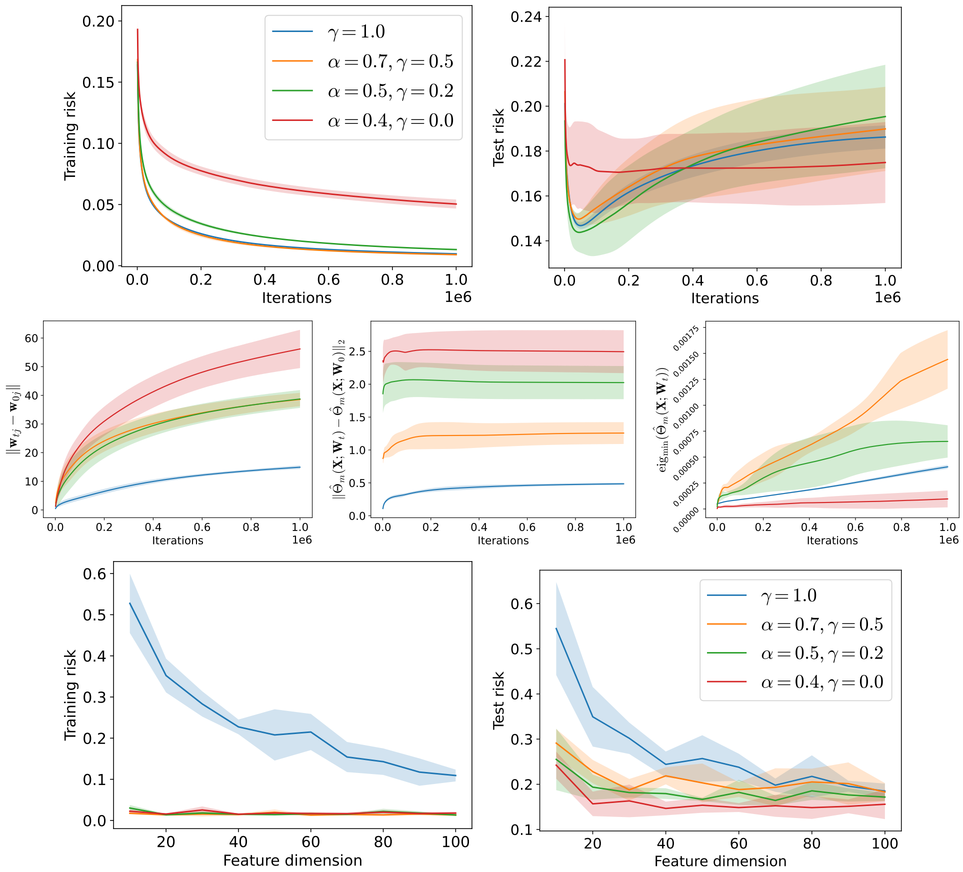

Figure 6: Results for the `concrete` dataset (swish). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

recovers the iid model. We can see from the experiments that pruning performance is improved as γ becomes smaller.

K Experimental results for a ReLU activation function

We provide here additional experimental results, as in Appendix J, but with a different activation function. Namely, we replace the swish activation function with the ReLU function. Although our theory does not cover the convergence of GD with the ReLU, the experimental results obtained in this section are quantitatively similar to those obtained with the swish function.

K.1 Regression

In Figures 13, 14, 15 and 16 we respectively provide detailed results for the datasets `concrete`, `energy`, `airfoil` and `plant`.

K.2 Classification

We provide in Figures 17, 18 and 19 detailed results for respectively the MNIST, CIFAR10 and CIFAR100 experiments.

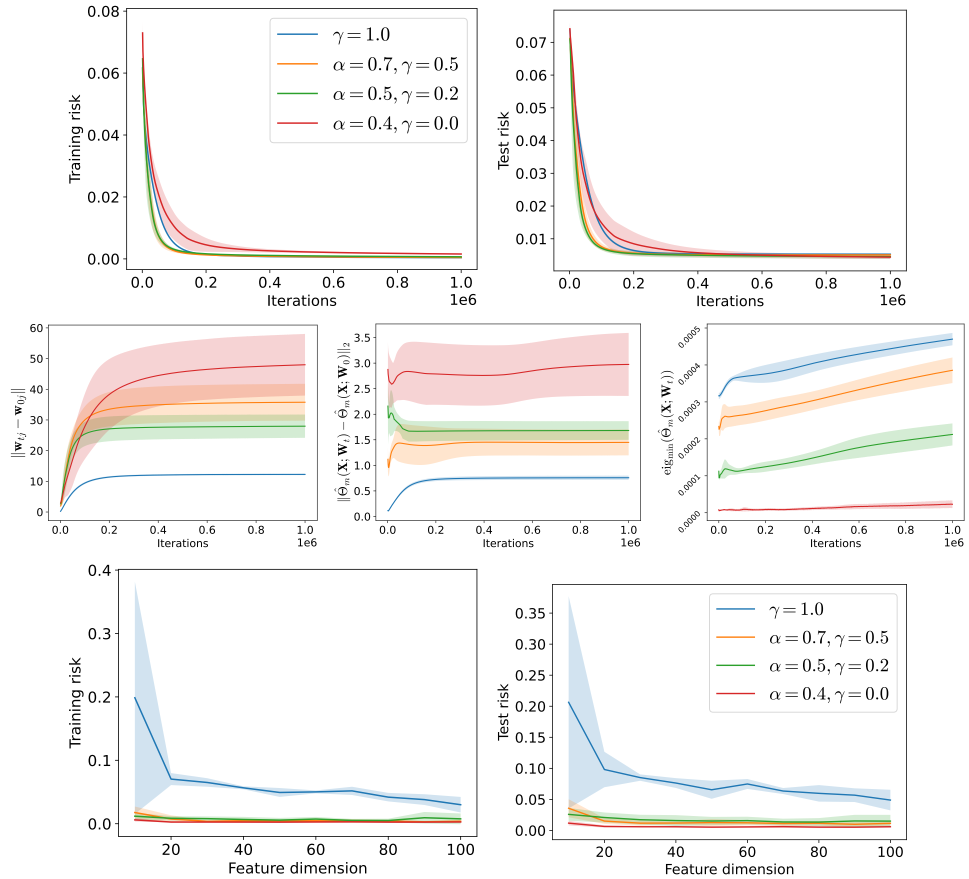

Figure 7: Results for the `energy` dataset (swish). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

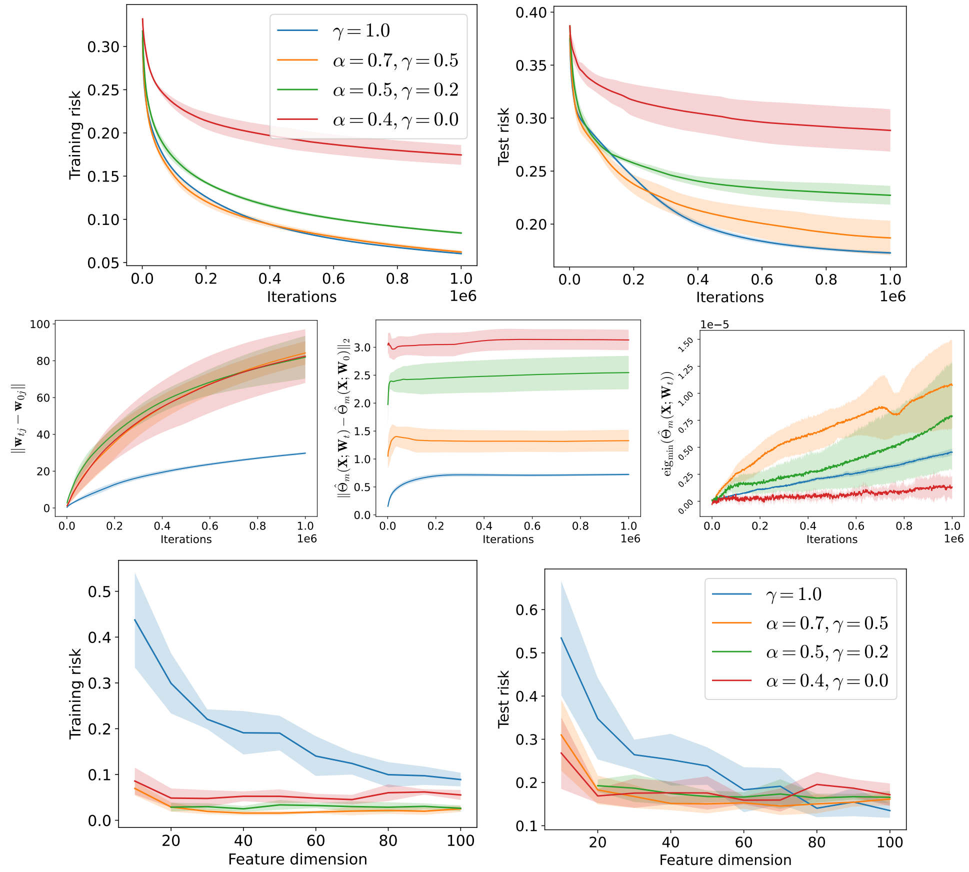

Figure 8: Results for the `airfoil` dataset (swish). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

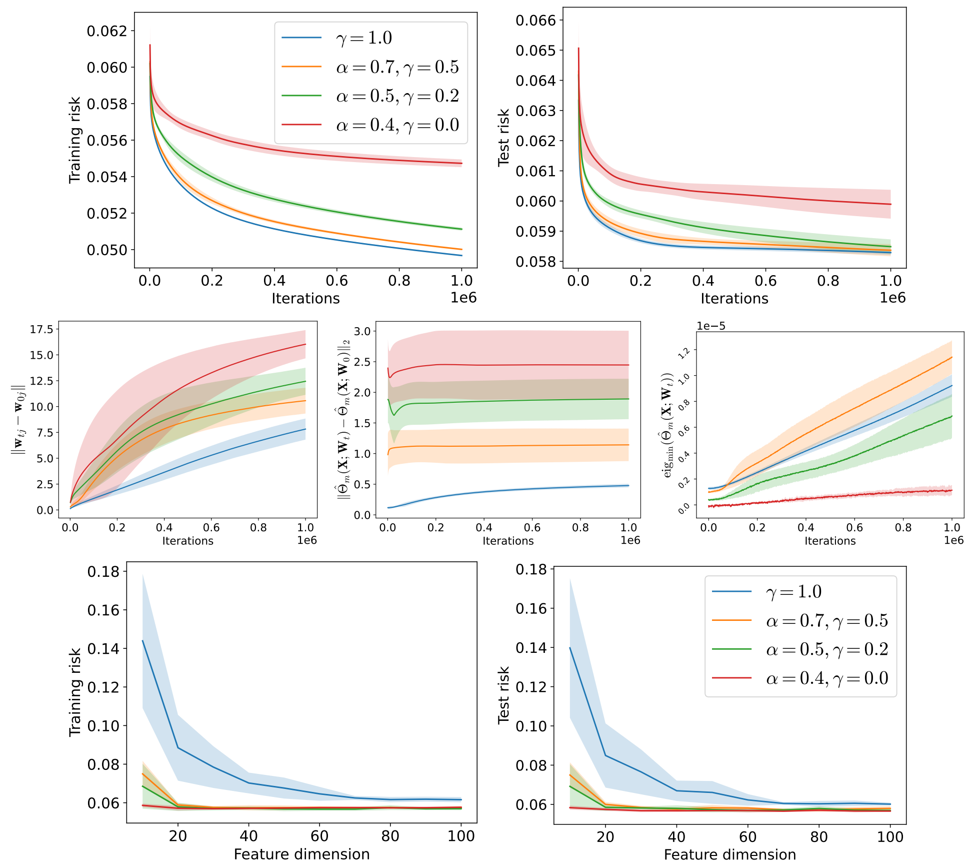

Figure 9: Results for the plant dataset (swish). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

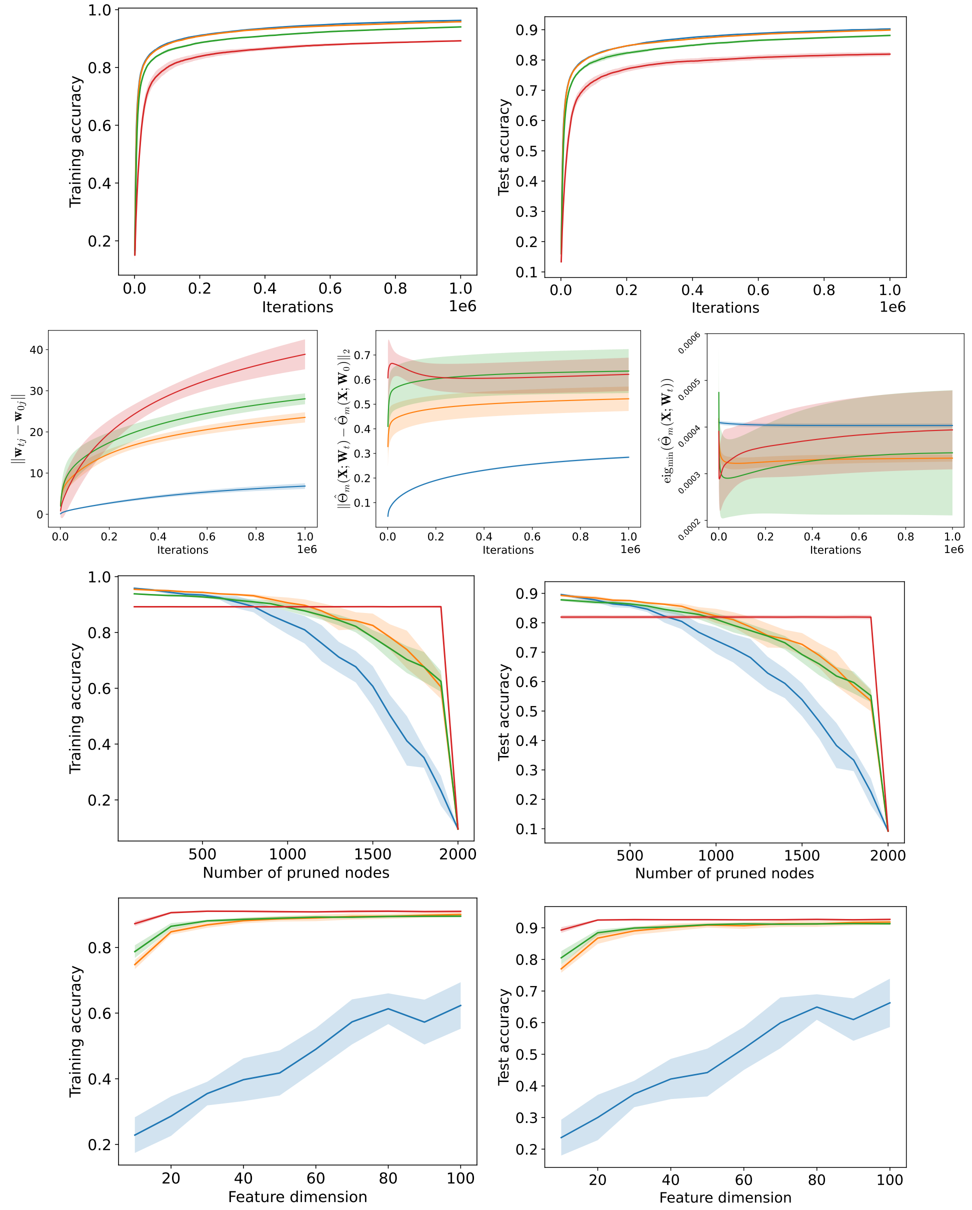

Figure 10: Results for the MNIST dataset (swish). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training accuracies for pruning, 7) test accuracies for pruning, 8) training accuracies for transfer learning, and 9) test accuracies for transfer learning.

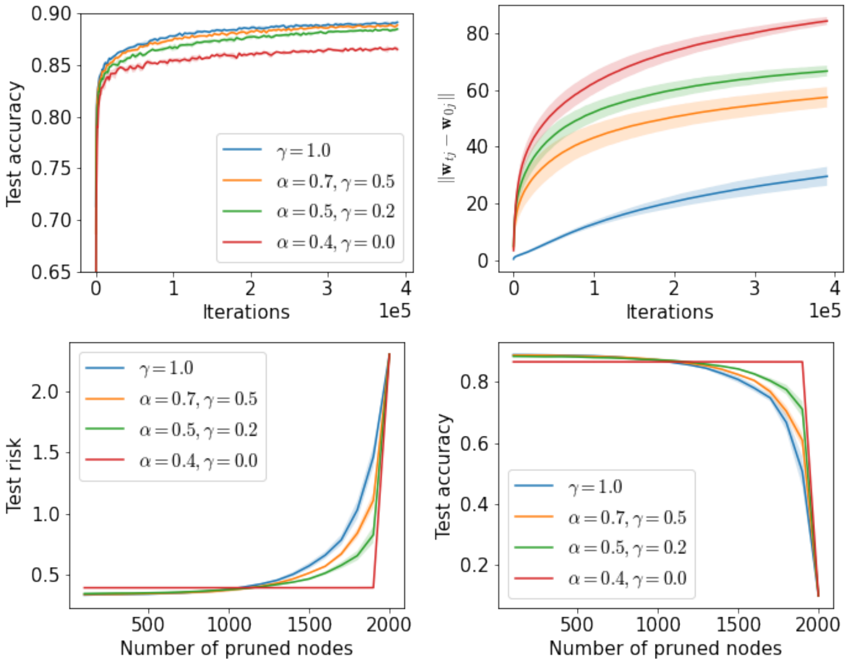

Figure 11: Results for the CIFAR-10 dataset (swish). From left to right and top to bottom, 1) test accuracies through training, 2) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 3) test risks of the pruned models, and 4) test accuracies of the pruned models.

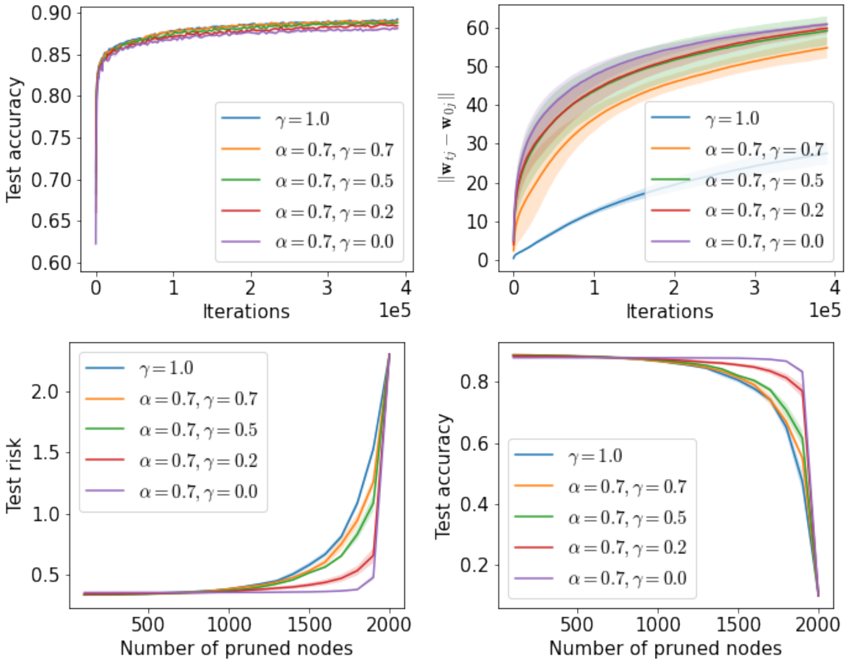

Figure 12: Results for the CIFAR-10 dataset (swish). Impact of the parameter γ. From left to right and top to bottom, 1) test accuracies through training, 2) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 3) test risks of the pruned models, and 4) test accuracies of the pruned models.

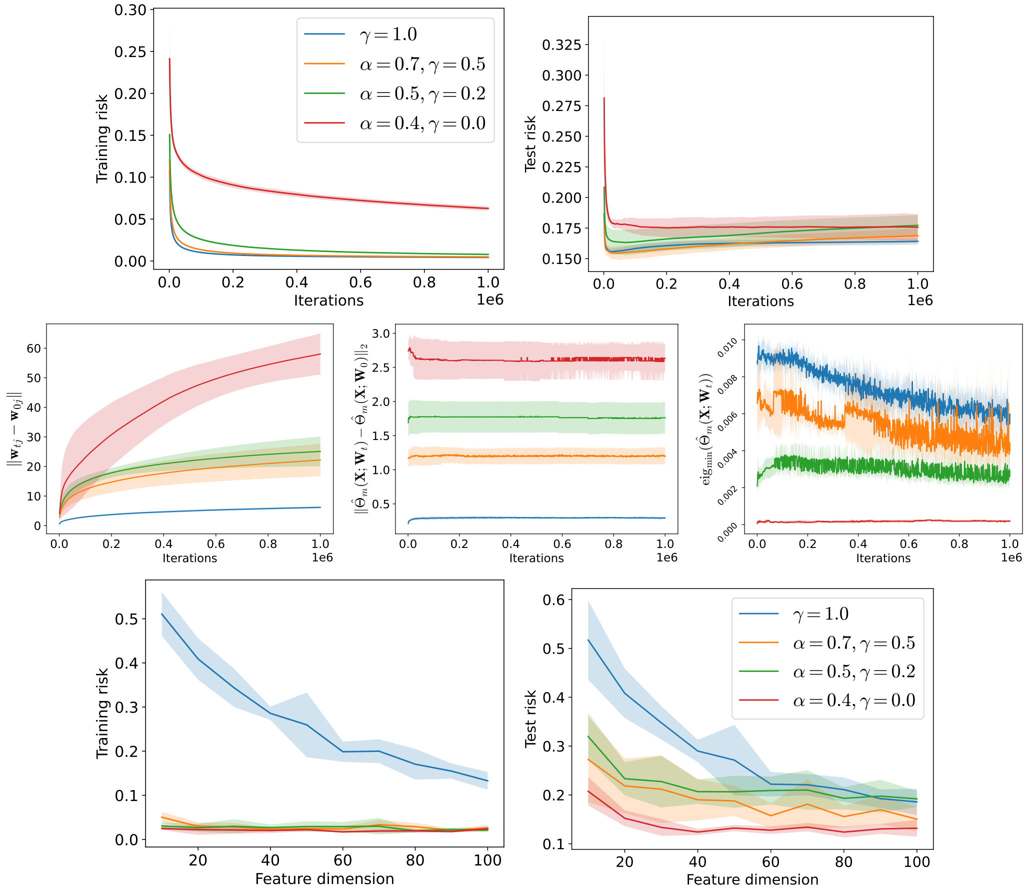

Figure 13: Results for the `concrete` dataset (ReLU). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

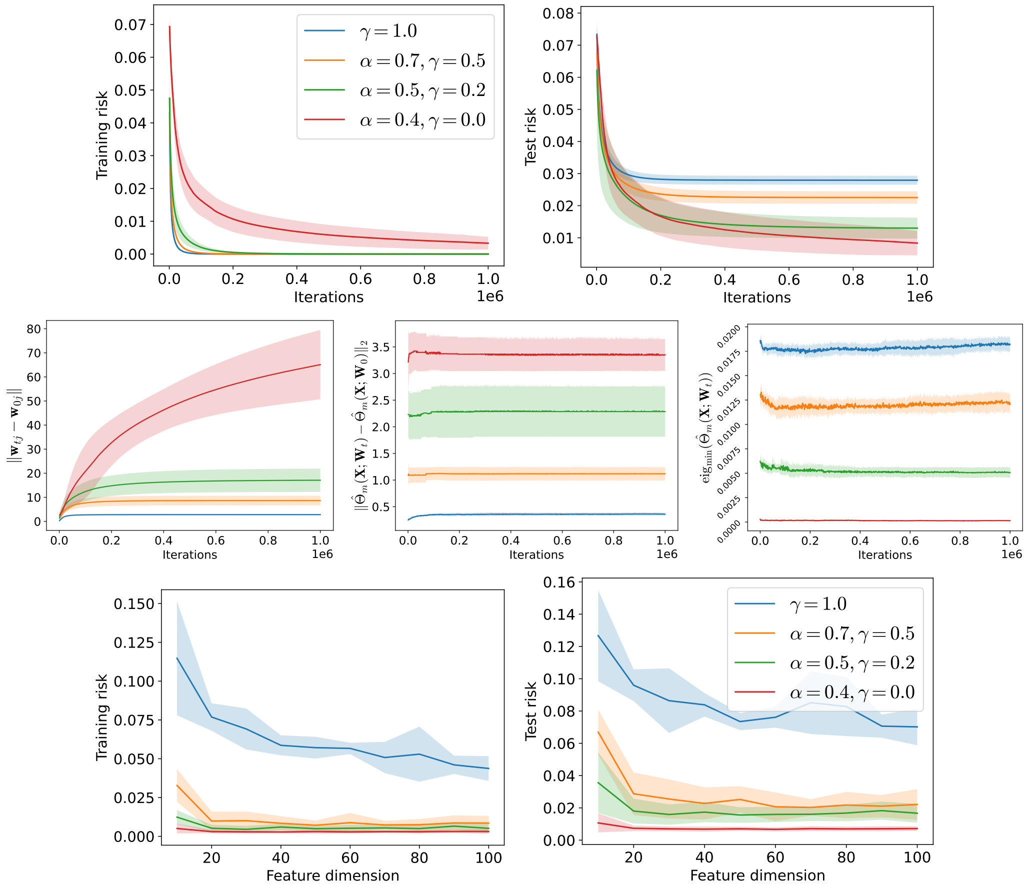

Figure 14: Results for the `energy` dataset (ReLU). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

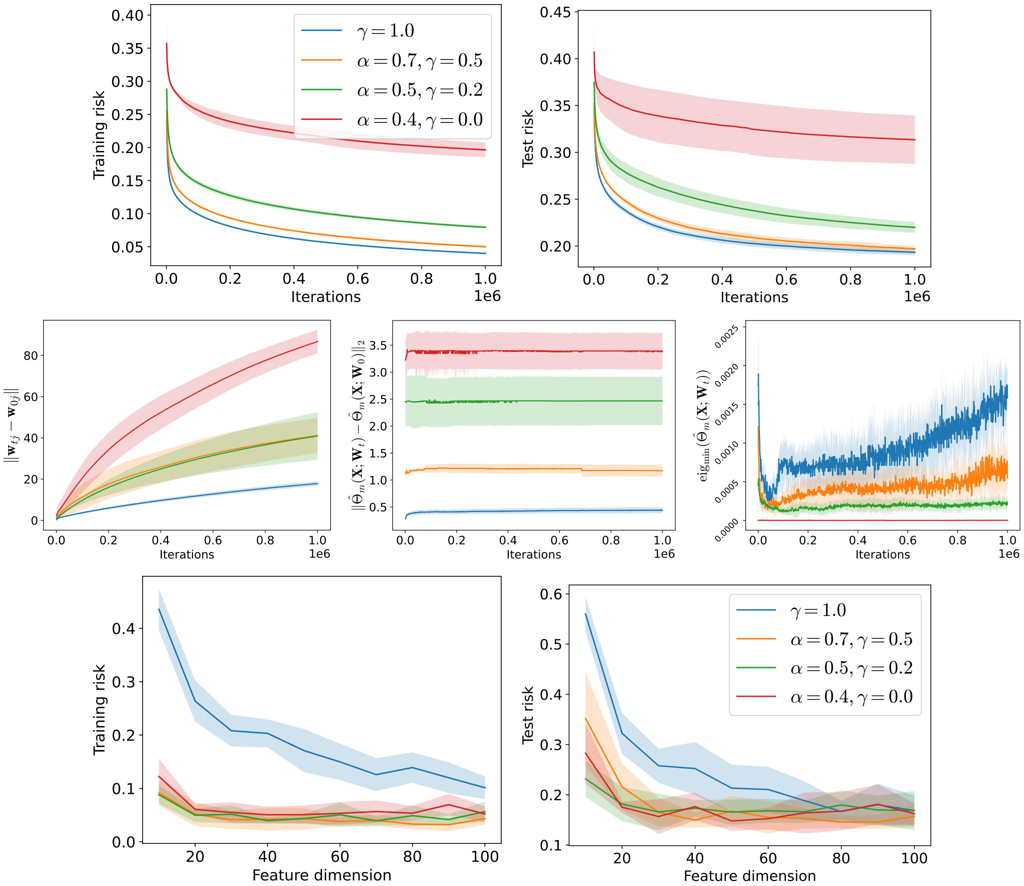

Figure 15: Results for the `airfoil` dataset (ReLU). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

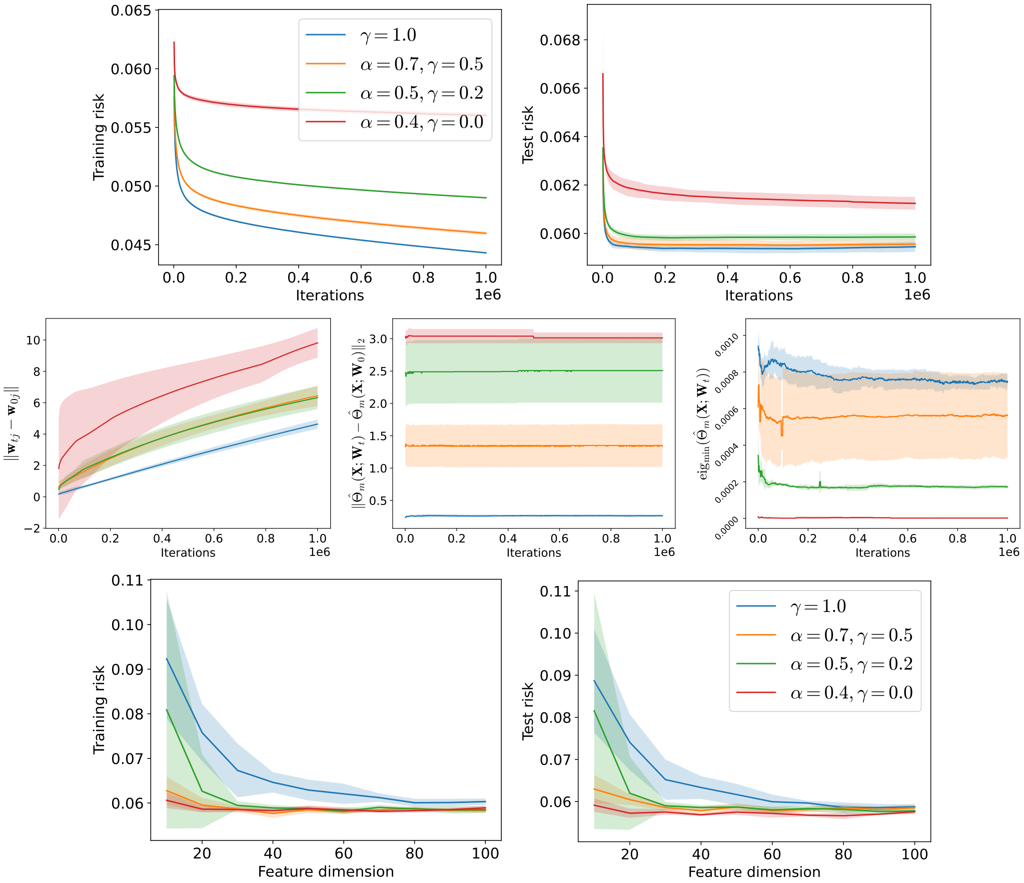

Figure 16: Results for the `plant` dataset (ReLU). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training risks for transfer learning, and 7) test risks for transfer learning.

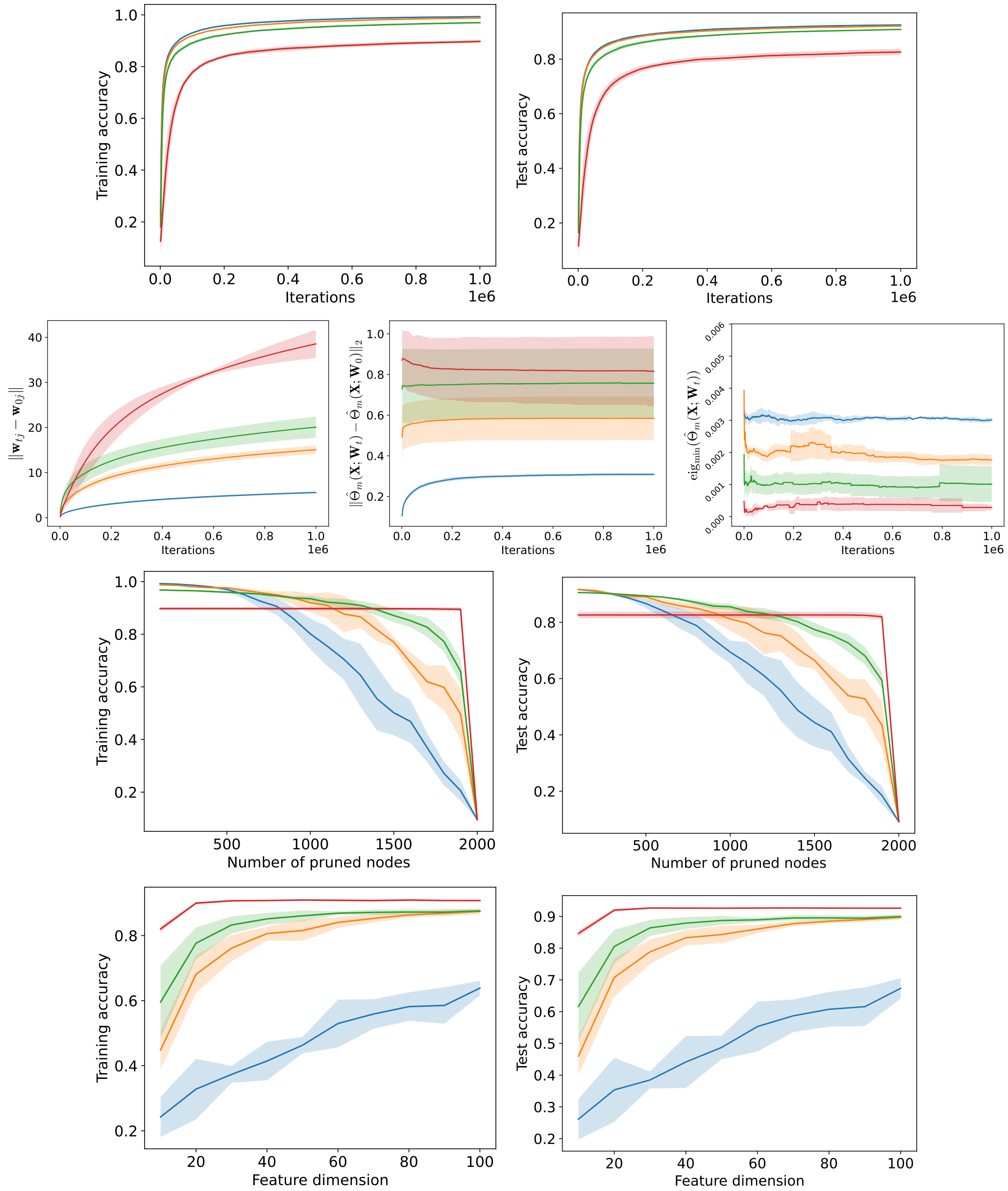

Figure 17: Results for the MNIST dataset (ReLU). From left to right and top to bottom, 1) training risks, 2) test risks, 3) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 4) difference in NTG matrices, 5) minimum NTG eigenvalues, 6) training accuracies for pruning, 7) test accuracies for pruning, 8) training accuracies for transfer learning, and 9) test accuracies for transfer learning.

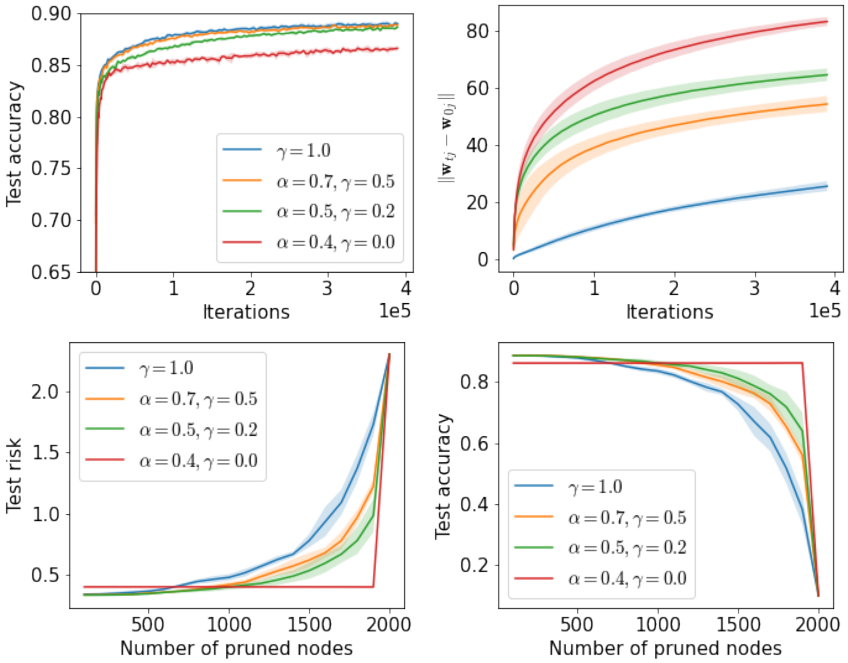

Figure 18: Results for the `CIFAR-10` dataset (ReLU). From left to right and top to bottom, 1) test accuracies through training, 2) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 3) test risks of the pruned models, and 4) test accuracies of the pruned models.

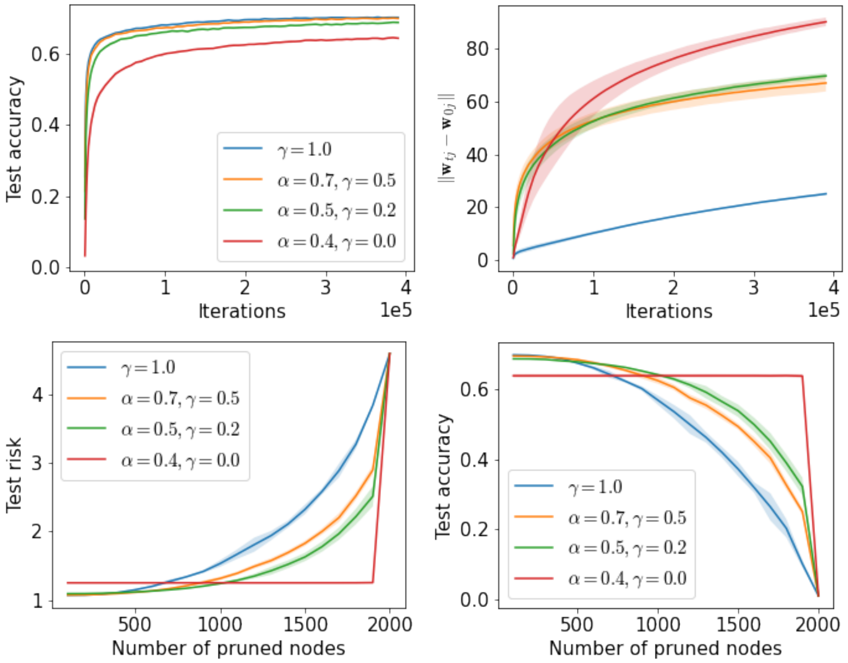

Figure 19: Results for the `CIFAR-100` dataset (ReLU). From left to right and top to bottom, 1) test accuracies through training, 2) differences in weight norms $\|\mathbf{w}_{tj} - \mathbf{w}_{0j}\|$ with j's being the neurons having the maximum difference at the end of the training, 3) test risks of the pruned models, and 4) test accuracies of the pruned models.