
Improved Convergence Guarantees for Shallow Neural Networks

Alexander Razborov*

Departments of Mathematics and Computer Science
University of Chicago
Chicago, IL 60637
razborov@uchicago.edu

Abstract

We continue a long line of research aimed at proving convergence of depth two neural networks, trained via gradient descent, to a global minimum. Like in many previous works, our model has the following features: regression with quadratic loss function, fully connected feedforward architecture, ReLU activations, Gaussian data instances and network initialization, adversarial labels. It is more general in the sense that we allow both layers to be trained simultaneously and at *different* rates. Our results improve on state-of-the-art [OS20] (training the first layer only) and [Ngu21, Section 3.2] (training both layers with Le Cun’s initialization). We also report several simple experiments with synthetic data. They strongly suggest that, at least in our model, the convergence phenomenon extends well beyond the “NTK regime”.

1 Introduction

Deep neural networks are remarkably successful at defying common wisdom inherited from the statistical learning theory and especially the theory of PAC learning. It would perhaps not entirely be an exaggeration to say that the wonder at their ability to generalize in a (highly) overparameterized regime is expressed in at least every other paper on the theory of deep learning so we confine ourselves to citing two expository papers [BHMM19, ZBH⁺21] *specifically* devoted to this phenomenon. While very interesting works [BFT17, NBS18, AGNZ18, LL18, ADH⁺19b, SY19] offer explanations of this “unbelievably good” generalization from different perspectives, it would be also fair to say that there is not clear consensus yet on the degree of their relevance to practical data sets.

Since there does not seem to be any reasons to expect a learning system to perform on test examples *better* than on training examples, an even more basic theoretical question is why an algorithm, which in this area almost always means gradient descent or its variants, achieves zero, or at least small, empirical risk. This question has met a certain degree of success, particularly in the situation described by various authors as “lazy training regime” (see e.g. [COB19]) or “neural tangent (kernel) regime” (see e.g. [MZ20]). Notwithstanding the name, this is the behavior described by the following (very much related) properties:

- only “a few” instance-neuron pairs change their activation during training;
- network parameters do not change “much” (say, in the Frobenius norm) during training.

*Also affiliated with Toyota Technological Institute at Chicago and Steklov Mathematical Institute, Moscow, Russia.

These properties in particular imply that the gradient of the loss function and hence the NTK² matrix H_t also do not change much from their value at initialization. Hence the training trajectory is almost linear and is driven, in the error space, by the value H_0 of the NTK matrix at the initialization. In the following discussion, we will adopt the terminology “NTK regime” to (loosely) describe this kind of behavior.

Previous work on the subject displays quite a diverse set of various assumptions on the activation and loss functions, data, initialization, as well as (also quite diverse) set of parameters featuring in the final results. They are not so easy to compare to each other. In order to get a more coherent picture, we find it convenient to start with a "theoretical benchmark" capturing a great deal of previous work (in the sense that their message is not distorted much when reduced to this scenario); we will separately mention several notable exceptions in Section 1.4.

1.1 Set-up

Feedforward fully connected neural networks of depth 2 (that is, with one hidden layer). ReLU activation function. Regression tasks with single output and the squared loss function. Data points (X^1, \dots, X^m) are sampled uniformly and independently from the unit sphere. Weights at initialization are Gaussian at the input level and Rademacher at the output level. Parameters:

- n – the problem dimension;
- S – the number of neurons at the middle level ("size" of the network);
- m – the number of training examples.

When $m \leq n$, the data is linearly separable. We will not exclude this case from statements and proofs but in Section 1 we will assume for simplicity that $m \gg n$. We will also tacitly assume $m \leq n^{O(1)}$; this assumption is indispensable in any work using NTK methods, including ours.

1.2 Previous work

The over-parameterized region is $m \ll nS$ and, clearly, no convergence or even representability is possible when $m \gg nS$. Several authors have loosely conjectured that perhaps this is actually tight and the (stochastic) gradient descent does converge in the whole region $m \ll nS$. This conjecture is not as far-fetched as it may seem since under this assumption alone one can derive, with overwhelming probability, quite strong conclusions at initialization like:

[SH18, Corollary 28] The activation matrix $A \in \{0, 1\}^{S \times m}$ at initialization possesses an intrinsic, purely combinatorial property ensuring the following. The corresponding NTK matrix is non-singular for *almost all* (in the Lebesgue sense) choices of data $X \in \mathbb{R}^{n \times m}$.

[XLS17, OS20, ADH⁺19a, NMM21, MZ20, BAM22] The NTK matrix w.r.t. the same data X that were used for computing the activation matrix A has the minimum eigenvalue³ $\Omega(S)$, i.e. well-separated from 0.

As for the dynamics of the gradient descent, known results are much less conclusive. Before briefly surveying them, let us again stress that many of these results pertain to more general situations such as deeper networks, less restrictive assumptions on data, more general loss or activation functions etc. In our simplified treatment we have deliberately reduced them to the “common denominator” outlined in Section 1.1.

Let us first consider the case when only the first layer is trained. [DZPS19] proved that convergence takes place whenever⁴ $m \leq \tilde{o}(S^{1/6})$ (plug $m \mapsto S$; $n \mapsto m$; $\lambda_0, \delta \mapsto O(1)$ in their Theorem 4.1); apparently, this can be approved to $m \leq \tilde{o}(n^{1/6} S^{1/6})$ by examining their proof more carefully. The paper [WDW19] improved the bound on the learning rate from [DZPS19] using more sophisticated

²In many works on the subject, this term is reserved for the value at initialization, that is H_0 , or even its limit version. We, however, prefer to use it in broader, dynamical sense as e.g. in [FDK⁺20].

³We do not normalize by S .

⁴We use the notation $\tilde{O}, \tilde{\Omega}, \tilde{\Theta}, \tilde{o}, \tilde{\omega}$ to hide factors that are poly-logarithmic in n, S , and their tildeless versions to hide constant factors. See Remark 1 below for clarifying remarks and examples.

(adaptive) gradient methods. The improvement $m \leq \tilde{o}(n^{1/2}S^{1/4})$ can be relatively easily extracted from [SY19] (plug $m \mapsto S$; $n \mapsto m$; $\lambda \mapsto O(1)$; $\alpha \mapsto \tilde{O}(m/n)$; $\theta \mapsto \tilde{O}(m^{1/2}/n^{1/2})$ in their Theorem 1.6). Finally, the paper [OS20] achieved the state-of-the-art result: convergence takes place whenever

$$m \leq \tilde{o}(n^{3/4}S^{1/4}). \quad (1)$$

Our description may look somewhat uneventful but we would like to stress that every new achievement along these lines required introducing new ingenious and technically advanced ideas and methods.

Less work seems to have been done on the case when both layers are trained simultaneously, and now the answer may depend on the normalization of weights at the initialization even if we disregard the learning rate. The original paper [DZPS19] gives the same estimate $m \leq \tilde{o}(S^{1/6})$ under the “standard” normalization, when the initial weights are supposed to be of the same order (say, $\tilde{O}(1)$) at both layers. The paper [Ngu21] is mostly devoted to deep networks but in Section 3.2 it also translates the main result to the shallow case under the so-called LeCun’s initialization. After performing slightly more careful calculations, in our notational system this amount to the condition

$$m \leq \min\left(\tilde{o}\left(\frac{S}{n}\right), \tilde{o}\left(S^{1/2}\right)\right). \quad (2)$$

1.3 Our contributions

We work in the natural multi-rate setting in which the first layer is trained at a rate $\eta_w \geq 0$ and the second layer is trained at possibly another rate $\eta_z \geq 0$. This appears to be very instructive and allows us to easily translate results from one normalization of initial weights to another. In particular, we choose to work in the “standard” normalization, when all weights are of order $\tilde{O}(1)$.

As a by-side remark, recent paper [VL22] expressed surprise, that we completely share, that this simple idea of multi-rate learning seems to have been largely overlooked in the theory of *deep* learning and sought to change that by providing, among other things, significant experimental evidence. One difference between our paper and [VL22] is that the latter restricts the *set of steps* in which a “slow” set of parameters is changed while we treat all steps uniformly. But this difference appears to be cosmetic.

Data and initialization. We pick the data X^1, \dots, X^m uniformly (under the Haar measure) and independently from \mathbb{S}^{n-1} . All entries in W_0 are chosen from $\mathcal{N}(0, 1)$, independently of each other. We prefer to allow for a slightly better flexibility in the choice of z_0 , primarily since the literature alternates between Gaussian and Rademacher initializations. Namely, we require that z_0 is independent of W_0 , and that the entries in z_0 are S i.i.d. copies of a one-dimensional distribution ζ such that:

$$\begin{aligned} \mathbf{E}[\zeta] &\stackrel{\text{def}}{=} 0; \\ \|\zeta\|_{\psi_2} &\leq \tilde{O}(1); \end{aligned} \quad (3)$$

$$\text{Var}(\zeta) \geq \tilde{\Omega}(1). \quad (4)$$

Our main contribution (all formal definitions will be given in Section 2) is as follows.

Theorem 1.1 *Let $m \geq 1$ and $n, S \geq \tilde{\omega}(1)$ be parameters such that $m \leq n^{O(1)}$ and $m \leq \tilde{o}(S)$. Let $\eta_w, \eta_z \geq 0$, $\eta_w + \eta_z > 0$ be such that*

$$\eta_w \leq \min\left(\tilde{o}\left(\frac{1}{S^2}\right), \tilde{o}\left(\frac{n}{mS^2}\right)\right), \quad \eta_z \leq \tilde{o}\left(\frac{1}{mS^2}\right) \quad (5)$$

and that at least one of the following three conditions holds:

$$m \leq \tilde{o}(nS^{1/4}), \quad \frac{\eta_z}{\eta_w} \leq \tilde{o}\left(\frac{S}{m}\right), \quad (6)$$

$$m \leq \tilde{o}(nS^{1/4}), \quad \frac{\eta_z}{\eta_w} \geq \tilde{\omega}\left(\frac{m^2}{nS}\right), \quad (7)$$

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(1 + \frac{m}{n}\right). \quad (8)$$

Then with probability $\geq 1 - (nS)^{-\omega(1)}$ the following holds. For almost all (in the Lebesgue sense) $y \in \mathbb{R}^m$ the gradient descent with multi-rates η_w, η_z avoids singular points and if additionally

$$\|y\| \leq \tilde{O}(m^{1/2} S^{1/2})$$

then it converges to a global minimum.

Remark 1 This paper uses the asymptotic notation $\tilde{O}, \tilde{\Omega}, \tilde{o}, \tilde{\omega}$ (and sometimes their tildeless versions) perhaps more systematically and heavily than most papers on the subject. Thus it would be appropriate to pause for a moment and reflect on what exactly it means.

We argue about infinite sequences $\left\{ \left(n^{(k)}, m^{(k)}, S^{(k)}, \eta_w^{(k)}, \eta_z^{(k)} \right) \right\}_{k \geq 1}$ of parameter values. Then (say) $m \leq \tilde{o}(nS^{1/4})$ means that for any fixed $C > 0$,

$$\lim_{k \rightarrow \infty} \frac{m^{(k)} \log(n^{(k)} S^{(k)})^C}{n^{(k)} (S^{(k)})^{1/4}} = 0 \quad (9)$$

while the error bound $(nS)^{-\omega(1)}$ simply states that it is asymptotically smaller than any fixed polynomial in n, S . The condition $n, S \geq \tilde{\omega}(1)$ says that S and n are (very mildly) balanced: $S \geq (\log n)^{\omega(1)}$ and $n \geq (\log S)^{\omega(1)}$; it appeared in most of the previous work.

Remark 2 The three cases (6), (8), and (7) are in fact three separate statements in which we set our hopes to converge fast on the first, second or both layers, respectively. We have combined them into one “mega-theorem” since the proofs share many steps.

Let us now compare our result with the previous work cited above. The case $\eta_z = 0$ (that is, training the first layer only) automatically implies the second condition in (6), and the only remaining assumptions are

$$m \leq \min \left(\tilde{o}(S), \tilde{o}(nS^{1/4}) \right) \quad (10)$$

which is an improvement on (1).

The condition $\frac{\eta_z}{\eta_w} \leq \tilde{o}\left(\frac{S}{m}\right)$ in (6) is automatically satisfied when $\eta_z = \eta_w$ (as $m \leq \tilde{o}(S)$), i.e. for the “ordinary” gradient descent. Thus our result improves on [DZPS19]. As an easy calculation shows, re-normalizing from LeCun’s initialization to the standard one gives the ratio $\frac{\eta_z}{\eta_w} = \frac{S}{m}$. Thus the condition (8) holds whenever $m \leq \min \left(\tilde{o}(S), \tilde{o}(n^{1/2} S^{1/2}) \right)$. This gives an improvement on (2).

Finally, note that when $m \leq \tilde{o}(n^{1/3} S^{2/3})$, the ranges for the relative learning rate $\frac{\eta_z}{\eta_w}$ in (6) and (7) overlap. Thus, under the assumptions

$$m \leq \min \left(\tilde{o}(S), \tilde{o}(n^{1/3} S^{2/3}), \tilde{o}(nS^{1/4}) \right),$$

convergence of the gradient descent is guaranteed *regardless* of the relative learning rate; we only have to make sure the rates are individually small enough. In particular, this holds regardless of the choice of normalization at the initialization.

Our improvements are admittedly not dramatic but our main motivation was to develop new techniques that might turn out useful for pushing the bound further or in other similar situations. This is what we consider our main technical contributions (some references below point out to the Appendix included as a supplementary material).

- Most previous work used, explicitly or implicitly, bounds on the spectral norm $\|X\|$, where X is the data matrix. We will heavily exploit that when X is random, analogous bounds also hold for all its sub-matrices X^J , where J is a subset of data instances (see (19)). We will also employ a dual lower bound on $\sigma_{\min}((X^J)^\top)$ for all sufficiently large J (see (20) and compare with [HY20, Assumption 2.2]).

⁵We do not include m as we are assuming $m \leq n^{O(1)}$ throughout the paper.

- We will also need to extend the lower bound $\lambda_{\min}(H_0) \geq \Omega(S)$ to the uniform lower bound $\lambda_{\min}(H_0|_{\Gamma}) \geq \Omega(S)$, where $\Gamma \subseteq [S]$ is a sufficiently large but otherwise arbitrary set of neurons and $H_0|_{\Gamma}$ is obtained from H_0 by disregarding neurons not in Γ . It turned out surprisingly difficult and required a version of the Lipschitz concentration inequality (see the second part of Appendix D).
- For the first two parts (6), (7), we also need to do finer analysis of the changes the activation matrix A incurs during training than in previous work. Informally speaking, there are two different ways the network may try to escape the NTK regime: by accumulating a large overall number of changes or a relatively large number concentrated on a small set of inputs J . Attempts of the first kind will be prevented by the assumption $m \leq \tilde{o}(nS^{1/4})$ while the second kind will be taken care of by $m \leq \tilde{o}(S)$. See Appendix E for details.
- Since in case of ReLU activations the gradient is discontinuous, we have to take care of the unpleasant situation when A and hence the Jacobian J_F of the feature map significantly change during one step. The previous work (see e.g. [OS20, Appendix G]) did it by comparing both Jacobians to the Jacobian at initialization but it does not work any longer with our relaxed assumptions. Instead, we adopt to our purposes the beautiful invariant from [ACH18, DHL18] concerning the behavior of weights associated to an individual neuron. See Appendix G for details.

In Section 4 and Appendix H we will also report several simple experiments with synthetic data. One strong conclusion we draw from these experiments is that there should be very significant room for further theoretical improvements, possibly (almost) all the way up to the representability barrier $m \sim nS$.

1.4 Related research

Our attempt at the uniformization has inevitably left behind some very interesting and at least somewhat related work. Very briefly:

Deeper networks. Several papers (notably [ALS19, ZG19]) are specifically devoted to the NTK regime in the context of deeper networks. When scaled down to shallow networks, however, the results do not seem to be as good as those obtained with methods specifically tailored to that situation.

Classification problems. Quite a bit of similar and generally stronger results have been obtained for binary classification problems, see e.g. [BGMS18, ZCZG20, JT20b, CCZG21, FCB22, LZ22] (the theoretical literature on *multi*-classification problems appears somewhat scarcer; see, however, [LL20, JT20a, VSS22]). While some methods and approaches are shared with regression problems, many techniques seem to be specific to classification.

Smooth activation. This case tends to be easier than the case of ReLU, and many technical difficulties disappear. For strong results in that direction not covered above see e.g. [HY20, SRP⁺21, BAM22].

2 Definitions and preliminaries

2.1 Notation

We let $[n] \stackrel{\text{def}}{=} \{1, 2, \dots, n\}$ and let $\binom{V}{n}$ be the collection of all n -element subsets of V . For a real number t , $\lfloor t \rfloor$ is its integer part and $\{t\} \stackrel{\text{def}}{=} t - \lfloor t \rfloor$ is its fractional part. All vectors and matrices are real, and all vectors are column vectors unless otherwise noted. Matrices are generally denoted by upper case Latin letters like A, B, H, W, X and vectors are typically denoted by lower case letters. We let $\mathbb{R}^{n \times m}$ be the space of $n \times m$ matrices. For $M \in \mathbb{R}^{n \times m}$ and $j \in [m]$, we let M^j denote its j th column vector, and for $i \in [n]$, M_i is its i th row vector. More generally, for $J \subseteq [m]$, M^J is the corresponding $n \times |J|$ matrix, and similarly for M_I . For $z \in \mathbb{R}^S$, $\text{diag}(z) \in \mathbb{R}^{S \times S}$ is the corresponding diagonal matrix.

The symbol \circ stands for the Hadamard (entrywise) product of vectors/matrices of the same size. The symbol $*$ will denote the *column-wise* Khatri-Rao product: for $A \in \mathbb{R}^{S \times m}$ and $X \in \mathbb{R}^{n \times m}$,

$(A * X) \in \mathbb{R}^{S \times n \times m}$ and $(A * X)_{(\nu i), j} \stackrel{\text{def}}{=} A_{\nu j} X_{ij}$. We will denote the standard Euclidean norm by $\|x\|$. The unit sphere $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ is given by $\mathbb{S}^{n-1} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid \|x\| = 1\}$; more generally, $r \cdot \mathbb{S}^{n-1} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid \|x\| = r\}$. The spectral norm $\|M\|$ of $M \in \mathbb{R}^{n \times m}$ is $\|M\| \stackrel{\text{def}}{=} \max_{\xi \in \mathbb{S}^{m-1}} \|M\xi\|$, its minimum singular value is $\sigma_{\min}(M) \stackrel{\text{def}}{=} \min_{\xi \in \mathbb{S}^{m-1}} \|M\xi\|$ and the Frobenius norm is $\|M\|_F \stackrel{\text{def}}{=} \left(\sum_{i \in [n]} \sum_{j \in [m]} M_{ij}^2\right)^{1/2}$. We let $\|x\|_\infty$, $\|M\|_\infty$ denote the maximum absolute value of an entry.

When H is a PSD (positive semi-definite) matrix, we will usually write $\lambda_{\min}(H)$ for $\sigma_{\min}(H)$; thus, for any real matrix M , $\sigma_{\min}(M) = \lambda_{\min}(M^\top M)^{1/2}$. For two symmetric matrices M, N of the same size, $M \succeq N$ means that $M - N$ is PSD.

We let $\mathbf{E}[*]$ be the expectation of a real random variable, $\text{Var}[*]$ be the variance of a (one-dimensional) random variable and $\mathbf{P}[E]$ be the probability of an event E . $\mathbf{1}(E)$ is the characteristic function of E . When we randomize over a part of the sample space, this will be indicated by the corresponding subscript, like $\mathbf{E}_{W_0}[*]$ or $\mathbf{P}_z[*]$. Let $\|\zeta\|_{\psi_2}$, $\|\zeta\|_{\psi_1}$ be the sub-gaussian and the sub-exponential norms of a one-dimensional ζ defined as

$$\|\zeta\|_{\psi_p} \stackrel{\text{def}}{=} \inf \left\{ s > 0 \mid \mathbf{E} \left[e^{(\zeta/s)^p} \right] \leq 2 \right\}.$$

2.2 The model

Let $n, m, S \geq 1$. We are given m data instances $X^1, \dots, X^m \in \mathbb{R}^n$ arranged as a matrix⁶ $X \in \mathbb{R}^{n \times m}$ and a *label vector* $y \in \mathbb{R}^m$. A (shallow) neural network is given by a pair $\theta = (W, z)$, where $W \in \mathbb{R}^{S \times n}$ and $z \in \mathbb{R}^S$. It computes as

$$\begin{aligned} F(W) &\stackrel{\text{def}}{=} \sigma(WX) \in \mathbb{R}^{S \times m}; \\ f(\theta) &\stackrel{\text{def}}{=} F(W)^\top z \in \mathbb{R}^m, \end{aligned}$$

where, as usual, the ReLU activation function $\sigma(x) \stackrel{\text{def}}{=} \max(x, 0)$ is applied to WX entrywise. We define the *error vector* as

$$e(\theta) \stackrel{\text{def}}{=} f(\theta) - y \in \mathbb{R}^m$$

and the (quadratic) *loss function* as

$$\ell(\theta) \stackrel{\text{def}}{=} \frac{1}{2} \|e(\theta)\|^2.$$

Let us call $\theta = (W, z)$ *regular* if WX does not contain zero entries and *singular* otherwise. All functions just defined are polynomial in a neighborhood of any regular θ , hence the gradient $\nabla \ell(\theta)$ is well-defined for regular θ . We split it as $\nabla \ell(\theta) = (\nabla^w \ell(\theta), \nabla^z \ell(\theta))$ in the obvious way: $\nabla^z \ell(\theta) \in \mathbb{R}^S$, and $\nabla^w \ell(\theta)$ will be treated, depending on the context, either as a long vector in \mathbb{R}^{nS} or as a matrix in $\mathbb{R}^{S \times n}$.

Assume now that we are also given initial values $\theta_0 = (W_0, z_0)$ and the *learning rates* $\eta_w, \eta_z \geq 0$. We define the *gradient descent* $\theta_\tau = (W_\tau, z_\tau)$ ($\tau \geq 1$ an integer) as

$$\begin{aligned} W_\tau &\stackrel{\text{def}}{=} W_{\tau-1} - \eta_w \nabla^w \ell(\theta_{\tau-1}); \\ z_\tau &\stackrel{\text{def}}{=} z_{\tau-1} - \eta_z \nabla^z \ell(\theta_{\tau-1}). \end{aligned} \tag{11}$$

There is a small caveat here since this definition makes sense only when all points θ_τ are regular. Fortunately, it is easy to take care of with a simple measure-theoretical argument that we will present in Appendix B.

Before giving a brief proof sketch of Theorem 1.1 (the complete proof is deferred to Appendix), let us make a few more remarks.

⁶In many works on the subject, data points are arranged as *rows* of the data matrix. We have found that, while also imperfect, the opposite convention leads to a cleaner and more natural notation.

Remark 3 The choice of labels is completely adversarial (they are only required to be of “right” order), and the adversary is also allowed to see the initialization, not only the data. This aligns well with the prominent experiments in [ZBH⁺21] strongly suggesting that the choice of labels should *not* be a defining factor in convergence.

Remark 4 As in all previous work, the error probability can be decreased to exponential by examining the proofs in Appendices C, D.

Remark 5 Since $m \leq S$, the bound in (8) implies the second bound in (7). Thus, we always have either $m \leq \tilde{o}(nS^{1/4})$ or $\frac{\eta_z}{\eta_w} \geq \tilde{\omega}\left(\frac{m^2}{nS}\right)$ and, in fact, the refinement (8) of this dichotomy will not be needed until Appendix F.

3 Some ideas of the proof

We frame the proof according to the paradigm known in combinatorics as *quasirandomness* (see [CGW89] for graphs and [CR23] for arbitrary combinatorial objects). The idea is to split a logically elaborated argument involving random objects into two totally independent parts. At the first stage we accumulate a list of so-called “quasirandom properties”, that is *completely deterministic* facts that hold for our random objects with overwhelming probability that we choose to be $\geq 1 - (nS)^{-\omega(1)}$. At the second stage, the argument proceeds completely deterministically, on the base of these properties only. This allows us to avoid unwanted serious complications, or even sheer mistakes, caused by the fact that quantifiers and randomization do not get along well in one argument.

This approach is of course not new to the area; it is known under various names like “assumptions”, “meta-theorems” (see e.g. [OS20]) etc. But we do it perhaps a bit more systematically and on a larger scale.

Almost all of our quasi-random properties are quite straightforward; their analogues have been heavily used in the previous work (usually as assumptions). The following two seem to be at least somewhat novel:

- We need to upper bound the spectral norm not only of the data matrix X itself but also of its *sub-matrices* of the form X^J , $J \subseteq [m]$, and this bound,

$$\|X^J\| \leq \tilde{O}\left(1 + \frac{|J|^{1/2}}{n^{1/2}}\right) \quad (12)$$

must hold uniformly over J . Dually, for *any* set J of examples that is just slightly larger than n , we postulate $\sigma_{\min}((X^J)^\top) \geq n/m$.

- As in the previous work, we need the bounds $\lambda_{\min}(H_0), \lambda_{\min}(G_0) \geq \Omega(S)$ for NTK matrices H_0, G_0 at the initialization. But for the matrix H_0 (corresponding to the first layer) we also need this property to hold *even after we “restrict” it* (see (29) for details) *to any sufficiently large set of neurons* Γ . The statement in italic has, somewhat surprisingly, turned out to be the most involved part of the proof; it occupies most of Appendix D. The main difficulty lies in the fact that we need our bound to hold uniformly over Γ that seems to preclude simple tools like the union bound. Instead, we employ the Lipschitz concentration inequality [Ver18, Theorem 5.1.4] or rather some of the ideas from its proof.

Let us now turn to the second, deterministic part of the proof. On the abstract level, it follows the same pattern as most other proofs in the NTK regime. Namely, we prove by joint induction on time t three intertwined properties:

1. The distances $\|W_t - W_0\|_F, \|z_t - z_0\|$ travelled by the trajectory stay small during the whole training (“lazy training”);
2. The activation matrix A_t (for precise definitions see Appendix A.2) does not change much w.r.t. A_0 ;

3. In the cases (6), (7) we have, again throughout the training, $\lambda_{\min}(H_t) \geq \Omega(S)$ while in cases (7), (8) we have $\lambda_{\min}(G_t) \geq \Omega(S)$.

In particular, 3) will immediately imply that the error vector e_t decays exponentially and will complete the proof.

As above, in this brief sketch we try to highlight what we view as relatively novel to our work. Assuming that properties 1), 2), 3) hold for all $t < T$, we have to prove them for $t := T$ (the base case $T = 0$ for 3) was taken care of at the first stage). We begin with 1), and we estimate⁷ $\|W_T - W_0\|_F$ using Cauchy-Schwarz as follows:

$$\|W_T - W_0\|_F \leq \eta_w \left(\sum_{t=0}^{T-1} \frac{\|\nabla^w \ell(\theta_t)\|^2}{\|e_t\|} \right)^{1/2} \cdot \left(\sum_{t=0}^{T-1} \|e_t\| \right)^{1/2}.$$

The first sum can be easily bounded by $O(\|e_0\|)$, which in turn is controlled by our quasi-random assumptions, as long as we know that the gradient $\nabla^w \ell(\theta_t)$ does not oscillate too widely during one step. (A priori, this may happen if too many pairs neuron-instance decide to fire at once.) The proof turns out to be surprisingly difficult and requires quite new ideas, namely a discredited version of the beautiful invariant from [ACH18, DHL18]. It is presented in Appendices F and G.

The estimate for the second sum $\sum_{t=0}^{T-1} \|e_t\|$ is pretty straightforward; it follows from the exponential decay of $\{e_t\}$.

The proof of 2) (for $t := T$) more or less follows the pattern from the previous work. Like before, we actually need the bound

$$\|(A_t - A_0) * X\| \leq \tilde{o}(S^{1/2}) \quad (13)$$

that, with a small caveat explained below, will also imply 3). For proving (13)), we also employ Shur's inequality. The novelty is that we have to split the set of examples $[m]$ as $[m] = J \cup co - J$ and then bound separately $\|((A_t - A_0) * X)^J\|$ and $\|((A_t - A_0) * X)^{co-J}\|$. Here J consists of n examples that undergo the heaviest activation changes, and this is where more general bound (12) is used (and is crucial to our improvement).

Finally, 3) almost follows from (13) but the caveat is that the NTK matrix H_t is computed not from A_t but from its normalized version $B_t = \text{diag}(z_t) A_t$ (see Appendix A.2). The difference is of little importance when only the first layer is trained. But in our multi-rate framework, the property 1) only controls ℓ_2 -norm of z_t , it still may contain unacceptably large values. This is where the second bullet in the description of the quasi-random stage comes into play: the set of neurons Γ referred to there will simply consist of those $\nu \in [S]$ for which $(z_t)_\nu$ is small; its size is controlled by the bound on $\|z_t\|$. Then we will get the restricted version $\|(B_t - B_0)_\Gamma * X\| \leq \tilde{o}(S^{1/2})$ of (13) which is good enough since $\sigma_{\min}(\|(B_0)_\Gamma * X\|) \geq \tilde{\Omega}(S^{1/2})$.

4 Experiments

Since our paper does not claim any direct relevance to practical data, we have confined our experiments to synthetic data generated precisely as in Section 1.3, with $\zeta \in_R \{\pm 1\}$. In most experiments, the vector y of label data was chosen from $\mathcal{N}(0, S)^m$. We also performed a few experiments with other natural choices:

low spectrum $e_0 = f_0 - y$ is the eigenvector of the matrix H_0 corresponding to the smallest eigenvalue and scaled in such a way that $\|e_0\| = m^{1/2} S^{1/2}$.

high spectrum $e_0 = n^{1/2} S^{1/2} (X^\top X^1)$. This is a ‘‘natural choice’’ for which $\|H_0 e_0\|$ is large.

⁷The proof for $\|z_T - z_0\|$ is similar. Note that the actual proof in Appendix A.4 uses integral notation, with the learning trajectory linearized in the obvious way.

local $e_0 = \begin{pmatrix} m^{1/2} S^{1/2} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. The initial error is completely concentrated on the first data instance (and is huge).

We considered only the case $\eta_z = 0$, i.e., training the first layer. All our experiments were equipped with the “safety valve”: they terminated once the situation $\|e_\tau\| > \|e_{\tau-1}\|$ was encountered. A computation was declared successful and stopped once $\|e_\tau\| < 10^{-3}$. We remark that due to our choice of normalization, the initial value was roughly of order $10^2 - 10^4$ so our stopping condition corresponds to the loss reduction by a factor of $10^{10} - 10^{14}$.

Like any other theoretical paper, ours is plugged with asymptotic notation (we would like to note in passing that attempts to replace it with explicit constants are usually not very instructive or even readable). The original purpose of our experiments was rather limited: verify that our theoretical findings are still reasonably relevant for relatively small values of parameters. What they have showed instead is that the crucial quantity $\lambda_{\min}(H_t)$ still behaves smoothly and nicely well beyond what could be optimistically called the “NTK regime”, almost all the way up to the representability barrier $m \sim nS$. In simpler words, there should be a significant room for improving theoretical results on convergence with methods yet to be developed. In retrospect, this is not surprising at all: the bounds like the first inequality in (42) are hopelessly pessimistic. But it is nice to have an experimental encouragement.

For finer details, including tables and figures, see Appendix H.

5 Conclusion

In this paper we have improved on previously known convergence guarantees of the plain gradient descent in the case of shallow networks. The real significance of this whole line of work depends on the yet elusive properties of data and labels that allow for the generalization. In the optimistic scenario this line of research will be eventually extended into more practical regions and then it might turn out to be helpful to explain why certain choices of labels allow us not only to converge but also to generalize (again, cf. [ZBH⁺21]). In the other scenario, generalization will be explained in such a way that the implied convergence will clearly follow from this ad-hoc, label-dependent explanation. That might render this whole line of research on label-independent convergence obsolete.

But while the jury is out, let us briefly sketch what, in our view, are worthy directions for future work assuming the first scenario.

First and foremost, it would be nice to be able to extend our knowledge for depth 2 networks to networks of larger constant depth, without significantly weakening the results.

It would be extremely nice (and probably very difficult, too) to prove convergence outside of the NTK regime, i.e. find more intelligent ways of controlling the learning trajectory than simply by the distance travelled. A significant improvement along these lines would be to remove or relax the assumption $m \ll S$ ubiquitously present in our and all previous work. The invariant from [ACH18, DHL18] that we will explore in Appendix G seems to be a paradigmatic example of the tools that are to be developed for the purpose.

It would be interesting to pinpoint those “quasirandom properties” from Appendix A.3, as well as the previous work, that are most problematic from the practical viewpoint. That might also give a good indication of what might be promising directions to generalize known results.

Finally, many, if not most, applications do not fit the rigid set-up in Section 1.1: they either deal with (multi)-classification problems or with more sophisticated network architectures, like CNN or Resnet, or both. A great deal of interesting work was done in these directions; some of it was cited in Section 1. But our impression is that less effort has been invested into working out more general united theory that would be less prone to this kind of changes in the model. That, in our view, is another interesting, and potentially more accessible, goal.

References

- [ACH18] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 244–253, 2018.
- [ADH⁺19a] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPC)*, pages 8139–8148, 2019.
- [ADH⁺19b] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 322–332, 2019.
- [AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 254–263, 2018.
- [ALS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 242–252, 2019.
- [BAM22] Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. Technical Report 2205.10217 [stat.ML], arxiv e-print, 2022.
- [BFT17] Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPC)*, pages 6240–6249, 2017.
- [BGMS18] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [CCZG21] Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep ReLU networks? In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [CGW89] F. Chung, R. Graham, and R. Wilson. Quasi-random graphs. *Combinatorica*, 9:345–362, 1989.
- [COB19] L  na  c Chizat, Edouard Oyallon, and Francis R. Bach. On lazy training in differentiable programming. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPC)*, pages 2933–2943, 2019.
- [CR23] L. Coregliano and A. Razborov. Natural quasirandom properties. *Random Structures and Algorithms*, 2023.
- [DHL18] Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPC)*, pages 382–393, 2018.
- [DZPS19] Simon S. Du, Xiyu Zhai, Barnab  s P  czos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

- [FCB22] Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, pages 2668–2703, 2022.
- [FDK⁺20] Stanislav Fort, Gintare K. Dziugaite, Sepideh Kharaghan, Daniel M. Roy, Mansheej Paul, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPC)*, 2020.
- [HY20] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 4542–4551, 2020.
- [JT20a] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPC)*, 2020.
- [JT20b] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.
- [LL18] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPC)*, pages 8168–8177, 2018.
- [LL20] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.
- [LZ22] Bochen Lyu and Zhanxing Zhu. Implicit bias of adversarial training for deep neural networks. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.
- [MZ20] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. Technical Report 2007.12826 [stat.ML], arxiv e-print, 2020.
- [NBS18] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [Ngu21] Quynh Nguyen. On the proof of global convergence of gradient descent for deep ReLU networks with linear widths. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8056–8062, 2021.
- [NMM21] Quynh Nguyen, Marco Mondelli, and Guido F. Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8119–8129, 2021.
- [OS20] Samet Oymak and Mahdi 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.
- [SH18] Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [SRP⁺21] Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPC)*, pages 11247–11259, 2021.

- [SY19] Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix Chernoff bound. Technical Report 1906.03593 [cs.LG], arxiv e-print, 2019.
- [Tro12] Joel Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.
- [Ver12] Roman Vershinin. *Introduction to the non-asymptotic analysis of random matrices*, chapter 5, pages 210–268. Cambridge University Press, 2012.
- [Ver18] Roman Vershinin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
- [VL22] Tiffany J. Vlaar and Benedict J. Leimkuhler. Multirate training of neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 22342–22360, 2022.
- [VSS22] G. Vardi, O. Shamir, and N. Srebro. On margin maximization in linear and relu networks. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPC)*, 2022.
- [WDW19] Xiaoxia Wu, Simon S. Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. Technical Report 1902.07111 [cs.LG], arxiv e-print, 2019.
- [XLS17] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, (AISTATS)*, pages 1216–1224, 2017.
- [ZBH⁺21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.
- [ZCZG20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.
- [ZG19] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPC)*, pages 2053–2062, 2019.

A Proof of the main result

As noted before, we defer the simple proof that the gradient descent avoids singular points a.e. to Appendix B and for now simply assume that $y \in \mathbb{R}^m$ is chosen in such a way that this is true, and such that $\|y\| \leq \tilde{O}(m^{1/2} S^{1/2})$.

A.1 Useful facts

Shur’s theorem and inequalities. The Hadamard product $M \circ N$ of two PSD $m \times m$ matrices is again PSD and satisfies

$$\begin{aligned} \|M \circ N\| &\leq \max_{j \in [m]} |M_{jj}| \cdot \|N\|; \\ \lambda_{\min}(M \circ N) &\geq \min_{j \in [m]} |M_{jj}| \cdot \lambda_{\min}(N). \end{aligned}$$

If $A \in \mathbb{R}^{S \times m}$ and $X \in \mathbb{R}^{n \times m}$ then

$$(A * X)^\top (A * X) = (A^\top A) \circ (X^\top X).$$

In particular, applying Shur’s inequalities with $M \mapsto A^\top A$, $N \mapsto X^\top X$, we get

$$\begin{aligned} \|A * X\| &\leq \max_{j \in [m]} \|A^j\| \cdot \|X\|; \\ \sigma_{\min}(A * X) &\geq \min_{j \in [m]} \|A^j\| \cdot \sigma_{\min}(X). \end{aligned} \tag{14}$$

Norm bounds. For $z \in \mathbb{R}^S$, the norm $\|(\text{diag}(z)A) * X\|$ can be bound in two different (dual) ways:

$$\|(\text{diag}(z)A) * X\| \leq \|z\|_\infty \cdot \|A * X\|; \quad (15)$$

$$\|(\text{diag}(z)A) * X\| \leq \|z\| \cdot \max_{\nu \in S} \|X \text{diag}(A_\nu)\| \leq \|z\| \cdot \|A\|_\infty \cdot \|X\|. \quad (16)$$

Another handy inequality is

$$\|MN\|_F \leq \|M\|_F \cdot \|N\|. \quad (17)$$

ϵ -nets [Ver12, Chapter 5.2.2]. An ϵ -net in \mathbb{S}^{n-1} is any subset $\mathcal{N} \subseteq \mathbb{S}^{n-1}$ such that $\forall v \in \mathbb{S}^{n-1} \exists \xi \in \mathcal{N} (\|v - \xi\| \leq \epsilon)$.

Fact A.1 ([Ver12, Lemma 5.2]) For any ϵ, n there exists an ϵ -net \mathcal{N} in \mathbb{S}^{n-1} with $|\mathcal{N}| \leq (1 + \frac{2}{\epsilon})^n$.

Fact A.2 (similar to [Ver12, Lemma 5.3]) Let $M \in \mathbb{R}^{k \times n}$, $\sigma > 0$ and \mathcal{N} be an $\frac{\sigma}{2\|M\|}$ -net in \mathbb{S}^{n-1} such that $\forall \xi \in \mathcal{N} (\|M\xi\| \geq \sigma)$. Then $\sigma_{\min}(M) \geq \frac{\sigma}{2}$.

A.2 More notation

It will be convenient to extend the trajectory (11) to continuous time in the piecewise linear manner. That is, for any $t \geq 0$ we set

$$\begin{aligned} W_t &\stackrel{\text{def}}{=} W_{\lfloor t \rfloor} - \eta_w \{t\} \nabla^w \ell(\theta_{\lfloor t \rfloor}); \\ z_t &\stackrel{\text{def}}{=} z_{\lfloor t \rfloor} - \eta_z \{t\} \nabla^z \ell(\theta_{\lfloor t \rfloor}). \end{aligned}$$

Since θ_τ is regular for all $\tau \in \mathbb{N}$, the interval $[\tau, \tau + 1]$ may contain only finitely many t for which θ_t is singular. Hence those t can and will be ignored in our estimates based on integration.

We abbreviate $F(W_t), f(\theta_t), e(\theta_t)$ etc. to F_t, f_t, e_t respectively. Whenever θ_t is regular, we let

$$A_t \stackrel{\text{def}}{=} \sigma'(W_t X) \quad (\text{the activation matrix}).$$

If θ_t is singular, we let $A_t \stackrel{\text{def}}{=} A_{t+0}$ which makes all our constructs *upper* semi-continuous.

Next, let

$$B_t \stackrel{\text{def}}{=} \text{diag}(z_t) A_t \quad (\text{the weighted activation matrix}),$$

then $(B_t * X, F_t)$ is the transposed Jacobian $J_f(\theta_t)$ and hence

$$\begin{aligned} \nabla^w \ell(\theta_t) &= (B_t * X) e_t; \\ \nabla^z \ell(\theta_t) &= F_t e_t. \end{aligned}$$

We also let

$$\begin{aligned} H_t &\stackrel{\text{def}}{=} (B_t * X)^\top (B_t * X) = (X^\top X) \circ (B_t^\top B_t), \\ G_t &\stackrel{\text{def}}{=} F_t^\top F_t; \end{aligned}$$

these are (for integer t) the two components of the NTK matrix.

A.3 Quasirandom properties

We claim that with our choice of X, W_0, z_0 (see Section 1.3), the following are satisfied with probability $\geq 1 - (nS)^{-\omega(1)}$. The proofs are deferred to Appendix C (straightforward proofs) and Appendix D (not so straightforward).

A.3.1 Properties of data

Almost orthogonality.

$$\max_{j \neq j' \in [m]} \langle X^j, X^{j'} \rangle \leq \tilde{O}(n^{1/2}). \quad (18)$$

In the following property, as well as (27), the constants assumed in the \tilde{O} notation do not depend on k and j, R respectively.

Uniform bounds on norm.

$$\forall k \in [m] \left(\max_{J \in \binom{[m]}{k}} \|X^J\| \leq \tilde{O} \left(1 + \frac{k^{1/2}}{n^{1/2}} \right) \right). \quad (19)$$

In the following property, as well as in (23), the bounds in the existential quantifiers are completely explicit, see the proofs in Appendix C.

Dual bound on the minimum singular value.

$$\exists n^* \leq \tilde{O}(n) \left(\min_{J \in \binom{[m]}{n^*}} \sigma_{\min}((X^J)^\top) \geq \frac{n}{m} \right). \quad (20)$$

A.3.2 Properties of initialization

Rows of W_0 are large.

$$\min_{\nu \in [S]} \|(W_0)_\nu\| \geq \tilde{\Omega}(n^{1/2}). \quad (21)$$

The following two properties are entirely obvious under either Rademacher or Gaussian initialization of z_0 .

Entries of θ_0 are small.

$$\|W_0\|_\infty, \|z_0\|_\infty \leq \tilde{O}(1). \quad (22)$$

z_0 contains sufficiently many large entries.

$$\exists \zeta_0 \geq \tilde{\Omega}(1) \left(|\{ \nu \in [S] \mid |(z_0)_\nu| \geq \zeta_0 \}| \geq \tilde{\Omega}(S) \right). \quad (23)$$

A.3.3 Smooth properties

Regularity.

$$\theta_0 \text{ is regular.} \quad (24)$$

Data is almost orthogonal to initialization.

$$\|W_0 X\|_\infty \leq \tilde{O}(1). \quad (25)$$

Right order of the output at initialization.

$$\|f_0\|_\infty \leq \tilde{O}(S^{1/2}). \quad (26)$$

“Good behavior” for any data instance.

$$\forall j \in [m] \forall R \geq 0 \left(|\{ \nu \in [S] \mid |(W_0 X)_{\nu j}| \leq R \}| \leq \tilde{O}(SR + 1) \right). \quad (27)$$

A.3.4 NTK properties at initialization

NTK: second layer.

$$\lambda_{\min}(G_0) \geq \Omega(S). \quad (28)$$

In order to formulate our last (and the most difficult) property, let us fix $\zeta_0 \geq \tilde{\Omega}(1)$ as in (23), and set

$$\Gamma_0 \stackrel{\text{def}}{=} \{ \nu \in [S] \mid |(z_0)_\nu| \geq \zeta_0 \};$$

thus, $|\Gamma_0| \geq \tilde{\Omega}(S)$.

NTK: first layer.

$$m \leq \tilde{o}(nS^{1/2}) \implies \exists S^* \geq \tilde{\Omega} \left(\frac{n^2 S}{n^2 + m} \right) \left(\min_{\Gamma \in \left(\frac{\Gamma_0}{|\Gamma_0| - S^*} \right)} \lambda_{\min} \left(((A_0)_\Gamma * X)^\top ((A_0)_\Gamma * X) \right) \geq \Omega(S) \right). \quad (29)$$

Remark 6 We have not been able to remove the annoying term m in the denominator of the bound on S^* ; we believe it can be done with a better analysis than ours.

A.4 Deterministic part

According to the plan outlined in Section 3, from now on we are assuming that X, θ_0 are chosen arbitrarily in such a way that all properties in Appendix A.3 are satisfied. The labels $y \in \mathbb{R}^m$ are also arbitrary, as long as $\|y\| \leq \tilde{O}(m^{1/2} S^{1/2})$ and the gradient descent avoids singular points. The learning rates η_w, η_z are assumed to satisfy (5) and one of (6), (7), (8). All the way until Appendix F, however, we will only need that either $m \leq \tilde{o}(nS^{1/4})$ or $\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right)$.

Remark 7 Before we begin, let us make one technical remark. The definition (9) of the assumption $m \leq \tilde{o}(nS^{1/4})$ can be equivalently re-written as

$$\lim_{k \rightarrow \infty} \frac{\log \left(\frac{n^{(k)} (S^{(k)})^{1/4}}{m^{(k)}} \right)}{\log \log (n^{(k)} S^{(k)})} = \infty. \quad (30)$$

Likewise, $m \geq \tilde{\Omega}(nS^{1/4})$ means that the quantity in (30) is bounded. Since from every infinite sequence we can extract an infinite subsequence for which one of the two is true, we can freely assume w.l.o.g. that either $m \leq \tilde{o}(nS^{1/4})$ or $m \geq \tilde{\Omega}(nS^{1/4})$ holds. The same dichotomy applies to $\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right)$ vs. $\frac{\eta_z}{\eta_w} \leq \tilde{O} \left(\frac{m^2}{nS} \right)$.

We are going to prove the following upper bounds on the distance travelled in the parameter space:

$$m \leq \tilde{o}(nS^{1/4}) \implies \sup_{T > 0} \|W_T - W_0\|_F < \tilde{O}(m^{1/2}); \quad (31)$$

$$m \leq \tilde{o}(nS^{1/4}) \implies \sup_{T > 0} \|z_T - z_0\| < \tilde{O} \left(\frac{\eta_z^{1/2}}{\eta_w^{1/2}} m^{1/2} \right); \quad (32)$$

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right) \implies \sup_{T > 0} \|W_T - W_0\|_F < \tilde{O} \left(\frac{\eta_w^{1/2}}{\eta_z^{1/2}} m^{1/2} \right); \quad (33)$$

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right) \implies \sup_{T > 0} \|z_T - z_0\| < \tilde{O} \left(m^{1/2} \right). \quad (34)$$

With a slight abuse of notation⁸, we prove (31)-(34) by induction on T ; that is, we fix $T > 0$ and assume that these bounds hold for all $t < T$. We then claim that for all $t < T$ we also have

$$m \leq \tilde{o}(nS^{1/4}) \implies \lambda_{\min}(H_t) \geq \tilde{\Omega}(S); \quad (35)$$

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right) \implies \lambda_{\min}(G_t) \geq \tilde{\Omega}(S). \quad (36)$$

Let us prove (36) since it is easier. Given (28), it is sufficient to show that

$$\|F_t - F_0\| \leq \tilde{o}(S^{1/2}). \quad (37)$$

⁸To be completely impeccable, we should first fix the constants assumed in the right-hand sides.

For that we perform the calculation

$$\begin{aligned} \|F_t - F_0\| &\leq \|F_t - F_0\|_F = \|\sigma(W_t X) - \sigma(W_0 X)\|_F \\ &\leq \|(W_t - W_0)X\|_F \stackrel{(17)}{\leq} \|W_t - W_0\|_F \cdot \|X\|, \end{aligned} \quad (38)$$

where the third inequality hold since σ is 1-Lipschitz. We now check that (38) indeed implies (37).

First,

$$\|W_t - W_0\|_F \stackrel{(33)}{\leq} \tilde{O}\left(\frac{\eta_w^{1/2}}{\eta_z^{1/2}} m^{1/2}\right) \leq \tilde{o}\left(\frac{n^{1/2} S^{1/2}}{m^{1/2}}\right), \quad (39)$$

where for the second inequality we used the assumption in (36).

Next, as noted in Remark 7, we can assume w.l.o.g. that either $m \geq \tilde{\Omega}(nS^{1/4})$ or $m \leq \tilde{o}(nS^{1/4})$.

In the first case, the bound (19) for $k = m$ simplifies to $\|X\| \leq \tilde{O}\left(\frac{m^{1/2}}{n^{1/2}}\right)$. This, along with (39), implies (37).

If, on the other hand, $m \leq \tilde{o}(nS^{1/4})$, we can also employ (31) and refine (39) to

$$\|W_t - W_0\|_F \leq \min\left(\tilde{O}(m^{1/2}), \tilde{o}\left(\frac{n^{1/2} S^{1/2}}{m^{1/2}}\right)\right) \leq \min\left(\tilde{o}(S^{1/2}), \tilde{o}\left(\frac{n^{1/2} S^{1/2}}{m^{1/2}}\right)\right).$$

Multiplying this with the full version $\|X\| \leq \tilde{O}\left(1 + \frac{m^{1/2}}{n^{1/2}}\right)$ of (19) again gives us (37). Thus, the proof of (37) and hence also of (36) is completed.

The proof of (35) is more elaborate and is deferred to Appendix E.

The dynamics of the gradient descent in the error space is ruled by the differential equation

$$\dot{\ell}_t = \langle \nabla \ell(\theta_t), \dot{\theta}_t \rangle,$$

where

$$\dot{\theta}_t = -(\eta_w \nabla^w \ell(\theta_{[t]}), \eta_z \nabla^z \ell(\theta_{[t]})).$$

Hence

$$\frac{d}{dt} \|e_t\| = \frac{\frac{d}{dt} \|e_t\|^2}{2\|e_t\|} = \frac{\dot{\ell}_t}{\|e_t\|} = \frac{\langle \nabla \ell(\theta_t), \dot{\theta}_t \rangle}{\|e_t\|} = \frac{\langle \nabla \ell(\theta_t), \dot{\theta}_{[t]} \rangle}{\|e_t\|}.$$

Let

$$\delta_t \stackrel{\text{def}}{=} \nabla \ell(\theta_t) - \nabla \ell(\theta_{[t]})$$

(this measures how much the gradient may change during one step). We now have

$$\langle \nabla \ell(\theta_t), \dot{\theta}_{[t]} \rangle = \langle \nabla \ell(\theta_{[t]}), \dot{\theta}_{[t]} \rangle + \langle \delta_t, \dot{\theta}_{[t]} \rangle = -\eta_w \|\nabla^w \ell(\theta_{[t]})\|^2 - \eta_z \|\nabla^z \ell(\theta_{[t]})\|^2 + \langle \delta_t, \dot{\theta}_{[t]} \rangle.$$

In Appendices F, G we will prove that

$$\left| \langle \delta_t, \dot{\theta}_{[t]} \rangle \right| < \frac{1}{2} \langle \nabla \ell(\theta_{[t]}), \dot{\theta}_{[t]} \rangle \quad (40)$$

(the proof for the z -part is pretty straightforward, given the absolute bounds (5), but will require a new idea for the W -part due to the discontinuity of $\nabla^w \ell$).

Once we have that, we know that $\|e_t\|$ is decreasing or, more specifically,

$$\frac{d}{dt} \|e_t\| \leq \frac{\langle \nabla \ell(\theta_{[t]}), \dot{\theta}_{[t]} \rangle}{2\|e_t\|} \leq \frac{\langle \nabla \ell(\theta_{[t]}), \dot{\theta}_{[t]} \rangle}{2\|e_{[t]}\|}, \quad (41)$$

where the last inequality holds since $\|e_t\|$ is decreasing.

Note that $\|f_0\| \leq \tilde{O}(m^{1/2} S^{1/2})$ by (26) and, therefore, $\|e_0\| \leq \tilde{O}(m^{1/2} S^{1/2})$ due to our assumption on y . Integrating (41) from 0 to T gives us

$$\int_0^T \frac{\langle \nabla \ell(\theta_{[t]}), -\dot{\theta}_{[t]} \rangle}{\|e_{[t]}\|} dt \leq 2 \int_0^T \left(-\frac{d}{dt} \|e_t\| \right) dt = 2(\|e_0\| - \|e_T\|) \leq \tilde{O}(m^{1/2} S^{1/2}),$$

where the equality holds since $\|e_t\|$ is decreasing.

Using Cauchy-Schwartz, we now estimate

$$\begin{aligned}
\|W_T - W_0\|_F &\leq \eta_w \int_0^T \|\nabla^w \ell(\theta_{[t]})\| dt \\
&\leq \eta_w \left(\int_0^T \frac{\|\nabla^w \ell(\theta_{[t]})\|^2}{\|e_{[t]}\|} dt \right)^{1/2} \cdot \left(\int_0^T \|e_{[t]}\| dt \right)^{1/2} \\
&\leq \eta_w \left(\int_0^T \frac{\langle \nabla \ell(\theta_{[t]}), -\dot{\theta}_{[t]} \rangle}{\eta_w \|e_{[t]}\|} dt \right)^{1/2} \cdot \left(\int_0^T \|e_{[t]}\| dt \right)^{1/2} \\
&\leq \tilde{O} \left(\eta_w^{1/2} m^{1/4} S^{1/4} \cdot \left(\int_0^T \|e_{[t]}\| dt \right)^{1/2} \right)
\end{aligned} \tag{42}$$

and, likewise,

$$\|z_T - z_0\| \leq \tilde{O} \left(\eta_z^{1/2} m^{1/4} S^{1/4} \cdot \left(\int_0^T \|e_{[t]}\| dt \right)^{1/2} \right). \tag{43}$$

In order to estimate the last remaining term $\int_0^T \|e_{[t]}\| dt$, let us first assume that $m \leq \tilde{o}(nS^{1/4})$. In that case we have (35) and then we can continue (41) as follows:

$$\begin{aligned}
\frac{d}{dt} \|e_t\| &\leq -\tilde{\Omega} \left(\frac{\langle \nabla \ell(\theta_{[t]}), -\dot{\theta}_{[t]} \rangle}{\|e_{[t]}\|} \right) \leq -\tilde{\Omega} \left(\eta_w \frac{\|\nabla^w \ell(\theta_{[t]})\|^2}{\|e_{[t]}\|} \right) \\
&\leq -\tilde{\Omega}(\eta_w \lambda_{\min}(H_{[t]}) \|e_{[t]}\|) \leq -\tilde{\Omega}(\eta_w S \|e_{[t]}\|).
\end{aligned}$$

This (sub)differential equation solves to

$$\begin{aligned}
m \leq \tilde{o}(nS^{1/4}) &\implies \|e_t\| \leq \|e_0\| \cdot \exp(-\tilde{\Omega}(\eta_w t S)) \\
&\leq \tilde{O} \left(m^{1/2} S^{1/2} \exp(-\tilde{\Omega}(\eta_w t S)) \right).
\end{aligned} \tag{44}$$

Integrating from 0 to T gives us

$$m \leq \tilde{o}(nS^{1/4}) \implies \int_0^T \|e_{[t]}\| dt \leq \tilde{O} \left(\frac{m^{1/2}}{\eta_w S^{1/2}} \right) \tag{45}$$

and, likewise (but applying (36) this time),

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right) \implies \|e_t\| \leq \tilde{O} \left(m^{1/2} S^{1/2} \exp(-\tilde{\Omega}(\eta_z t S)) \right); \tag{46}$$

$$\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right) \implies \int_0^T \|e_{[t]}\| dt \leq \tilde{O} \left(\frac{m^{1/2}}{\eta_z S^{1/2}} \right). \tag{47}$$

Plugging (45), (47) into (42), (43) gives us all four inequalities (31)-(34). This completes their proof by joint induction on t , modulo facts deferred to Appendices B-G.

In particular, in the first case (6) we have (44), in the third case (8) we have (46) (and in the second case (7) we have both). This completes the proof of Theorem 1.1, with exponential speed of convergence.

B Gradient descent avoids singular points

Fix X, W_0, z_0 such that θ_0 is regular, that is, $W_0 X$ does not contain zero entries. Let $\eta_w, \eta_z \geq 0$ be arbitrary. Our goal in this section is to show that for Lebesgue almost all $y \in \mathbb{R}^m$, all points θ_τ ($\tau = 1, 2, 3, \dots$) are also regular.

Due to countable additivity, it is sufficient to prove that

$$\lambda(\{y \mid \theta_0(y), \theta_1(y), \dots, \theta_{\tau-1}(y) \text{ are regular, } \theta_\tau(y) \text{ is singular}\}) = 0$$

for any fixed positive integer τ . By the same token, it is sufficient to show that for any fixed collection $A_0, A_1, \dots, A_{\tau-1}$ of 0-1 $(S \times m)$ matrices and for any fixed $\nu \in [S]$, $j \in [m]$,

$$\lambda(\{y \mid \theta_0(y), \theta_1(y), \dots, \theta_{\tau-1}(y) \text{ are regular, } A_0(y) = A_0, A_1(y) = A_1, \dots, A_{\tau-1}(y) = A_{\tau-1}, (W_\tau(y)X)_{\nu j} = 0\}) = 0.$$

But once the activation matrices $A_0, A_1, \dots, A_{\tau-1}$ are fixed, a simple induction on τ shows that all our mappings, including $(W_\tau(y)X)_{\nu j}$, become polynomial in y . Moreover, if we initialize $y := f_0$ then θ_0 becomes the global minimum, $\theta_0(y) = \theta_1(y) = \dots = \theta_\tau(y)$ and hence $(W_\tau(y)X)_{\nu j} = (W_0 X)_{\nu j} \neq 0$. Hence this polynomial is not identically zero and therefore the set of its zeros has Lebesgue measure 0.

C Easy quasirandom properties

In this section we prove that all properties in Appendix A.3, except for (28), (29), hold with probability $\geq 1 - (nS)^{-\omega(1)}$; all proofs in this section are routine exercises. As a preliminary remark, note that the nature of the error probability $(nS)^{-\omega(1)}$ (along with $m \leq n^{O(1)}$) allows us to freely use the union bound over sets whose cardinality is polynomial in n, m, S .

(18) This is essentially [Ver18, Remark 3.2.5] but let us do estimates slightly more carefully.

By the remark just made, we can assume that j, j' are fixed. Let $x, y \sim \mathcal{N}(0, \frac{1}{n}I_n)$ be two independent samples. Then, due to isotropy, X^j and $X^{j'}$ can be alternately represented as $X^j \sim \frac{x}{\|x\|}$, $X^{j'} \sim \frac{y}{\|y\|}$ and hence $|\langle X^j, X^{j'} \rangle| = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$.

Now, for any $i \in [n]$, the random variables $x_i^2 - 1/n$, $y_i^2 - 1/n$, $x_i y_i$ are centered and $\|x_i^2 - 1/n\|_{\psi_1}$, $\|y_i^2 - 1/n\|_{\psi_1}$, $\|x_i y_i\|_{\psi_1} \leq O(1)$. [Ver18, Example 2.5.8(i); Lemma 2.7.7; Exercise 2.7.10]. Hence by Bernstein's inequality ([Ver18, Theorem 2.8.1]; plug $N \mapsto n$, $t \mapsto \frac{\log(nS)}{n^{1/2}}$), with probability $\geq 1 - \exp(-\Omega(\log(nS))^2) \geq 1 - (nS)^{-\omega(1)}$ we have $\|x\|^2, \|y\|^2 \geq 1 - \frac{\log(nS)}{n^{1/2}} \geq \frac{1}{2}$ and $|\langle x, y \rangle| \leq \frac{\log(nS)}{n^{1/2}} \leq \tilde{O}(n^{1/2})$. The bound (18) follows.

(19) We can treat each value of k individually (and then apply the union bound).

Let us first consider the case $k = m$. By [Ver12, Theorem 5.39] (plug $N \mapsto m$, $A \mapsto n^{1/2} \cdot X^\top$), there are absolute constants $c, C > 0$ such that for every $t > 0$,

$$\mathbf{P}\left[\|n^{1/2} \cdot X\| \leq C(m^{1/2} + n^{1/2}) + t\right] \geq 1 - 2\exp(-ct^2).$$

Set $t \stackrel{\text{def}}{=} m^{1/2} \log(nS)$. Then we get that $\|X\| \leq \tilde{O}\left(1 + \frac{m^{1/2}}{n^{1/2}}\right)$ with probability $\geq 1 - \exp(-\Omega(m \log(nS)^2)) \geq 1 - (nS)^{-\omega(1)}$.

Let now $k \in [m]$ be arbitrary and $J \in \binom{[m]}{k}$. Then, plugging in the above argument $m \mapsto k$, we see that $\|X^J\| \leq \tilde{O}\left(1 + \frac{k^{1/2}}{n^{1/2}}\right)$ with probability $\geq 1 - \exp(-\Omega(k \log(nS)^2))$, and $\exp(\Omega(k \log(nS)^2))$ dominates the number $\binom{m}{k} \leq \exp(k \log m)$ of choices of J (recall again that $m \leq n^{O(1)}$). Hence we can apply the union bound to complete the proof of (19).

(20) This time we have not been able to find a convenient off-the-shelf inequality so we will do a simple ad hoc argument using ϵ -nets.

We set $n^* \stackrel{\text{def}}{=} n(\log n)^2$ and $\epsilon \stackrel{\text{def}}{=} \frac{1}{m}$. By Fact A.1, we can find an ϵ -net \mathcal{N} in \mathbb{S}^{n-1} with $|\mathcal{N}| \leq \exp(O(n \log n))$. Then it is sufficient to prove that for any fixed $\xi \in \mathcal{N}$ we have

$$\mathbf{P}\left[\forall J \in \binom{[m]}{n^*} \|(X^J)^\top \xi\| \geq \frac{2n}{m}\right] \geq 1 - \exp(-\omega(n \log n)) \quad (48)$$

since then we can apply the union bound over all $\xi \in \mathcal{N}$, followed by an application of Fact A.2 (with $\sigma := n/m$; note that $\|X^J\| \leq \tilde{O}(1)$ by the already proven (19)).

In order to prove (48), by isotropy we can assume that $\xi = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, that is $X^\top \xi = X_1^\top$. Let

$R \stackrel{\text{def}}{=} \frac{n^{1/2}}{m}$, then (by an argument similar to the one used in the proof of (18)) $\mathbf{P}[|X_{1j}| \leq R] \leq 3Rn^{1/2} \leq \frac{3n}{m}$ and hence $\mathbf{E}[|\{j \mid |X_{1j}| \leq R\}|] \leq 3n \leq \frac{n^*}{3}$. Hence, by the plain Chernoff bound,

$$\mathbf{P}\left[|\{j \mid |X_{1j}| \leq R\}| \leq \frac{n^*}{2}\right] \geq 1 - \exp(-\Omega(n^*)) \geq 1 - \exp(-\omega(n \log n)).$$

On the other hand, this event logically implies that for every $J \in \binom{[m]}{n^*}$, $(X^J)^\top \xi$ contains at least $\frac{n^*}{2}$ entries X_{1j} with $|X_{1j}| \geq R$. Hence $\|(X^J)^\top \xi\| \geq R \left(\frac{n^*}{2}\right)^{1/2} \geq \frac{2n}{m}$. This completes the proof of (48) and hence of (20) as well.

(21) Fix $\nu \in [S]$. Like in the proof of (18), for every $i \in [n]$, $(W_0)_{\nu i}^2 - 1$ is a centered distribution with $\|(W_0)_{\nu i}^2 - 1\|_{\psi_1} \leq O(1)$. Applying to it Bernstein's inequality ([Ver18, Theorem 2.8.1], plug $N \mapsto m$, $X_i \mapsto (W_0)_{\nu i}^2 - 1$, $t \mapsto n/2$), we get

$$\mathbf{P}[\|(W_0)_\nu\|^2 \leq n/2] \geq 1 - \exp(-\Omega(n)) \geq 1 - (nS)^{-\omega(1)}.$$

(22) The inequality $\|W_0\|_\infty \leq \tilde{O}(1)$ is obvious. For $\|z_0\|_\infty \leq \tilde{O}(1)$, (3) implies that for sufficiently large C ,

$$\mathbf{P}[|\zeta| \leq (\log nS)^C] \geq 1 - (nS)^{-\omega(1)}. \quad (49)$$

(23) The bound (4) on the variance allows us to conclude that for some $\zeta_0 \geq \tilde{\Omega}(1)$, $\mathbf{E}[|\zeta|^2 \cdot \mathbf{1}(|\zeta| \geq \zeta_0)] \geq \tilde{\Omega}(1)$ and then the bound (49) implies $\mathbf{P}[|\zeta| \geq \zeta_0] \geq \tilde{\Omega}(1)$. The desired bound (23) now follows from the plain Chernoff inequality.

(24) is obvious (in fact, it holds with probability 1).

(25) Due to isotropy, every individual entry $(W_0 X)_{\nu j}$ is a standard Gaussian.

(26) By the just proven (25), we also have $\|F_0\|_\infty \leq \tilde{O}(1)$ and then, for any fixed F_0 , the j th entry of $f_0 = F_0^\top z_0$ is of the form $a_1 \zeta_1 + \dots + a_S \zeta_S$, where $a_\nu \stackrel{\text{def}}{=} (F_0)_{\nu j}$ is fixed, $|a_\nu| \leq \tilde{O}(1)$ and ζ_1, \dots, ζ_S are S independent copies of ζ . Now we only have to apply the sub-Gaussian Hoeffding inequality [Ver18, Theorem 2.6.2].

(27) Fix $j \in [m]$. Due to the term $+1$ in $RS + 1$, we can assume w.l.o.g. that $R \geq \frac{1}{S}$; we can also assume that $R \leq 1$ as otherwise the bound is trivial. By replacing R with the nearest rational of the form 2^{-h} ($h = 0, 1, 2, \dots$), we can assume that R takes on only $\tilde{O}(1)$ different values. Hence we can apply the union bound on R and assume that $R \in [\frac{1}{S}, 1]$ is fixed.

Now, for an fixed $j \in [m]$, the values $(W_0 X)_{\nu j}$ are i.i.d. from $\mathcal{N}(0, 1)$. Hence $\mathbf{P}[|(W_0 X)_{\nu j}| \leq R] \leq \tilde{O}(R)$ and now (27) follows from the plain Chernoff bound.

D NTK matrix at initialization

In this section we prove (28) and (29).

Let $w \sim \mathcal{N}(0, I_n)$. Recall that the *limit NTK matrices* (that we will denote by $H^w(X)$ and $H^z(X)$, respectively) are defined as the following expectations:

$$\begin{aligned} H^w(X) &\stackrel{\text{def}}{=} \mathbf{E}_w[(X^\top X) \circ (\sigma'(X^\top w)\sigma'(w^\top X))] \\ H^z(X) &\stackrel{\text{def}}{=} \mathbf{E}_w[\sigma(X^\top w)\sigma(w^\top X)]. \end{aligned}$$

The first step is to show that (18) implies $\lambda_{\min}(H^w(X)), \lambda_{\min}(H^z(X)) \geq \Omega(S)$. The argument is essentially the same as in [NMM21] but since we are working in way less general set-up, we prefer to present a self-contained (and simpler) proof.

The entries $H^w(X)_{jj'}, H^z(X)_{jj'}$ of those matrices depend only on $\langle X^j, X^{j'} \rangle$ and this dependence is provided by analytical (near 0) functions f^w, f^z , respectively such that all coefficients in their Taylor expansions are non-negative.

Indeed, by isotropy we can assume that

$$X^j = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X^{j'} = \begin{pmatrix} \cos \phi \\ \sin \phi \\ \vdots \\ 0 \end{pmatrix},$$

where $\phi \stackrel{\text{def}}{=} \arccos(\langle X^j, X^{j'} \rangle)$. Next, (w_1, w_2) is distributed as $(r \cos \psi, r \sin \psi)$, where r is the squared χ -distribution with 2 degrees of freedom, $\psi \sim_R [0, 2\pi]$ and r, ψ are independent. We now compute

$$(H^w(X))_{jj'} = (X^\top X)_{jj'} \cdot \frac{1}{2\pi} \int_{\phi-\pi/2}^{\pi/2} 1 d\psi = \langle X^j, X^{j'} \rangle \left(\frac{1}{2} - \frac{\phi}{2\pi} \right)$$

that is,

$$f^w(\gamma) = \gamma \left(\frac{1}{2} - \frac{\arccos \gamma}{2\pi} \right)$$

(this computation already appeared several times in the literature). Also, assuming $\phi \in [0, \pi]$,

$$(H^z(X))_{jj'} = \mathbf{E}[r^2] \cdot \frac{1}{2\pi} \int_{\phi-\pi/2}^{\pi/2} \cos(\phi - \psi) d\psi = \frac{\cos(\phi) \cdot (\pi - \phi) + \sin(\phi)}{2\pi},$$

that is

$$f^z(\gamma) = \frac{\gamma(\pi - \arccos(\gamma)) + \sqrt{1 - \gamma^2}}{2\pi}.$$

These functions have the following explicit Taylor expansions:

$$\begin{aligned} f^w(\gamma) &= \frac{1}{4}\gamma + \frac{1}{2\pi} \cdot \sum_{r=0}^{\infty} \frac{(2r)!}{4^r (r!)^2 (2r+1)} \gamma^{2r+2}, \\ f^z(\gamma) &= \frac{1}{2\pi} + \frac{1}{4}\gamma + \frac{1}{48\pi} \gamma^2 + \frac{1}{2\pi} \sum_{r=2}^{\infty} \frac{(2r-3)!!}{(2r)!!} \gamma^{2r} \end{aligned}$$

which verifies our claim.

Now, since $H^w(X)$ is obtained from the ordinary kernel matrix $X^\top X$ by entrywise applications of f^w , we have

$$H^w(X) = \frac{1}{4} X^\top X + \frac{1}{2\pi} \cdot \sum_{r=0}^{\infty} \frac{(2r)!}{4^r (r!)^2 (2r+1)} \underbrace{(X^\top X \circ \dots \circ X^\top X)}_{2r+2 \text{ times}}$$

and, by Shur's theorem, all summands in the right-hand side are positive semi-definite. Hence for every

fixed $r \geq 1$, $H^w(X) \succeq \Omega \left(\underbrace{(X^\top X \circ \dots \circ X^\top X)}_{2r \text{ times}} \right)$. Also, for a fixed $r > 0$ all off-diagonal entries

in $\underbrace{(X^\top X \circ \dots \circ X^\top X)}_{2r \text{ times}}$ are bounded as $\tilde{O}(n^{r/2})$ (due to (18)) which is $o(m)$ if $r > 0$ is large enough.

On the other hand, all diagonal entries are still $\Omega(1)$. Hence $\lambda_{\min} \left(\underbrace{(X^\top X \circ \dots \circ X^\top X)}_{2r \text{ times}} \right) \geq \Omega(1)$

which completes the proof of $\lambda_{\min}(H^w(X)) \geq \Omega(1)$. The argument for $\lambda_{\min}(H^z(X)) \geq \Omega(1)$ is identical.

The proof of (28) is now easy to complete by the matrix Chernoff inequality. Fix an X satisfying (18). Then G_0 is the sum of S independent copies of the random (rank 1) PSD matrix $\sigma(w^\top X)^\top \sigma(w^\top X)$ while $H^z(X)$ is its expectation. Moreover,

$$\|\sigma(w^\top X)^\top \sigma(w^\top X)\| = \|\sigma(w^\top X)\|^2 \leq \|w^\top X\|^2.$$

Let

$$H^z(X, w) \stackrel{\text{def}}{=} \begin{cases} \sigma(w^\top X)^\top \sigma(w^\top X) & \text{if } \|w^\top X\|^2 \leq m(\log(nS))^2 \\ 0 & \text{otherwise.} \end{cases}$$

Then $\|H^z(X, w)\| \leq \tilde{O}(m)$ and

$$\begin{aligned} \|\mathbf{E}_w[H^z(X, w)] - H^z(X)\| &\leq \mathbf{E}_w[\|w^\top X\|^2 \cdot \mathbf{1}(\|w^\top X\|^2 \geq m(\log(nS))^2)] \\ &\leq \mathbf{E}_w[\|w^\top X\|^2 \cdot \mathbf{1}(\|w^\top X\|_\infty \geq \log(nS))] \leq o(1) \end{aligned}$$

(for the last inequality, observe that $w^\top X$ is a tuple of standard Gaussian, albeit not necessary independent). This in particular implies that

$$\lambda_{\min}(\mathbf{E}_w[H^z(X, w)]) \geq \Omega(1),$$

and now we can apply Matrix Chernoff Inequality ([Tro12, Theorem 1.1]; plug $d \mapsto m$, $X_k \mapsto H^z(X, w)$, $R \mapsto \tilde{O}(m)$, $\mu_{\min} \mapsto \Omega(S)$, $\delta \mapsto 1/2$) to conclude that (28) holds with probability $\geq 1 - \exp(-\Omega(S/m)) \geq 1 - \exp(-\tilde{\omega}(1)) \geq 1 - (nS)^{-\omega(1)}$.

The proof of (29) is trickier, and, as we indicated above, there probably should be an easier way of doing this.

We first note that Γ_0 depends only on z_0 while $(A_0 * X)^\top (A_0 * X)$ depends only on X , W_0 . Hence we can fix an arbitrary $\Gamma_0 \subseteq [S]$ with $|\Gamma_0| \geq \tilde{\Omega}(S)$ and prove (29) conditioned by $\Gamma_0(z_0) = \Gamma_0$. After that we can simplify our notation by assuming w.l.o.g. that $\Gamma_0 = [S]$ (for the Rademacher initialization this holds automatically anyway).

For any $\Gamma \subseteq [S]$,

$$\|(A_0)_\Gamma * X\| \leq \|A_0 * X\| = \|(A_0^\top A_0) \circ (X^\top X)\|^{1/2} \stackrel{(14)}{\leq} S^{1/2} \cdot \|X\| \stackrel{(19)}{\leq} \tilde{O}\left(S^{1/2} \left(1 + \frac{m^{1/2}}{n^{1/2}}\right)\right).$$

Let now

$$\begin{aligned} \delta &\stackrel{\text{def}}{=} \lambda_{\min}(H^w(X)) \geq \Omega(1); \\ \sigma &\stackrel{\text{def}}{=} \min\left(\frac{\delta^{1/2}}{3}, \frac{1}{2}\right) S^{1/2}; \\ \epsilon &\stackrel{\text{def}}{=} \frac{\sigma}{2 \cdot \|(A_0)_\Gamma * X\|} \geq \Omega(m^{-1/2}). \end{aligned}$$

Fix an arbitrary ϵ -net $\mathcal{N} \subseteq \mathbb{S}^{m-1}$ of cardinality $\exp(\tilde{O}(m))$ (by Fact A.1). Set also

$$S^* \stackrel{\text{def}}{=} \lfloor \frac{n^2 S}{(n^2 + m) \log(nS)^C} \rfloor,$$

where $C \geq 1$ is a sufficiently large constant, also to be specified later. Note that

$$S^* \geq \tilde{\omega}(m), \tag{50}$$

due to $S \geq \tilde{\omega}(m)$ and the assumption $m \leq \tilde{\omega}(nS^{1/2})$ incorporated in (29). Then it is sufficient to show that for any *fixed* $\xi \in \mathcal{N}$ we have

$$\mathbf{P} \left[\forall \Gamma \in \binom{[S]}{S-S^*} \|((A_0)_\Gamma * X)\xi\| \geq \sigma \right] \geq 1 - \exp(-\tilde{\omega}(m)) \quad (51)$$

since after that we can apply the union bound over \mathcal{N} , and then Fact A.2.

Before proceeding with the formal proof, let us briefly explain the predicament we are facing. The bound $\lambda_{\min}(H^w(X)) \geq \Omega(1)$ we have already proven only implies that the *expectation* of the random variable $\xi^\top (X^\top X \circ (\sigma'(X^\top w)\sigma'(w^\top X)))\xi$ is bounded away from zero, and $\xi^\top (A_0 * X)^\top (A_0 * X)\xi$ is a sum of S independent copies of that variable. Since $S \geq \tilde{\omega}(m)$, we would have been done by a simple application of Chernoff's inequality if we knew that this random variable does not behave abnormally; say, if we knew that the *probability* that it is separated from 0 is also $\Omega(1)$ (cf. [SY19, Assumption 1.2.2]). We can not directly apply Markov's inequality since we only have an upper bound of $\tilde{O}(m/n)$ on the value of this random variable. Our way to rule out this pathological situation (i.e. when the large expectation is made by large values occurring with small probability) is to apply some ideas from the proof of the Lipschitz concentration inequality [Ver18, Theorem 5.1.4].

Returning to the formal argument, let us first express our random variable in more compact way:

$$\xi^\top (X^\top X \circ (\sigma'(X^\top w)\sigma'(w^\top X)))\xi = \|X(\sigma'(X^\top w) \circ \xi)\|^2.$$

Thus,

$$\mathbf{E}_w [\|X(\sigma'(X^\top w) \circ \xi)\|^2] \geq \lambda_{\min}(H^w(X)) = \delta.$$

For the rest of this section, it will be more convenient to assume that $w \in_R \sqrt{n} \cdot \mathbb{S}^{n-1}$; since σ' is invariant under positive scalings, this will not change anything. We will denote by μ the standard (Haar) measure on $\sqrt{n} \cdot \mathbb{S}^{n-1}$.

Let

$$K \stackrel{\text{def}}{=} \frac{\delta S}{2S^*}$$

and

$$\mathcal{W} \stackrel{\text{def}}{=} \{w \in \sqrt{n} \cdot \mathbb{S}^{n-1} \mid \|X(\sigma'(X^\top w) \circ \xi)\|^2 \geq K\}.$$

We split the analysis according to whether \mathcal{W} is small or large.

Case 1. $\mu(\mathcal{W}) \leq 1/m$.

This case is easy. Since

$$\|X(\sigma'(X^\top w) \circ \xi)\|^2 \leq \|X\|^2 \cdot \|\sigma'(X^\top w) \circ \xi\|^2 \leq \|X\|^2 \stackrel{(19)}{\leq} \tilde{O}\left(1 + \frac{m}{n}\right) \leq \tilde{\omega}(m),$$

we have that $\mathbf{E}[\|X(\sigma'(X^\top w) \circ \xi)\|^2 \cdot \mathbf{1}(w \in \mathcal{W})] \leq \tilde{\omega}(1)$. Hence if we consider the truncated (and scaled by K) function

$$f(w) \stackrel{\text{def}}{=} \min\left(\frac{1}{K}\|\sigma'(X^\top w) \circ \xi\|^2, 1\right),$$

we will still have $\mathbf{E}[f(w)] \geq \frac{2}{3} \frac{\delta}{K} = \frac{4}{3} \frac{S^*}{S}$. Noting that $f(w) \in [0, 1]$, we have by the plain Chernoff bound:

$$\mathbf{P}\left[f(w_1) + \dots + f(w_S) \geq \frac{5}{4}S^*\right] \geq 1 - \exp(-\Omega(S^*)) \stackrel{(50)}{\geq} 1 - \exp(-\tilde{\omega}(m)).$$

Now, if we remove from this sum S^* terms, it will get decreased by at most S^* and hence

$$\|((A_0)_\Gamma * X)\xi\|^2 \geq K \cdot \sum_{\nu \in \Gamma} f(w_\nu) \geq \frac{1}{4}KS^* = \frac{1}{8}\delta S,$$

for any $\Gamma \in \binom{[S]}{S-S^*}$. This completes the analysis of Case 1.

Case 2. $\mu(\mathcal{W}) \geq 1/m$.

In this case we will be able to achieve our dream goal and show that $\|X(\sigma'(X^\top w) \circ \xi)\| \geq 1$ with probability at least $1/2$.

For $\rho \geq 0$, let

$$\mathcal{W}_\rho \stackrel{\text{def}}{=} \{w \in \sqrt{n} \cdot \mathbb{S}^{n-1} \mid \exists w^* \in \mathcal{W}(\|w - w^*\| \leq \rho)\}.$$

Then, by the standard isoperimetric inequality we can fix $\rho \leq \tilde{O}(1)$ in such a way that $\mu(\mathcal{W}_\rho) \geq 2/3$. Our goal is to show that for all but a negligible fraction of “bad” $w \in \mathcal{W}_\rho$, we have $\|X(\sigma'(X^\top w) \circ \xi)\| \geq 1$.

In order to identify the first set of “bad” points in \mathcal{W}_ρ , set first

$$\phi \stackrel{\text{def}}{=} \frac{\log(nS)^{C_1}}{n^{1/2}},$$

where $C_1 \geq 1$ is large enough. Let

$$E(w) \stackrel{\text{def}}{=} \{j \in [m] \mid \|w^\top X^j\| \leq \phi\},$$

and let $\chi_w \in \{0, 1\}^m$ be the characteristic function of this set. Then, since $\sum_{j \in [m]} \xi_j^2 = 1$, we have the estimate

$$\mathbf{E}_w[\|\chi_w \circ \xi\|^2] = \sum_{j \in [m]} \xi_j^2 \cdot \mathbf{P}_w[\|w^\top X^j\| \leq \phi] \leq \tilde{O}(n^{-1/2}).$$

Hence we can choose $\psi \leq \tilde{O}(n^{-1/4})$ such that

$$\mathbf{P}_w[\|\chi_w \circ \xi\| \geq \psi] \leq \frac{1}{12},$$

and this is our first “bad” event.

For the second “bad” event we need to identify one more quasirandom property of the data X that in a sense is a uniform version of a dual to (27). Namely, for $R > 0$, $w \in \sqrt{n} \cdot \mathbb{S}^{n-1}$ and $X \in (\mathbb{S}^{n-1})^m$, let

$$\text{Bad}_R(w, X) \stackrel{\text{def}}{=} |\{j \in [m] \mid \|w^\top X^j\| \leq R\}| \geq \log(nS)^2 \cdot (mR + 1). \quad (52)$$

Similarly to the proof of (27), for any *fixed* w we have $\mathbf{P}_X[\text{Bad}_R(w, X)] \leq \exp(-\Omega(\log(nS))^2)$ and, averaging over w ,

$$\mathbf{P}[\text{Bad}_R(w, X)] \leq \exp(-\Omega((\log(nS))^2)) \leq 1 - (nS)^{-\omega(1)}.$$

On the other hand, if we set

$$\mathcal{B}_{R,X} = \{w \in \sqrt{n} \cdot \mathbb{S}^{n-1} \mid \text{Bad}_R(w, X)\},$$

then

$$\mathbf{P}[\text{Bad}_R(w, X)] = \mathbf{E}_X[\mu(\mathcal{B}_{R,X})].$$

Therefore,

$$\mathbf{P}_X[\mu(\mathcal{B}_{R,X}) \leq \tilde{o}(1)] \geq 1 - (nS)^{-\omega(1)}.$$

Let now

$$\mathcal{B}_X \stackrel{\text{def}}{=} \bigcup_{R>0} \mathcal{B}_{R,X}.$$

Then we still have

$$\mathbf{P}_X[\mu(\mathcal{B}_X) \leq \tilde{o}(1)] \geq 1 - (nS)^{-\omega(1)}$$

since (cf. the proof of (27)) it is sufficient to consider only $\tilde{O}(1)$ different values of R in this union. Thus we also require that

$$\mu(\mathcal{B}_X) \leq \frac{1}{12}.$$

We stress that this is the quasirandom property of the *data* X *only* and it does *not* depend in any way on the actual initialization W_0 .

The set \mathcal{B}_X is our second “bad event”, and now we let

$$\widetilde{\mathcal{W}}_\rho \stackrel{\text{def}}{=} \{w \in \mathcal{W}_\rho \mid \|\chi_w \circ \xi\| \leq \psi \wedge w \notin \mathcal{B}_X\};$$

note that $\mu(\widetilde{\mathcal{W}}_\rho) \geq \frac{2}{3} - 2 \cdot \frac{1}{12} = \frac{1}{2}$. We claim that

$$\forall w \in \widetilde{\mathcal{W}}_\rho \quad \|X(\sigma'(X^\top w) \circ \xi)\| \geq 1. \quad (53)$$

Indeed, fix $w \in \widetilde{\mathcal{W}}_\rho$ and let $w^* \in \mathcal{W}$ be such that $\|w - w^*\| \leq \rho \leq \widetilde{O}(1)$. Let

$$d \stackrel{\text{def}}{=} \sigma'(X^\top w) - \sigma'(X^\top w^*); \quad d \in \{-1, 0, 1\}^m.$$

We start with the obvious estimate

$$\|X(\sigma'(X^\top w) \circ \xi)\| \geq \|X(\sigma'(X^\top w^*) \circ \xi)\| - \|X(d \circ \xi)\| \geq K^{1/2} - \|X(d \circ \xi)\|$$

(the second inequality holds since $w^* \in \mathcal{W}$). Let now $D \stackrel{\text{def}}{=} \text{supp}(d)$, and split d as $d = d' + d''$, where $d' \stackrel{\text{def}}{=} d \circ \chi_w$ is the part corresponding to $E(w)$ and d'' corresponds to $co - E(w)$. Then we have the bound

$$\begin{aligned} \|X(d \circ \xi)\| &\leq \|X(d' \circ \xi)\| + \|X(d'' \circ \xi)\| \\ &\leq \|X^D\| \cdot \|d' \circ \xi\| + \|X^{D \setminus E(w)}\| \cdot \|d'' \circ \xi\| \\ &\leq \|X^D\| \cdot \|\chi_w \circ \xi\| + \|X^{D \setminus E(w)}\| \\ &\leq \psi \cdot \|X^D\| + \|X^{D \setminus E(w)}\|. \end{aligned} \quad (54)$$

Let us estimate $|D|$ and $|D \setminus E(w)|$ (and then we will apply (19)).

For any $j \in D$, we have

$$\|(w^* - w)^\top X^j\| \geq \|w^\top X^j\| \quad (55)$$

since $w^\top X^j$ and $(w^*)^\top X^j$ have the opposite sign. Thus, $\|(w^* - w)^\top X^D\| \geq \|w^\top X^D\|$; recalling that $\|w^* - w\| \leq \widetilde{O}(1)$, we obtain

$$\|w^\top X^D\| \leq \widetilde{O}(\|X^D\|) \stackrel{(19)}{\leq} \widetilde{O}\left(1 + \frac{|D|^{1/2}}{n^{1/2}}\right).$$

On the other hand, let

$$R \stackrel{\text{def}}{=} \frac{1}{m} \left(\frac{|D|}{2 \log(nS)^2} - 1 \right).$$

so that the right-hand side in (52) becomes $\frac{|D|}{2}$. Since $w \notin \mathcal{B}_{R,X}$, there are at least $\frac{|D|}{2}$ indices $j \in D$ such that $\|w^\top X^j\| \geq R$ which implies

$$\|w^\top X^D\| \geq \frac{1}{\sqrt{2}} R |D|^{1/2} \geq \widetilde{\Omega}\left(\frac{|D|^{3/2}}{m}\right) - \widetilde{O}\left(\frac{|D|^{1/2}}{m}\right).$$

Comparing now the upper and lower bounds on $\|w^\top X^D\|$, we get $\frac{|D|^{3/2}}{m} \leq \widetilde{O}\left(1 + \frac{|D|^{1/2}}{n^{1/2}} + \frac{|D|^{1/2}}{m}\right)$ which solves to $|D| \leq \widetilde{O}\left(m^{2/3} + \frac{m}{n^{1/2}}\right)$. By (19), this implies

$$\|X^D\| \leq \widetilde{O}\left(1 + \frac{m^{1/3}}{n^{1/2}} + \frac{m^{1/2}}{n^{3/4}}\right) \leq \widetilde{O}\left(1 + \frac{m^{1/2}}{n^{3/4}}\right), \quad (56)$$

where the second inequality holds simply because $\frac{m^{1/3}}{n^{1/2}} = \left(\frac{m^{1/2}}{n^{3/4}}\right)^{2/3}$.

To bound $|D \setminus E(w)|$, we again use (55) which, along with the definition of $E(w)$, gives us

$$\|(w - w^*)^\top X^{D \setminus E(w)}\| \geq \phi \cdot |D \setminus E(w)|^{1/2}.$$

On the other hand,

$$\|(w - w^*)^\top X^{D \setminus E(w)}\| \leq \rho \cdot \|X^{D \setminus E(w)}\| \stackrel{(19)}{\leq} \widetilde{O}\left(1 + \frac{|D \setminus E(w)|^{1/2}}{n^{1/2}}\right).$$

If the constant C_1 in the definition of ϕ is large enough (namely, exceeds the constant assumed in (19)), the second term $\tilde{O}\left(\frac{|D \setminus E(w)|^{1/2}}{n^{1/2}}\right)$ is dominated by $\phi|D \setminus E(w)|^{1/2}$ and therefore $\phi|D \setminus E(w)|^{1/2} \leq \tilde{O}(1)$. Thus $|D \setminus E(w)| \leq \tilde{O}(n)$ and (as always, by (19)),

$$\|X^{D \setminus E(w)}\| \leq \tilde{O}(1). \quad (57)$$

Plugging (56), (57) and $\psi \leq \tilde{O}(n^{-1/4})$ into (54), we finally conclude that

$$\|X(d \circ \xi)\| \leq \tilde{O}\left(1 + \frac{m^{1/2}}{n}\right) \leq K^{1/2} - 1,$$

provided the constant C in the definition of S^* is large enough. This completes the proof of (53).

The rest is easy. By plain Chernoff bound, $\left|\left\{\nu \in [S] \mid w_\nu^\top \in \widetilde{\mathcal{W}}_\rho\right\}\right| \geq S/3$ with probability $\geq 1 - \exp(-\Omega(S)) \geq 1 - \exp(-\tilde{\omega}(m))$. Since $S^* \leq o(S)$, removing any S^* entries will still leave us with $\geq \frac{S}{4}$ neurons $\nu \in [S]$ such that $\|X(\sigma'(X^\top w_\nu^\top) \circ \xi)\| \geq 1$. This completes the proof of (51) (since $\sigma \leq \frac{S^{1/2}}{2}$) and (29).

E NTK matrix at the first layer

In this section we prove (35), assuming that (31), (32) and (34) hold (for the same t).

Let

$$\Gamma \stackrel{\text{def}}{=} \left\{ \nu \in \Gamma_0 \mid |(z_t)_\nu| \geq \frac{1}{2}\zeta_0 \right\}.$$

Our first task is to check that this Γ satisfies the condition $|\Gamma_0 \setminus \Gamma| \leq \tilde{o}\left(\frac{n^2 S}{n^2 + m}\right)$ (and hence can be chosen in (29)). For that we note that for any $\nu \in \Gamma_0 \setminus \Gamma$, $|(z_t)_\nu - (z_0)_\nu| \geq \frac{1}{2}\zeta_0 \geq \tilde{\Omega}(1)$ and hence $\|z_t - z_0\| \geq \tilde{\Omega}(|\Gamma_0 \setminus \Gamma|^{1/2})$. So we only have to check that

$$\|z_t - z_0\| \leq \min\left(\tilde{o}(S^{1/2}), \tilde{o}\left(\frac{nS^{1/2}}{m^{1/2}}\right)\right). \quad (58)$$

We can assume (see Remark 7) that either $\frac{\eta_z}{\eta_w} \geq \tilde{\omega}\left(\frac{m^2}{nS}\right)$ or $\frac{\eta_z}{\eta_w} \leq \tilde{O}\left(\frac{m^2}{nS}\right)$.

If $\frac{\eta_z}{\eta_w} \geq \tilde{\omega}\left(\frac{m^2}{nS}\right)$, we can apply (34) and use the condition $m \leq \tilde{o}(nS^{1/4})$ in (35).

If on the other hand $\frac{\eta_z}{\eta_w} \leq \tilde{O}\left(\frac{m^2}{nS}\right)$ then the cases (7), (8) are ruled out and hence (6) must hold. Now we apply (32) to conclude that $\|z_t - z_0\| \leq \tilde{O}\left(\frac{m^{3/2}}{n^{1/2}S^{1/2}}\right) \leq \tilde{o}\left(\frac{nS^{1/2}}{m^{1/2}}\right)$, where the last inequality follows from the calculation $m \leq \tilde{o}\left((nS^{1/4})^{3/4} \cdot (S)^{1/4}\right) \leq \tilde{o}(n^{3/4}S^{7/16}) \leq \tilde{o}(n^{3/4}S^{1/2})$. Alternatively, from (6) we have $\frac{\eta_z}{\eta_w} \leq \tilde{o}\left(\frac{S}{m}\right)$ which, also by (32), gives us $\|z_t - z_0\| \leq \tilde{o}(S^{1/2})$. Thus in either case we have (58).

We can now apply (29) to our particular Γ , and we get $\sigma_{\min}((A_0)_\Gamma * X) \geq \Omega(S^{1/2})$. Further, since $(z_t)_\nu \geq \tilde{\Omega}(1)$ whenever $\nu \in \Gamma$, we have

$$\sigma_{\min}((B_t)_\Gamma * X) \geq \tilde{\Omega}(\sigma_{\min}((A_t)_\Gamma * X)).$$

Thus, all that remains to prove is that $\sigma_{\min}((A_t)_\Gamma * X) \geq \frac{1}{2}\sigma_{\min}((A_0)_\Gamma * X)$, and for that it is sufficient to establish

$$\|(A_t - A_0) * X\| \leq \tilde{o}(S^{1/2}). \quad (59)$$

Let

$$D \stackrel{\text{def}}{=} \sup(A_t - A_0) = \left\{ (\nu, j) \in [S] \times [m] \mid \sigma'((W_t)_\nu X^j) \neq \sigma'((W_0)_\nu X^j) \right\}$$

and

$$D^j \stackrel{\text{def}}{=} \{ \nu \in [S] \mid (\nu, j) \in D \}.$$

Let $J \subseteq [m]$ consist of $\min(m, n)$ data instances $j \in [m]$ with the maximum value of $|D^j|$ and let $co - J \stackrel{\text{def}}{=} [m] \setminus J$. We will bound $\|(A_t - A_0)^J * X^J\|$ and $\|(A_t - A_0)^{co-J} * X^{co-J}\|$ as $\tilde{o}(S^{1/2})$ separately.

For the first term, (19) implies that $\|X^J\| \leq \tilde{O}(1)$. Hence, by Shur's inequality (14), it is sufficient to prove that $\max_{j \in J} |D^j| \leq \tilde{o}(S)$. Fix $j \in J$, and fix $R \geq \tilde{\Omega}\left(\frac{|D^j|}{S}\right)$ such that the right-hand side in (27) is $\leq \frac{|D^j|}{2}$. Then for at least $\frac{|D^j|}{2}$ values $\nu \in D^j$ we have $|(W_0 X)_{\nu j}| \geq R$ and hence

$$\|(W_0)_{D^j} X^j\| \geq \Omega\left(R \cdot |D^j|^{1/2}\right) \geq \tilde{\Omega}\left(\frac{|D^j|^{3/2}}{S}\right).$$

Also (cf. (55)) $\forall \nu \in D^j (|(W_0)_{\nu} X^j| \leq |(W_0 - W_t)_{\nu} X^j|)$. This gives us

$$\|(W_0 - W_t) X^j\| \geq \|(W_0 - W_t)_{D^j} X^j\| \geq \|(W_0)_{D^j} X^j\| \geq \tilde{\Omega}\left(\frac{|D^j|^{3/2}}{S}\right). \quad (60)$$

On the other hand,

$$\|(W_0 - W_t) X^j\| \leq \|W_0 - W_t\| \leq \|W_0 - W_t\|_{\text{F}} \stackrel{(31)}{\leq} \tilde{O}(m^{1/2}) \leq \tilde{o}(S^{1/2}).$$

Comparing these two bounds proves $\max_{j \in J} |D^j| \leq \tilde{o}(S)$ and thus $\|(A_t - A_0)^J * X^J\| \leq \tilde{o}(S^{1/2})$.

For the second term $\|(A_t - A_0)^{co-J} * X^{co-J}\|$, we can assume that $m \geq n$ (and hence $|J| = n$) as otherwise the statement is void. Let s be such that

$$\min_{j \in J} |D^j| \geq s \geq \max_{j \in co-J} |D^j|$$

(it exists due to our choice of J). We can now continue (60) as $\|(W_0 - W_t) X^j\| \geq \tilde{\Omega}\left(\frac{s^{3/2}}{S}\right)$ ($j \in J$) and then conclude $\|(W_0 - W_t) X^J\|_{\text{F}} \geq \tilde{\Omega}\left(\frac{s^{3/2} n^{1/2}}{S}\right)$.

On the other hand,

$$\|(W_0 - W_t) X^J\|_{\text{F}} \stackrel{(17)}{\leq} \|W_0 - W_t\|_{\text{F}} \cdot \|X^J\| \stackrel{(19), (31)}{\leq} \tilde{O}(m^{1/2}).$$

Comparing these two bounds gives us

$$s \leq \tilde{O}\left(\frac{S^{2/3} m^{1/3}}{n^{1/3}}\right). \quad (61)$$

Applying now (14), we see that

$$\begin{aligned} \|(A_t - A_0)^{co-J} * X^{co-J}\| &\leq \left(\max_{j \in co-J} |D^j|\right)^{1/2} \cdot \|X\| \stackrel{(19)}{\leq} \tilde{O}\left(\frac{s^{1/2} m^{1/2}}{n^{1/2}}\right) \\ &\leq \tilde{O}\left(\frac{S^{1/3} m^{2/3}}{n^{2/3}}\right) \stackrel{m \leq \tilde{o}(n S^{1/4})}{\leq} \tilde{o}(S^{1/2}). \end{aligned}$$

This completes the proof of (59) and hence also of (35).

F Gradient does not change radically between steps

In this section we prove (40) assuming that the bounds (31)-(34) hold for all $s < t$. By applying another auxiliary induction⁹ we can assume that (40) also holds for all $s < t$. In particular, for $s < t$ we may freely use all the conclusions made in Appendix A.4. Finally, if t is an integer then the

⁹Recall that due to our convention the gradient is upper semi-continuous and hence the set of those t for which (40) fails is violated is either empty or contains the minimum element.

statement is trivially true. Hence we can assume w.l.o.g. that all inequalities in Appendix A.4 hold for $\tau \stackrel{\text{def}}{=} \lfloor t \rfloor$.

Recall that

$$\begin{aligned}\nabla \ell(\theta_t) &= ((B_t * X)e_t, F_t e_t); \\ \dot{\theta}_\tau &= -(\eta_w, (B_\tau * X)e_\tau, \eta_z F_\tau e_\tau); \\ \left| \langle \nabla \ell(\theta_\tau), \dot{\theta}_\tau \rangle \right| &= \eta_w \|(B_\tau * X)e_\tau\|^2 + \eta_z \|F_\tau e_\tau\|^2.\end{aligned}$$

Let us introduce the uniform notation

$$\begin{aligned}\bar{\eta}_w &= \begin{cases} \eta_w & \text{if either (6) or (7) holds} \\ 0 & \text{otherwise} \end{cases} \\ \bar{\eta}_z &= \begin{cases} \eta_z & \text{if either (7) or (8) holds} \\ 0 & \text{otherwise} \end{cases}.\end{aligned}$$

Then by (35) and (36), the right-hand side of (40) gets bounded from below as

$$\begin{aligned}\left| \langle \nabla \ell(\theta_\tau), \dot{\theta}_\tau \rangle \right| &\geq \tilde{\Omega} \left(S^{1/2} \|e_\tau\| (\bar{\eta}_w \|(B_\tau * X)e_\tau\| + \bar{\eta}_z \|F_\tau e_\tau\|) \right) \\ &\geq \tilde{\Omega} (S \|e_\tau\|^2 (\bar{\eta}_w + \bar{\eta}_z)).\end{aligned}\tag{62}$$

In order to upper bound the left-hand side in (40), let us first collect some useful estimates; in what follows, $s < t$ is arbitrary.

First we have

$$\|W_s - W_0\|_F \leq \tilde{O}(m^{1/2}).\tag{63}$$

In the cases (6), (7) it follows from (31), and in the case (8) – from (33) (since $\eta_z \geq \eta_w$ in that case).

The following is the first part¹⁰ of (58):

$$\|z_s - z_0\| \leq \tilde{O}(S^{1/2});\tag{64}$$

along with (22) this gives us

$$\|z_s\| \leq \tilde{O}(S^{1/2}).$$

We can now bound the Jacobian. Namely, by (16) we have

$$\|B_s * X\| \leq \|z_s\| \cdot \|X\| \leq \tilde{O} \left(S^{1/2} \left(1 + \frac{m^{1/2}}{n^{1/2}} \right) \right)\tag{65}$$

and

$$\begin{aligned}\|F_s - F_0\|_F &\leq \|W_s X - W_0 X\|_F \leq \|W_s - W_0\|_F \cdot \|X\| \\ &\stackrel{(63), (19)}{\leq} \tilde{O} \left(m^{1/2} + \frac{m}{n^{1/2}} \right) \leq \tilde{O} \left(m^{1/2} S^{1/2} \right)\end{aligned}$$

which, along with (25), implies

$$\|F_s\|_F \leq \tilde{O} \left(m^{1/2} S^{1/2} \right).$$

Now we can also control the evolution from τ to t , both in the parameter space and in the feature space. Namely, recalling that we have already proved that $\|e_s\|$ is decreasing for $s < t$ and that $\|e_s\| \leq \tilde{O}(m^{1/2} S^{1/2})$, we have

$$\begin{aligned}\|W_t - W_\tau\|_F &= \eta_w(t - \tau) \|\nabla^w \ell(\theta_\tau)\| \leq \eta_w \cdot \|B_\tau * X\| \cdot \|e_\tau\| \\ &\leq \tilde{O} \left(\eta_w m^{1/2} S \left(1 + \frac{m^{1/2}}{n^{1/2}} \right) \right)\end{aligned}\tag{66}$$

¹⁰That part did not use the assumption $m \leq \tilde{O}(nS^{1/4})$.

and thus

$$\|F_t - F_\tau\|_F \leq \|W_t - W_\tau\|_F \cdot \|X\| \leq \tilde{O}\left(\eta_w m^{1/2} S \left(1 + \frac{m}{n}\right)\right). \quad (67)$$

In the W -department, we claim that

$$m \leq \tilde{o}(nS^{1/4}) \implies \|(B_t - B_\tau) * X\| \leq \tilde{o}(S^{1/2}). \quad (68)$$

Since our proof of this fact requires quite new ideas, it is postponed to Appendix G.

Next, for $s \in [\tau, t)$ we have

$$\dot{e}_s = -(\eta_w(B_s * X))^\top (B_\tau * X) + \eta_z F_s^\top F_\tau e_\tau$$

and hence by the above bounds on $\|B_s * X\|, \|F_s\|$,

$$\begin{aligned} \|\dot{e}_s\| &\leq \tilde{O}(\|e_\tau\| \cdot (\eta_w \|B_s * X\| \cdot \|B_\tau * X\| + \eta_z \|F_s\| \cdot \|F_\tau\|)) \\ &\leq \tilde{O}\left(S \cdot \|e_\tau\| \cdot \left(\eta_w \left(1 + \frac{m}{n}\right) + \eta_z m\right)\right) \stackrel{(5)}{\leq} \tilde{o}\left(\frac{\|e_\tau\|}{S}\right). \end{aligned}$$

Therefore,

$$\|e_t - e_\tau\| \leq \int_\tau^t \|\dot{e}_s\| ds \leq \tilde{o}\left(\frac{\|e_\tau\|}{S}\right). \quad (69)$$

Equipped with all this knowledge, we can now proceed to completing the proof of (40).

First, we claim that in (62) we can now replace $\bar{\eta}_w$ with η_w , i.e. that

$$\begin{aligned} \left| \langle \nabla \ell(\theta_\tau), \dot{\theta}_\tau \rangle \right| &\geq \tilde{\Omega}\left(S^{1/2} \|e_\tau\| (\eta_w \|B_\tau * X\| + \bar{\eta}_z \|F_\tau e_\tau\|)\right) \\ &\geq \tilde{\Omega}\left(S \|e_\tau\|^2 (\eta_w + \bar{\eta}_z)\right). \end{aligned} \quad (70)$$

Indeed, we only need to consider the last case (8). But then $\bar{\eta}_z = \eta_z$,

$$\eta_z \|F_\tau e_\tau\| \stackrel{(36)}{\geq} \tilde{\Omega}(\eta_z S^{1/2} \|e_\tau\|)$$

and

$$\eta_w \|(B_\tau * X)e_\tau\| \leq \tilde{O}\left(\eta_w S^{1/2} \left(1 + \frac{m^{1/2}}{n^{1/2}}\right) \|e_\tau\|\right).$$

Now, the condition in (8) implies that the former expression dominates the latter.

Second, let $\delta_t = (\delta_t^w, \delta_t^z)$, where

$$\begin{aligned} \delta_t^w &= (B_t * X)e_t - (B_\tau * X)e_\tau \\ \delta_t^z &= F_t e_t - F_\tau e_\tau; \end{aligned}$$

we bound their contributions separately.

More specifically,

$$\|\delta_t^w\| \leq \|(B_\tau * X)(e_t - e_\tau)\| + \|((B_t - B_\tau) * X)e_t\| \leq \|B_\tau * X\| \cdot \|e_t - e_\tau\| + \|(B_t - B_\tau) * X\| \cdot \|e_\tau\|.$$

We now do some case analysis.

If $m \leq \tilde{o}(nS^{1/4})$, we can apply (65), (69) and (68) to conclude that

$$\|\delta_t^w\| \leq \tilde{o}\left(\frac{\|e_\tau\|}{S^{1/2}} \left(1 + \frac{m^{1/2}}{n^{1/2}}\right) + S^{1/2} \|e_\tau\|\right) \leq \tilde{o}(S^{1/2} \|e_\tau\|).$$

Hence

$$\left| \langle \delta_t^w, \dot{\theta}_\tau \rangle \right| \leq \eta_w \|\delta_t^w\| \cdot \|(B_\tau * X)e_\tau\| \leq \tilde{o}\left(\eta_w S^{1/2} \|e_\tau\| \cdot \|(B_\tau * X)e_\tau\|\right) \leq \tilde{o}\left(\left| \langle \nabla \ell(\theta_\tau), \dot{\theta}_\tau \rangle \right|\right)$$

by (70).

If, on the other hand, (8) takes place then $\bar{\eta}_z = \eta_z$ and we bound $\|\delta_t^w\|$ trivially as

$$\|\delta_t^w\| \leq O((\|B_\tau * X\| + \|B_t * X\|) \cdot \|e_\tau\|) \leq \tilde{O}\left(S^{1/2} \left(1 + \frac{m^{1/2}}{n^{1/2}}\right) \|e_\tau\|\right)$$

and then

$$\begin{aligned} \left|\langle \delta_t^w, \dot{\theta}_\tau \rangle\right| &\leq \tilde{O}\left(\eta_w S^{1/2} \left(1 + \frac{m^{1/2}}{n^{1/2}}\right) \|e_\tau\| \cdot \|(B_\tau * X)e_\tau\|\right) \\ &\stackrel{(65)}{\leq} \tilde{O}\left(\eta_w \|e_\tau\|^2 S \left(1 + \frac{m}{n}\right)\right) \stackrel{(8)}{\leq} \tilde{o}(\eta_z \|e_\tau\|^2 S) \stackrel{(70)}{\leq} \tilde{o}\left(\|\nabla \ell(\theta_\tau), \dot{\theta}_\tau\|\right). \end{aligned}$$

This completes the analysis of the contribution of δ_t^w in (40).

The analysis of δ_t^z is analogous (and easier):

$$\begin{aligned} \|\delta_t^z\| &\leq \|F_\tau(e_t - e_\tau)\| + \|(F_t - F_\tau)e_t\|_F \leq \|F_\tau\| \cdot \|e_t - e_\tau\| + \|F_t - F_\tau\|_F \cdot \|e_\tau\| \\ &\stackrel{(69), (67), (66)}{\leq} \tilde{o}\left(\frac{m^{1/2}}{S^{1/2}} \|e_\tau\| + \eta_w m^{1/2} S \left(1 + \frac{m}{n}\right) \|e_\tau\|\right) \\ &\stackrel{(5)}{\leq} \tilde{o}\left(\frac{m^{1/2}}{S^{1/2}} \|e_\tau\|\right) \end{aligned}$$

and then

$$\left|\langle \delta_t^z, \dot{\theta}_\tau \rangle\right| \leq \tilde{o}\left(\eta_z \frac{m^{1/2}}{S^{1/2}} \|e_\tau\| \cdot \|F_\tau e_\tau\|\right).$$

If (7) or (8) takes place then $\bar{\eta}_z = \eta_z$ and we are done by (70).

If, on the other hand, (6) takes place then we can continue this estimate as

$$\left|\langle \delta_t^z, \dot{\theta}_\tau \rangle\right| \leq \tilde{o}\left(\eta_w \frac{S^{1/2}}{m^{1/2}} \|e_\tau\| \cdot (m^{1/2} S^{1/2}) \cdot \|e_\tau\|\right) = \tilde{o}(\eta_w S \|e_\tau\|^2),$$

and we are done again since $\bar{\eta}_w = \eta_w$.

This completes the proof of (40).

G Not too many activation changes

In this section we prove (68); as noted in the introduction, our proof uses (a discretized version of) the beautiful invariant discovered in [ACH18, DHL18].

Let us first remind the set-up: we are given a non-integer $t > 0$ such that for all $s < t$ we have all the facts and inequalities proven in Appendix A.4 as well as Appendix E. By continuity, we also have the seed inequalities (31)-(34) for our chosen t as well. The bound (40) is not guaranteed for this t ; in fact, this is exactly what we are proving. But it is used only in the integral form which means that we still have all conclusions from Appendix A.4 and Appendix E for our chosen t as well. We are specifically interested in (59): $\|(A_t - A_0) * X\| \leq \tilde{o}(S^{1/2})$.

Let us now start the argument. First, we have

$$\|(B_t - B_\tau) * X\| \leq \|(\text{diag}(z_t - z_\tau)A_\tau) * X\| + \|(\text{diag}(z_t)(A_t - A_\tau)) * X\|.$$

The estimate of the first term is immediate:

$$\|(\text{diag}(z_t - z_\tau)A_\tau) * X\| \stackrel{(16)}{\leq} \|z_t - z_\tau\| \cdot \|X\| \stackrel{(19)}{\leq} \tilde{O}\left(\|z_t - z_\tau\| \left(1 + \frac{m^{1/2}}{n^{1/2}}\right)\right).$$

The upper bound on $\|z_t - z_\tau\|$ is obtained via a computation completely analogous to (66):

$$\|z_t - z_\tau\| = \eta_z(t - \tau) \|\nabla^z \ell(\theta_\tau)\| \leq \eta_z \|F_\tau\| \cdot \|e_\tau\| \leq \tilde{O}(\eta_z m S); \quad (71)$$

note for the record (we will need it later) that (66) and (71) also hold for $t \mapsto s$, $\tau \mapsto \lfloor s \rfloor$, for an arbitrary $s < t$. Hence

$$\|(\text{diag}(z_t - z_\tau)A_\tau) * X\| \leq \tilde{O}\left(\eta_z m S \left(1 + \frac{m^{1/2}}{n^{1/2}}\right)\right) \stackrel{(5)}{\leq} \tilde{O}(1).$$

It remains to handle the second term, that is $\|(\text{diag}(z_t)(A_t - A_\tau)) * X\|$. We can identify yet another contribution that we already know how to handle:

$$\begin{aligned} \|(\text{diag}(z_0)(A_t - A_\tau)) * X\| &\stackrel{(15)}{\leq} \|z_0\|_\infty \cdot \|(A_t - A_\tau) * X\| \\ &\stackrel{(22)}{\leq} \tilde{O}(\|(A_t - A_0) * X\| + \|(A_\tau - A_0) * X\|) \\ &\stackrel{(59)}{\leq} \tilde{o}(S^{1/2}). \end{aligned}$$

Thus, all that remains to show is

$$\|(\text{diag}(z_t - z_0)(A_t - A_\tau)) * X\| \leq \tilde{o}(S^{1/2}), \quad (72)$$

and the difficulty is that we do not have good enough bound on $\|z_t - z_0\|_\infty$.

In order to circumvent this difficulty, let

$$\begin{aligned} D &\stackrel{\text{def}}{=} \sup(A_t - A_\tau); \\ D_\nu &\stackrel{\text{def}}{=} \{j \in [m] \mid (\nu, j) \in D\} \end{aligned}$$

and

$$\Gamma \stackrel{\text{def}}{=} \{\nu \in [S] \mid |D_\nu| \leq n^*\},$$

where $n^* \leq \tilde{O}(n)$ is as in (20). We split $(\text{diag}(z_t - z_0)(A_t - A_\tau)) * X$ in two parts, $(\text{diag}(z_t - z_0)(A_t - A_\tau))_\Gamma * X$ and $(\text{diag}(z_t - z_0)(A_t - A_\tau))_{co-\Gamma} * X$ and bound their norms separately.

The first one follows from what we already know:

$$\begin{aligned} \|(\text{diag}(z_t - z_0)(A_t - A_\tau))_\Gamma * X\| &\stackrel{(16)}{\leq} \|z_t - z_0\| \cdot \max_{\nu \in \Gamma} \|X \text{diag}((A_t - A_\tau)_\nu)\| \\ &\stackrel{(64)}{\leq} \tilde{o}\left(S^{1/2} \cdot \max_{\nu \in \Gamma} \|X^{D_\nu}\|\right) \stackrel{(19)}{\leq} \tilde{o}(S^{1/2}) \end{aligned}$$

(recall that $|D_\nu| \leq \tilde{O}(n)$ for $\nu \in \Gamma$ by our choice of Γ).

We bound the last remaining term as

$$\begin{aligned} \|(\text{diag}(z_t - z_0)(A_t - A_\tau))_{co-\Gamma} * X\| &\stackrel{(15)}{\leq} \|(z_t - z_0)_{co-\Gamma}\|_\infty \cdot \|(A_t - A_\tau) * X\| \\ &\stackrel{(59)}{\leq} \tilde{o}\left(S^{1/2} \|(z_t - z_0)_{co-\Gamma}\|_\infty\right). \end{aligned}$$

So it only remains to prove that $\|(z_t - z_0)_{co-\Gamma}\|_\infty \leq \tilde{O}(1)$ which, given (22), amounts to proving

$$\|(z_t)_{co-\Gamma}\|_\infty \leq \tilde{O}(1). \quad (73)$$

In words: for every particular neuron $\nu \in [S]$ we need to prove the dichotomy: either its weight on the second layer is small or it changes activation only on a small number of input data. This is where [ACH18, DHL18] steps in.

Let us fix a neuron $\nu \in co - \Gamma$. Pick an arbitrary $\tilde{D}_\nu \subseteq D_\nu$ of cardinality *exactly* n^* . Then we have the following chain of inequalities:

$$\begin{aligned} \|(W_\tau)_\nu\| &\leq \frac{m}{n} \|(W_\tau)_\nu X^{\tilde{D}_\nu}\| \leq \frac{m}{n} \|(W_\tau - W_t)_\nu X^{\tilde{D}_\nu}\| \leq \tilde{O}\left(\frac{m}{n} \|(W_\tau - W_t)_\nu\|\right) \\ &\leq \tilde{O}\left(\frac{m}{n} \|W_\tau - W_t\|_F\right) \leq \tilde{O}\left(\frac{m}{n} \eta_w m^{1/2} S \left(1 + \frac{m^{1/2}}{n^{1/2}}\right)\right) \leq \tilde{O}(1). \end{aligned}$$

The first inequality holds since $\sigma_{\min} \left(\left(X \tilde{D}_\nu \right)^\top \right) \geq \frac{n}{m}$ by (20). The second inequality holds since for any $j \in \tilde{D}_\nu$, $(W_\tau)_{\nu j}$ and $(W_t)_{\nu j}$ have opposite signs. The third inequality is true since $\|X \tilde{D}_\nu\| \leq \tilde{O}(1)$ by (19), the fourth is obvious and the fifth is (66). Finally, the sixth inequality follows from $\eta_w \leq \tilde{O}\left(\frac{n}{mS^2}\right)$ in (5). Hence by (21) we have

$$\|(W_0)_\nu\|^2 - \|(W_\tau)_\nu\|^2 \geq \tilde{\Omega}(n). \quad (74)$$

For any $s < t$ such that θ_s is regular, we have

$$\frac{d}{ds}(z_s)_\nu^2 = 2(z_s)_\nu \cdot (\dot{z}_s)_\nu = -2\eta_z(z_s)_\nu (F_{[s]})_\nu e_{[s]}$$

and

$$\frac{d}{ds}\|(W_s)_\nu\|^2 = 2\langle (W_s)_\nu, (\dot{W}_s)_\nu \rangle = -2\eta_w(z_{[s]})_\nu (W_s)_\nu X (e_{[s]} \circ (A_{[s]})_\nu^\top).$$

We note that

$$(W_{[s]})_\nu X (e_{[s]} \circ (A_{[s]})_\nu^\top) = (F_{[s]})_\nu e_{[s]}.$$

Hence if we let

$$R_s \stackrel{\text{def}}{=} \eta_w(z_s)_\nu^2 - \eta_z\|(W_s)_\nu\|^2$$

then

$$\begin{aligned} \|\dot{R}_s\| &\leq \tilde{O} \left(\eta_w \eta_z \left(|(z_{[s]})_\nu| \cdot |(W_{[s]} - W_s)_\nu X (e_{[s]} \circ (A_{[s]})_\nu^\top)| \right. \right. \\ &\quad \left. \left. + |(z_{[s]} - z_s)_\nu| \cdot |(F_{[s]})_\nu e_{[s]}| \right) \right) \\ &\leq \tilde{O} \left(\eta_w \eta_z \|e_{[s]}\| \cdot (\|z_{[s]}\| \cdot \|W_{[s]} - W_s\| \cdot \|X\| + \|z_{[s]} - z_s\| \cdot \|F_{[s]}\|) \right) \\ &\leq \tilde{O} \left(\eta_w \eta_z \|e_{[s]}\| \cdot m^{1/2} S^{3/2} \left(\eta_w \left(1 + \frac{m}{n} \right) + \eta_z m \right) \right), \end{aligned}$$

where for the last inequality we used the bounds from Appendix F, as well as (71). Let us now integrate this.

In the case (6) we have $\eta_w \left(1 + \frac{m}{n} \right) + \eta_z m \leq \tilde{O}(\eta_w S)$, and applying (45), $|R_\tau - R_0| \leq \int_0^\tau \|\dot{R}_s\| ds \leq \tilde{O}(\eta_w \eta_z m S^2)$.

In the two remaining cases (7), (8) we have (see Remark 5) $\frac{\eta_z}{\eta_w} \geq \tilde{\omega} \left(\frac{m^2}{nS} \right)$ and then $\eta_w \left(1 + \frac{m}{n} \right) + \eta_z m \leq \tilde{O} \left(\eta_z \left(\frac{nS}{m^2} + \frac{S}{m} + m \right) \right)$. Hence (47) gives us $|R_\tau - R_0| \leq \tilde{O} \left(\eta_w \eta_z m S \left(\frac{nS}{m^2} + \frac{S}{m} + m \right) \right)$. Thus in either case we have the bound

$$|R_\tau - R_0| \leq \tilde{O} \left(\eta_w \eta_z S^2 \left(m + \frac{n}{m} \right) \right) \stackrel{(5)}{\leq} \tilde{O}(n\eta_z). \quad (75)$$

On the other hand,

$$\begin{aligned} R_\tau - R_0 &= \eta_z(\|(W_0)_\nu\|^2 - \|(W_\tau)_\nu\|^2) - \eta_w((z_0)_\nu^2 - (z_\tau)_\nu^2) \\ &\stackrel{(74)}{\geq} \tilde{\Omega}(n\eta_z) - \eta_w((z_0)_\nu^2 - (z_\tau)_\nu^2). \end{aligned} \quad (76)$$

Comparing (75) and (76), we conclude¹¹ that $\|(z_\tau)_\nu\| \leq \|(z_0)_\nu\| \stackrel{(22)}{\leq} \tilde{O}(1)$. This concludes the proof of (73), (72) and (68).

H Experiments (details)

In all experiments the dimension was set to $n := 100$ and, unless otherwise noted, the learning rate was $\eta_w := 10^{-3}$. We did not try to optimize on the latter since this is not the main focus of the paper. But several sporadic experiments showed that for smaller values of S, m it can be significantly increased.

We kept track of the following control quantities (here T is the stopping time):

¹¹If $\eta_w = 0$ then $D = \emptyset$ and (68) is trivial.

- $\kappa_H \stackrel{\text{def}}{=} \frac{\lambda_{\min}(H_T)}{\lambda_{\min}(H_0)}$. Computing this quantity is computationally costly so we did not track the evolution of $\lambda_{\min}(H_T)$ systematically. But a spike-like behavior ($\lambda_{\min}(H_T)$ goes down and then up again) was never observed in several sporadic experiments in which it was tracked.
- $|D|$, where $D \subseteq [S] \times [m]$ is the set of all pairs (ν, j) that change their activation at least once during training;
- $\|W_T - W_0\|_F$. Same remark as in the first item applies (except that it is not very costly).

In the main series of our experiments, the vector $y \in \mathbb{R}^m$ was chosen from $\mathcal{N}(0, S)^m$, and we considered four values of S : $S = 100, 200, 500, 1000$. The number of input data m ranged from 100 to 1000, with step $\frac{S}{10}$. For every triple (n, S, m) we performed 10 experiments, with fresh values of (X, y, W_0, z_0) each time. *All these experiments resulted in convergence as defined above* but for $S \in \{500, 1000\}$, $m \geq 900$ we had to decrease the learning rate to $\eta_w = 5 \cdot 10^{-4}$ when $S = 500$ and to $\eta_w = 2 \cdot 10^{-4}$ when $S = 1000$.

The ranges for our control parameters are partially¹² reported in Table 1, and their average values are depicted on Figures 1, 2 and 3; for further comments and explanations see the respective captions.

In less systematic way, for $S = 100$ we also tried really large values (note that $nS = 10000$) $m = 1000, 2000, 3000, 4000, 5000$, with 10, 10, 6, 3 and 1 experiments, respectively. All of them have converged, Table 2 tabulates the results of these experiments; note that it quite smoothly extends the part of Table 1 for $S = 100$.

We also ran a few experiments (again, for $S = 100$) with other choices of the labels y mentioned at the beginning of Section 4. No significant discrepancies have been found with the high spectrum case and lower spectrum case.

¹²To make it comprehensible, we confine ourselves to $m = 100, 200, \dots, 1000$ for all four values of S . Our main purpose for presenting results in the tabular form is to demonstrate that there is sufficiently sharp concentration in 10 experiments performed for every pair (S, m) .

S	m	T	κ_H	$ D $	$\ W_T - W_0\ _F$
100	100	468-624; 528	0.89- 1.00; 0.96	657-1024; 805	15.91-22.86; 18.70
100	200	664-772; 699	0.88- 0.96; 0.93	2091-2482; 2232	27.10-31.74; 28.59
100	300	825-905; 861	0.84- 0.91; 0.89	3801-4449; 4074	34.81-39.73; 37.14
100	400	931-1054; 991	0.82- 0.88; 0.84	5763-6952; 6072	40.81-48.48; 43.33
100	500	1091-1176; 1129	0.80- 0.84; 0.82	8133-8802; 8397	47.43-51.65; 49.81
100	600	1216-1388; 1282	0.74- 0.84; 0.80	10858-12893; 11311	55.23-63.03; 57.30
100	700	1362-1492; 1418	0.76- 0.82; 0.78	13496-15722; 14282	60.32-68.55; 63.96
100	800	1489-1623; 1547	0.72- 0.78; 0.76	15974-17881; 16747	61.82-70.55; 67.37
100	900	1591-1816; 1685	0.71- 0.80; 0.75	19126-21581; 19958	68.92-78.38; 72.85
100	1000	1784-1915; 1840	0.71- 0.75; 0.73	22411-24654; 23255	76.52-81.95; 79.26
200	100	235-264; 245	0.95- 1.00; 0.98	888-1451; 1084	14.68-22.89; 17.77
200	200	306-330; 314	0.92- 0.98; 0.96	2785-3308; 3045	26.09-29.24; 27.51
200	300	362-384; 368	0.91- 0.94; 0.93	5163-5971; 5530	32.50-38.23; 34.80
200	400	406-441; 425	0.87- 0.92; 0.89	8116-9286; 8740	39.43-45.76; 42.61
200	500	446-505; 480	0.83- 0.90; 0.86	11173-13078; 12373	43.74-52.91; 49.63
200	600	508-538; 524	0.82- 0.86; 0.84	15004-17063; 16056	51.39-58.14; 54.92
200	700	547-590; 570	0.81- 0.84; 0.82	19367-21210; 20345	57.03-63.41; 60.18
200	800	610-654; 632	0.77- 0.82; 0.79	23235-25827; 24723	61.05-67.69; 65.17
200	900	674-707; 691	0.75- 0.77; 0.76	28437-31031; 29753	68.31-74.35; 70.89
200	1000	706-785; 728	0.72- 0.78; 0.76	32606-36192; 34320	70.12-80.08; 74.74
500	100	91-98; 95	0.98- 1.00; 0.99	1654-2065; 1822	16.23-20.71; 18.64
500	200	114-122; 117	0.98- 0.99; 0.98	4356-5386; 4869	24.75-30.71; 27.61
500	300	129-137; 133	0.95- 0.98; 0.97	8453-9347; 8898	32.96-37.27; 35.12
500	400	144-150; 148	0.94- 0.97; 0.95	13192-14614; 13684	39.68-43.53; 41.36
500	500	159-166; 163	0.92- 0.94; 0.93	18258-20345; 19450	44.44-49.85; 47.62
500	600	172-178; 175	0.90- 0.93; 0.92	24445-26661; 25699	48.71-55.22; 52.66
500	700	182-196; 187	0.87- 0.93; 0.90	30777-35985; 32429	54.55-65.46; 57.59
500	800	193-209; 201	0.86- 0.91; 0.88	36625-42393; 39962	57.16-66.52; 62.20
500	900	427-459; 437	0.84- 0.88; 0.86	44111-48943; 46445	62.97-70.43; 66.44
500	1000	463-478; 470	0.83- 0.85; 0.84	52602-58394; 55638	68.06-75.92; 72.29
1000	100	94-101; 96	0.99- 1.00; 1.00	2123-3066; 2546	15.78-22.89; 18.33
1000	200	112-119; 115	0.98- 1.00; 0.99	5831-7786; 6688	22.94-31.31; 26.70
1000	300	125-131; 130	0.98- 1.00; 0.98	11194-14368; 12620	31.81-40.68; 35.31
1000	400	137-146; 142	0.96- 0.99; 0.98	17651-19942; 18908	37.78-43.71; 40.45
1000	500	147-156; 152	0.95- 0.99; 0.97	25194-27879; 26485	43.62-49.31; 45.75
1000	600	158-166; 162	0.95- 0.98; 0.96	33262-37519; 35391	48.37-53.29; 51.03
1000	700	168-180; 174	0.92- 0.98; 0.95	43723-48665; 45819	54.62-59.54; 56.94
1000	800	180-189; 185	0.92- 0.95; 0.94	53662-60491; 56434	57.42-66.78; 61.55
1000	900	487-503; 498	0.91- 0.94; 0.92	62774-67471; 64791	63.01-68.03; 65.22
1000	1000	511-532; 520	0.90- 0.93; 0.91	74727-78658; 76410	67.13-70.88; 69.24

Table 1: Average values are shown in bold.

m	T	κ_H	$ D $	$\ W_T - W_0\ _F$
1000	1739 – 1948; 1859	0.7 – 0.76; 0.73	21835 – 24793; 23321	74.62 – 83.61; 79.73
2000	3363 – 3731; 3581	0.62 – 0.68; 0.64	57838 – 61459; 59157	125.67 – 136.35; 131.03
3000	6482 – 7088; 6689	0.53 – 0.57; 0.56	99676 – 102597; 100896	186 – 192.5; 189.2
4000	11558 – 12199; 11984	0.48 – 0.5; 0.49	146339 – 146489; 146410	245 – 255.3; 251.8
5000	28217	0.36	210147	360.8

Table 2: $S = 100$, m is large. Average values are shown in bold. The values for $m = 1000$ were recomputed anew.

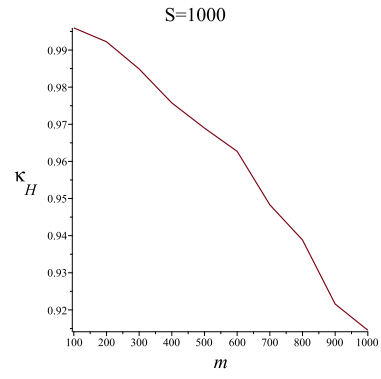
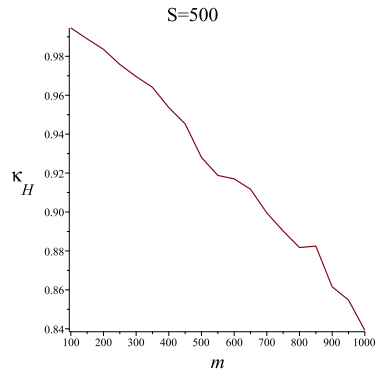
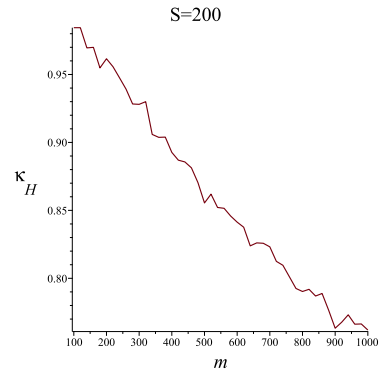
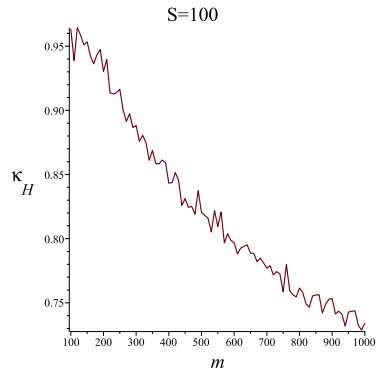


Figure 1: The dependence of κ_H on m .

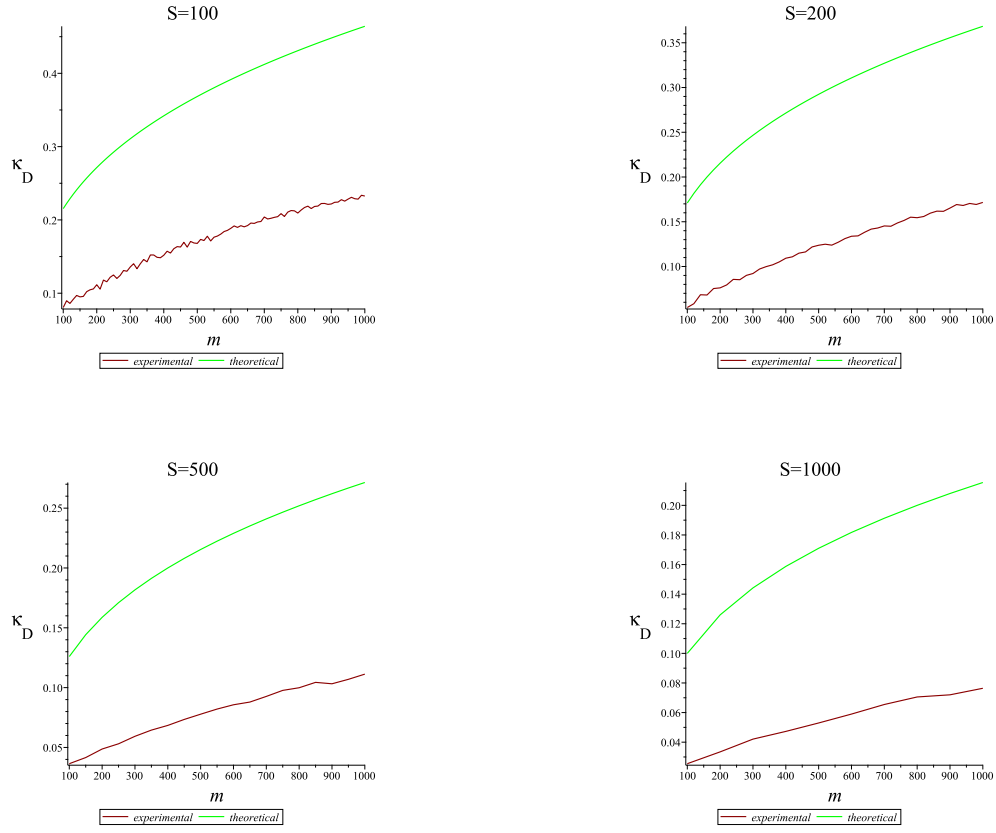


Figure 2: $\kappa_D \stackrel{\text{def}}{=} \frac{|D|}{|mS|}$ is the *density* of changes. The green line shows the theoretical prediction $\left(\frac{m}{nS}\right)^{1/3}$ obtained by dividing (61) by S .

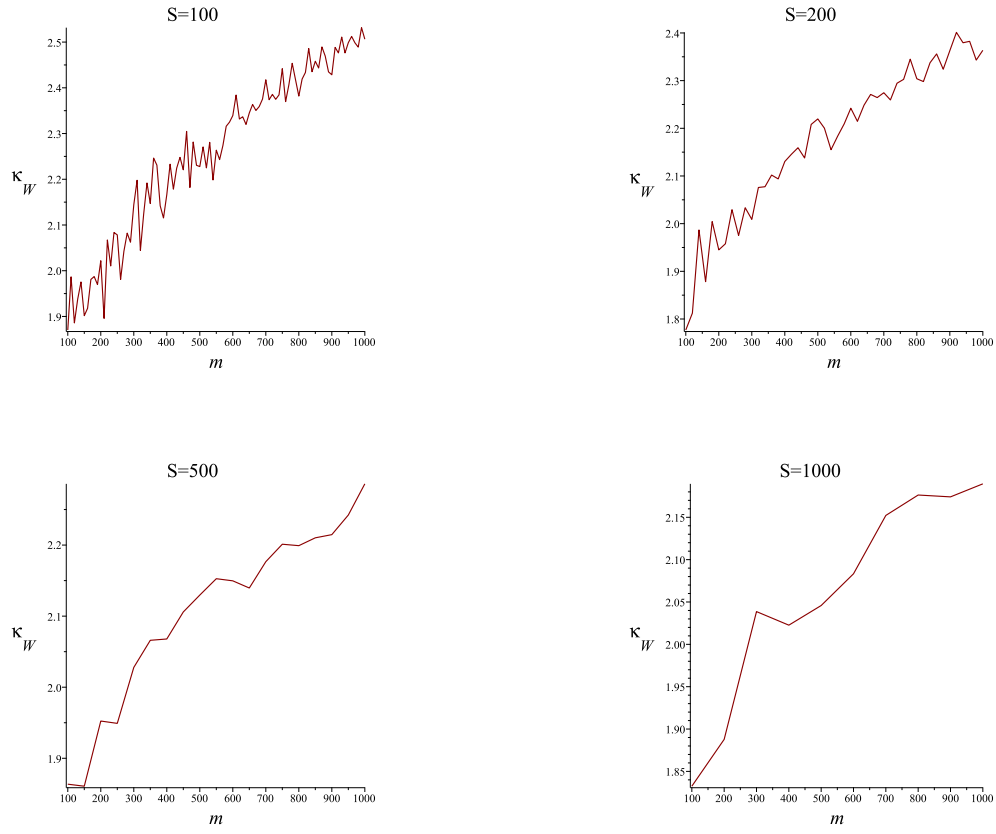


Figure 3: $\kappa_W \stackrel{\text{def}}{=} \frac{\|W_T - W_0\|_F}{m^{1/2}}$ (cf. (31))

The situation appears more interesting in the local case, that is when the initial error is placed on just one data point; in fact, this is the only situation where we have discovered anything that looks like a sharp transition. To report on the most striking experiment, for $n = S = 100$, $m = 1000$, $\eta_w = 10^{-3}$ the gradient descent converged 19 times out of 20. While for $m = 2000$, even after decreasing η_w to 10^{-4} , it aborted (according to the stopping rule $\|e_\tau\| > \|e_{\tau-1}\|$) in 10 (out of 10) attempts. In yet another experiment, decreasing the learning rate even further did not help convergence. But since these effects seem to be happening well beyond the range $m \ll S$ covered by our (and all previous) paper, we have not tried to investigate this systematically but rather defer this to future research.