Large-width asymptotics and training dynamics of α-Stable ReLU neural networks

Anonymous authors
Paper under double-blind review

Abstract

There is a recent literature on large-width properties of Gaussian neural networks (NNs), namely NNs with Gaussian distributed weights. Two popular results are: i) the characterization of the large-width asymptotic behavior of NNs in terms of Gaussian processes; ii) the characterization of the large-width training dynamics of NNs in terms of the so-called neural tangent kernel (NTK). In this paper, we investigate large-width asymptotics and training dynamics of α-Stable NNs, namely NNs whose weights are distributed according to α-Stable distributions, with $\alpha \in (0, 2]$. First, for α-Stable NNs with a ReLU activation function, we show that if the NN's width goes to infinity then a rescaled NN converges weakly to an α-Stable process, generalizing Gaussian processes. Differently from the Gaussian setting, our result shows that the choice of the activation function affects the scaling of the NN, that is: to achieve the infinitely wide α-Stable process, the ReLU activation requires an additional logarithmic term in the scaling with respect to sub-linear activations. Then, we characterize the large-width training dynamics of α-Stable ReLU-NNs in terms of a random kernel, referred to as the α-Stable NTK, showing that, for a sufficiently large width, the gradient descent achieves zero training error at a linear rate. The randomness of the α-Stable NTK is a further difference with respect to the Gaussian setting, that is: in the α-Stable setting, the randomness of the NN at initialization does not vanish in the large-width regime of the training.

1 Introduction

There is a growing literature on large-width properties of Gaussian neural networks (NNs), namely NNs with weights are Gaussian distributed (Neal, 1996; Williams, 1997; Der and Lee, 2006; Garriga-Alonso et al., 2018; Jacot et al., 2018; Lee et al., 2018; Matthews et al., 2018; Novak et al., 2018; Arora et al., 2019; Lee et al., 2019; Yang, 2019;a;b; Bracale et al., 2021; Eldan et al., 2021; Klukowski, 2021; Yang and Hu, 2021; Yang and Littwin, 2021; Basteri and Trevisan, 2022). Consider this setting: i) for $d, k \geq 1$ let X be the $d \times k$ NN's input, with $x_j = (x_{j1}, \dots, x_{jd})^T$ being the j-th input (column vector); ii) let ϕ be an activation function; iii) for $m \geq 1$ let $W = (w_1^{(0)}, \dots, w_m^{(0)}, w)$ be the NN's weights, such that $w_i^{(0)} = (w_{i1}^{(0)}, \dots, w_{id}^{(0)})$ and $w = (w_1, \dots, w_m)$ with the $w_{ij}^{(0)}$'s and the w_i's being i.i.d. as a Gaussian distribution with mean 0 and variance σ^2. If

$$f_m(x_j) = \sum_{i=1}^{m} w_i \phi(\langle w_i^{(0)}, x_j \rangle)$$

for $j = 1, \dots, k$, then $f_m(X) = (f_m(x_1), \dots, f_m(x_k))$ defines a (fully connected feed-forward) Gaussian ϕ-NN of width m. Neal (1996) investigated the large-width asymptotic behaviour of Gaussian ϕ-NNs. In particular, under suitable assumptions on ϕ, Neal (1996) showed that an application of the central limit theorem (CLT) leads to characterize the large-width distribution of the NN as follows: if $m \to +\infty$ then the rescaled NN $m^{-1/2} f_m(X)$ converges weakly to a Gaussian stochastic process with covariance function $\Sigma_{X,\phi}$ such that $\Sigma_{X,\phi}[r, s] = \sigma^2 \mathbb{E}[\phi(\langle w_i^{(0)}, x_r \rangle \phi(\langle w_i^{(0)}, x_s \rangle]$. Extensions are obtained for deep NNs (Matthews et al., 2018), general NN's architectures such as convolutional NNs (Yang, 2019a;b), and infinite-dimensional inputs (Bracale et al., 2021; Eldan et al., 2021).

In addition to the large-width asymptotic behavior, recent works have investigated the large-width training dynamics of Gaussian NNs, with the training being performed through gradient descent (Jacot et al., 2018; Arora et al., 2019; Du et al. , 2019; Lee et al., 2019). Let (X, Y) be the training set, where $Y = (y_1, \ldots, y_k)$ is the (training) output, with y_j being the (training) output for the j-th input x_j, let $f_m(X)$ be the Gaussian ϕ-NN, with ϕ to be the ReLU activation function, and set

$$\tilde{f}_m(W, X) = \frac{1}{m^{1/2}} f_m(X).$$

In particular, at random initialization $W(0)$ for the NN's weights, and assuming the squared-error loss function, the gradient flow of $W(t)$ leads to the training dynamics of $\tilde{f}_m(W(t), X)$, that is for $t \geq 0$

$$\frac{\mathrm{d}\tilde{f}_m(W(t), X)}{\mathrm{d}t} = -(\tilde{f}_m(W(t), X) - Y)\eta_m H_m(W(t), X), \tag{1}$$

where $\eta_m > 0$ is the (continuous) learning rate, and $H_m(W(t), X)$ is a $k \times k$ matrix whose (j, j') entry is $\langle \partial \tilde{f}_m(W(t), x_j)/\partial W, \partial \tilde{f}_m(W(t), x_{j'})/\partial W \rangle$. In such a context, Du et al. (2019) showed that if $\eta_m = 1$, then: i) the kernel $H_m(W(0), X)$ converges in probability, as $m \to +\infty$, to a deterministic kernel $H^*(X, X)$, which is referred to as the neural tangent kernel (NTK) (Jacot et al., 2018; Arora et al., 2019); ii) the least eigenvalue of $H^*(X, X)$ is bounded from below by a positive constant λ_0; iii) for m sufficiently large, the gradient descent achieves zero training error at a linear rate, i.e.

$$\|Y - \tilde{f}_m(W(t), X)\|_2^2 \leq \exp(-\lambda_0 t)\|Y - \tilde{f}_m(W(0), X)\|_2^2,$$

with high probability. We refer to the works of Arora et al. (2019), Yang (2019) and Yang and Littwin (2021) for several extensions of these results to deep NNs and also to more general architectures.

1.1 Our contributions

In this paper, we study large-width properties of α-Stable ReLU-NNs, namely NNs with a ReLU activation function and weights distributed according to α-Stable distributions (Samoradnitsky and Taqqu, 1994). For $\alpha \in (0, 2]$, α-Stable distributions form a class of heavy tails distributions, with $\alpha = 2$ being the Gaussian distribution. In his seminal work, Neal (1996) first considered α-Stable distributions to initialize NNs' weights, showing that while all Gaussian weights vanish in the infinitely wide limit, some α-Stable weights retain a non-negligible contribution, allowing to represent "hidden features" (Der and Lee, 2006; Fortuin et al., 2019; Lee et al., 2022). This is attributed to the diversity of the NN's path properties as $\alpha \in (0, 2]$ varies, which makes α-Stable NNs more flexible than Gaussian NNs. Motivated by these works, Favaro et al. (2020; 2021) characterized the large-width distribution of α-Stable ϕ-NN $f_m(X; \alpha)$ as follows: for $\alpha \in (0, 2)$ and a sub-linear ϕ, if $m \to +\infty$ then the rescaled NN $m^{-1/\alpha} f_m(X; \alpha)$ converges weakly to an α-Stable stochastic process, that is a process with α-Stable finite-dimensional distributions. Here, we extend this result to the ReLU activation, this being the most popular linear activation function. We show that if $m \to +\infty$, then the α-Stable ReLU-NN $(m \log m)^{-1/\alpha} f_m(X; \alpha)$ converges weakly to an α-Stable process. While for NNs with a single input, i.e. $k = 1$, such a result follows by an application of the generalized CLT for heavy tails distributions (Uchaikin and Zolotarev, 2011; Bordino et al., 2022), for $k > 1$ the generalized CLT does not apply, leading us to develop an alternative proof that may be of independent interest for multidimensional α-Stable distributions. It turns out that in the α-Stable setting, differently from the Gaussian setting, the choice of ϕ affects the scaling of the NN, that is: to achieve the infinitely wide α-Stable process, the use of the ReLU activation in place of a sub-linear activation results in a change of the scaling $m^{-1/\alpha}$ of the NN through the additional $(\log m)^{-1/\alpha}$ term.

Then, our main contribution consists in the study of the large-width training dynamics of α-Stable ReLU-NNs, generalizing to the α-Stable setting the result of Du et al. (2019), as well as soome results of Jacot et al. (2018) and Arora et al. (2019). For $\alpha \in (0, 2)$ and a training set (X, Y), we denote by

$$\tilde{f}_m(W, X; \alpha) = \frac{1}{(m \log m)^{1/\alpha}} f_m(X; \alpha)$$

the rescaled (model) output, and we consider the training of the NN performed through gradient descent under the squared-error loss function. By writing the training dynamics of $\tilde{f}_m(W(t), X; \alpha)$ as in equation 1, with η_m being the (continuous) learning rate and $H_m(W(t), X)$ the kernel in the α-Stable setting, we show that if $\eta_m = (\log m)^{2/\alpha}$ then: i) the rescaled kernel $(\log m)^{2/\alpha} H_m(W(0), X)$ converges in distribution, as $m \to +\infty$, to an $(\alpha/2)$-Stable (almost surely) positive definite random kernel $\tilde{H}^*(X, X; \alpha)$, which is referred to as the α-Stable NTK; ii) during training $t > 0$, for every $\delta > 0$ the least eigenvalue of $(\log m)^{2/\alpha} \tilde{H}_m(W(t), X; \alpha)$ remains bounded away from zero, for m sufficiently large, with probability $1 - \delta$; iii) for every $\delta > 0$ the gradient descent achieves zero training loss at a linear rate, for m sufficiently large, with probability $1-\delta$. The randomness of the α-Stable NTK is a further difference with respect to the Gaussian setting, and it makes the convergence analysis of the gradient descent more challenging than in the Gaussian setting. Our work is the first to investigate the large-width training dynamics of NNs with weights initialized through heavy tails distributions, and it shows that, within the α-Stable setting, the randomness of the NN at initialization does not vanish in the large-width regime of the training. Such a behaviour may be viewed as the counterpart, at the training level, of the large-width behaviour described in Neal (1996).

1.2 Organization of the paper

The paper is organized as follows. Section 2 contains some preliminary definitions on the multidimensional α-Stable distribution. In Section 3 we the study of the large-width distributions of α-Stable ReLU-NNs, characterizing the infinitely wide limit of a rescaled NN in terms of an α-Stable process. In Section 4 we study the large-width training dynamics of α-Stable ReLU-NNs, characterizing the infinitely wide dynamics in terms of the α-Stable NTK, and showing that, for a sufficiently large width, the gradient descent achieves zero training error at a linear rate, with high probability. Section 5 contains a discussion of our results, their extension to deep α-Stable NNs, and some directions for future work. Appendices contain the proofs and a brief review of α-Stable distributions.

2 Preliminaries on multidimensional α-Stable distributions

We recall the definition of the multidimensional α-Stable distribution. See Samoradnitsky and Taqqu (1994, Chapter 1 and Chapter 2). For $\alpha \in (0, 2]$, a random variable $S \in \mathbb{R}$ is distributed as a symmetric and centered 1-dimensional α-Stable distribution with scale $\sigma > 0$ if its characteristic function is

$$\mathbb{E}(\exp\{\mathrm{i}zS\}) = \exp\left\{-\sigma^{\alpha}|z|^{\alpha}\right\},$$

and we write $S \sim \mathrm{St}(\alpha, \sigma)$. The parameter α is typically referred to as the stability parameter. In particular, if $\alpha = 2$ then S is distributed according to a Gaussian distribution with mean 0 and variance σ^2. Let $\mathbb{S}^{k-1}$ be the unit sphere in $\mathbb{R}^k$, with $k \geq 1$, and let Γ be a symmetric finite measure on $\mathbb{S}^{k-1}$. For $\alpha \in (0, 2]$, we say that a random variable $S \in \mathbb{R}^k$ is distributed as a symmetric and centered k-dimensional α-Stable distribution with spectral measure Γ if its characteristic function is

$$\mathbb{E}(\exp\{\mathrm{i}\langle z, S\rangle\}) = \exp\left\{-\int_{\mathbb{S}^{k-1}} |\langle z, s\rangle|^{\alpha}\Gamma(\mathrm{d}s)\right\},$$

and we write $S \sim \mathrm{St}_k(\alpha, \Gamma)$. Let 1_r be the r-dimensional (column) vector with 1 in the r-th entry and 0 elsewhere, for any $r = 1, \ldots, k$. Then, the r-th element of S, that is $S1_r$ is distributed as an α-Stable distribution with scale

$$\sigma = \left(\int_{\mathbb{S}^{k-1}} |\langle 1_r, s\rangle|^{\alpha}\Gamma(\mathrm{d}s)\right)^{1/\alpha}.$$

We deal mostly with k-dimensional α-Stable distributions with discrete spectral measure, that is we consider measures of the form $\Gamma(\cdot) = \sum_{1 \leq i \leq n} \gamma_i \delta_{s_i}(\cdot)$ with $n \in \mathbb{N}$, $\gamma_i \in \mathbb{R}$ and $s_i \in \mathbb{S}^{k-1}$, for $i = 1, \ldots, n$ (Samoradnitsky and Taqqu, 1994, Chapter 2). Throughout this paper, it is assumed that all the random variables are defined on a common probability space, say $(\Omega, \mathcal{F}, \mathbb{P})$, unless otherwise stated.

We make use several times of the following characterization of the spectral measure of α-stable distributions: if $S \sim \mathrm{St}_k(\alpha, \Gamma)$, then for every Borel set B of $\mathbb{S}^{k-1}$ such that $\Gamma(\partial B) = 0$, it holds true that

$$\lim_{r\to\infty} r^{\alpha}\mathbb{P}\left(\|S\| > r, \frac{S}{\|S\|} \in B\right) = C_{\alpha}\Gamma(B),$$

where

$$C_{\alpha} = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} & \alpha \neq 1 \\ \frac{2}{\pi} & \alpha = 1. \end{cases} \tag{2}$$

The proof of this result is reported in Appendix B for completeness (Samoradnitsky and Taqqu, 1994, Chapter 2). Moreover, the distribution of a random vector ξ belongs to the domain of attraction of the $\mathrm{St}_k(\alpha, \Gamma)$ distribution, with $\alpha \in (0, 2)$ and Γ simmetric finite measure on $\mathbb{S}^{k-1}$, if and only if

$$\lim_{n\to\infty} n\mathbb{P}\left(||\xi|| > n^{1/\alpha}, \frac{\xi}{||\xi||} \in A\right) = C_{\alpha}\Gamma(A) \tag{3}$$

for every Borel set A of S such that $\Gamma(\partial A) = 0$. We refer to Appendix B for more details. In genera, we refer to the monograph Samoradnitsky and Taqqu (1994, Chapter 1 and Chapter 2) for further details on C_{α} within the context of the definition of the class of multidimensional α-Stable distributions.

3 Large-width asymptotics of α-Stable ReLU-NNs

To define an α-Stable ReLU-NNs, consider the following setting: i) for any $d, k \geq 1$ let X be the $d \times k$ NN's input, with $x_j = (x_{j1}, \ldots, x_{jd})^T$ being the j-th input (column vector); ii) for $m \geq 1$ let $W = (w_1^{(0)}, \ldots, w_m^{(0)}, w)$ be the NN's weights, such that $w_i^{(0)} = (w_{i1}^{(0)}, \ldots, w_{id}^{(0)})$ and $w = (w_1, \ldots, w_m)$. If

$$f_m(W, x_j; \alpha) = \sum_{i=1}^{m} w_i \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)$$

for $j = 1, \ldots, k$, where $I(\cdot)$ being the indicator function, then the ReLU-NN of width m is $f_m(W, X; \alpha) = (f_m(W, x_1; \alpha), \ldots, f_m(W, x_k; \alpha))$. Throughout the paper, we denote by $W(0) = (w_1^{(0)}(0), \ldots, w_m^{(0)}(0), w(0))$ the NN weights at random initialization. In particular, if the weight $w_{ij}^{(0)}$'s and w_i's are initialized as i.i.d. α-Stable random variables, with $\alpha \in (0, 2)$ and $\sigma > 0$, then $f_m(W(0), X; \alpha)$ defines an α-Stable ReLU-NN of width m. Without loss of generality we assume the scale parameter $\sigma = 1$. Further, the case $\alpha = 2$, which corresponds to the Gaussian setting, is excluded by our analysis, though some of our results remain valid also for $\alpha = 2$. The next theorem characterizes the infinitely wide limit of α-Stable ReLU-NNs. In particular, we denote by $Z_m \xrightarrow{w} Z$ the weak convergence, as $m \to +\infty$, of the sequence of random vectors $(Z_m)_{n\geq 1}$ to the random vector Z.

Theorem 3.1. *Let $f_m(W(0), X; \alpha)$ be an α-Stable ReLU-NN. If $m \to +\infty$ then*

$$\frac{1}{(m\log m)^{1/\alpha}} f_m(W(0), X; \alpha) \xrightarrow{w} f(X),$$

where $f(X) \sim St_k(\alpha, \Gamma_X)$, with the spectral measure Γ_X being of the following form:

$$\Gamma_X = \frac{C_{\alpha}}{4}\sum_{i=1}^{d}(\|[x_{ji}I(x_{ji} > 0)]_j\|^{\alpha})D_i^{+}(X) + \|[x_{ji}I(x_{ji} < 0)]_j\|^{\alpha})D_i^{-}(X)$$

such that

$$D_i^{+}(X) = \delta\left(\frac{[x_{ji}I(x_{ji} > 0)]_j}{\|[x_{ji}I(x_{ji} > 0)]_j\|}\right) + \delta\left(-\frac{[x_{ji}I(x_{ji} > 0)]_j}{\|[x_{ji}I(x_{ji} > 0)]_j\|}\right)$$

and

$$D_i^{-}(X) = \delta\left(\frac{[x_{ji}I(x_{ji} < 0)]_j}{\|[x_{ji}I(x_{ji} < 0)]_j\|}\right) + \delta\left(-\frac{[x_{ji}I(x_{ji} < 0)]_j}{\|[x_{ji}I(x_{ji} < 0)]_j\|}\right),$$

where, for any $s \in \mathbb{S}^{k-1}$, $\delta(s)$ is probability measure degenerate in s, and C_α is a constant defined in equation 2. The stochastic process $f(X) = (f(x_1), \ldots, f(x_k))$, as a process indexed by the NN's input X, is an α-Stable process with spectral measure Γ_X.

See Appendix A.1 for the proof of Theorem 3.1. For a broad class of bounded or sub-linear activation functions, Favaro et al. (2021) characterizes the large-width distribution of deep α-Stable NNs. In particular, let

$$f_m(x_j; \alpha) = \sum_{i=1}^{m} w_i \phi \langle w_i^{(0)}, x_j \rangle$$

be the α-Stable NN of width m for the input x_j, for $j = 1, \ldots, k$, with ϕ being a bounded activation function. Let $f_m(X; \alpha) = (f_m(x_1; \alpha), \ldots, f_m(x_k; \alpha))$. From Favaro et al. (2021, Theorem 1.2), if $m \to +\infty$ then

$$\frac{1}{m^{1/\alpha}} f_m(X; \alpha) \xrightarrow{\text{w}} f(X), \tag{4}$$

with $f(X)$ being an α-Stable process with spectral measure $\Gamma_{X,\phi}$. Theorem 3.1 provides an extension of Favaro et al. (2021, Theorem 1.2) to the ReLU activation function, which is one of the most popular unbounded activation function. It is useful to discuss Theorem 3.1 with respect to the scaling $(m \log m)^{-1/\alpha}$, which is required to achieve the infinitely wide α-Stable process. In particular, Theorem 3.1 shows that the use of the ReLU activation in place of a bounded activation results in a change of the scaling $m^{-1/\alpha}$ in equation 4, through the inclusion of the $(\log m)^{-1/\alpha}$ term. This is a critical difference between the α-Stable setting and Gaussian setting, as in the latter the choice of the activation function ϕ does not affect the scaling $m^{-1/2}$ required to achieve the infinitely wide Gaussian process. For $k = 1$, we refer to Bordino et al. (2022) for a detailed analysis of infinitely wide limits of α-Stable NNs with general classes of sub-linear, linear and super-linear activation functions.

4 Large-width training dynamics of α-Stable ReLU-NNs

Let $f_m(W, X; \alpha)$ be an α-Stable ReLU-NN, and let (X, Y) be the training set, such that $Y = (y_1, \ldots, y_k)$ is the (training) output, with y_j being the (training) output for the j-th input x_j. We consider

$$\tilde{f}_m(W, X; \alpha) = \frac{1}{(m \log m)^{1/\alpha}} f_m(W, X; \alpha),$$

and denote by $\tilde{f}_m(W, x_j; \alpha) = (m \log m)^{-1/\alpha} f_m(W, x_j; \alpha)$ the (model) output of x_j, for $j = 1, \ldots, k$. With the squared-error loss function $\ell(y_j, \tilde{f}_m(W, x_j; \alpha)) = 2^{-1} \sum_{1 \le j \le k} (\tilde{f}_m(W, x_j; \alpha) - y_j)^2$, a direct application of the chain rule leads to the NN's training dynamics. That is for any $t \ge 0$ we write

$$\frac{\mathrm{d}\tilde{f}_m(W(t), X; \alpha)}{\mathrm{d}t} = -(\tilde{f}_m(W(t), X; \alpha) - Y) \eta_m H_m(W(t), X), \tag{5}$$

where the kernel $H_m(W(t), X)$ in the NN's training dynamics is a $k \times k$ matrix whose (j, j') entry is

$$H_m(W(t), X)[j, j'] = \left\langle \frac{\partial \tilde{f}_m(W(t), x_j; \alpha)}{\partial W}, \frac{\partial \tilde{f}_m(W(t), x_{j'}; \alpha)}{\partial W} \right\rangle, \tag{6}$$

and η_m is the (continuous) learning rate. We show that if $\eta_m = (\log m)^{2/\alpha}$ then: i) the rescaled kernel at initialization $\tilde{H}_m(W(0), X) = \eta_m H_m(W(0), X)$ converges in distribution to an $(\alpha/2)$-Stable (almost surely) positive definite random kernel $\tilde{H}^*(X, X; \alpha)$, as $m \to \infty$; ii) during training $t > 0$, for every $\delta > 0$ the least eigenvalue of the kernel $\tilde{H}_m(W(t), X)$ remains bounded away from zero, for m sufficiently large, with probability $1 - \delta$; iii) for every $\delta > 0$ the gradient descent achieves zero training loss at a linear rate, with probability $1 - \delta$. Denote by $\lambda_{\min}(\cdot)$, $\|\cdot\|_F$ and $\|\cdot\|_2$ the minimum eigenvalue, the Frobenius and operator norms of symmetric and positive semi-definite matrices.

4.1 Infinitely wide limits of $\tilde{H}_m(W(0), X)$

For α-Stable ReLU-NNs, we study the large-width behaviour of the kernel $H_m(W(0), X)$ in equation 6. In particular, if

$$\tilde{H}_m(W, X) = (\log m)^{2/\alpha} H_m(W, X), \tag{7}$$

then $\tilde{H}_m(W(0), X)$ converges in distribution, as $m \to \infty$, to a positive definite random matrix $\tilde{H}^*(X, X, \alpha)$, with $(\alpha/2)$-stable distribution. In particular, this result allows to prove that the minimum eigenvalue of $\tilde{H}_m(W(0), X)$ is bounded away from zero, with arbitrarily high probability, for m sufficiently large. Critical for these results is the fact that $\tilde{H}_m(W, X)$ can be decomposed as follows:

$$\tilde{H}_m(W, X) = \tilde{H}_m^{(1)}(W, X) + \tilde{H}_m^{(2)}(W, X), \tag{8}$$

with $\tilde{H}_m^{(1)}(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$ being two matrices whose (j, j') entries are of the following form:

$$\tilde{H}_m^{(1)}(W, X)[j, j'] = \frac{1}{m^{2/\alpha}} \sum_{i=1}^{m} w_i^2 \langle x_j, x_{j'} \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) I(\langle w_i^{(0)}, x_{j'} \rangle > 0), \tag{9}$$

and

$$\tilde{H}_m^{(2)}(W, X)[j, j'] = \frac{1}{m^{2/\alpha}} \sum_{i=1}^{m} \langle w_i^{(0)}, x_j \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) \langle w_i^{(0)}, x_{j'} \rangle I(\langle w_i^{(0)}, x_{j'} \rangle > 0), \tag{10}$$

respectively. The next theorem characterizes the infinitely wide limits of the random matrices $\tilde{H}_m^{(1)}(W(0), X)$, $\tilde{H}_m^{(2)}(W(0), X)$, and $\tilde{H}_m(W(0), X)$, and provides expressions for their spectral measures.

Theorem 4.1. *Let $\tilde{H}_m(W, X)$, $\tilde{H}_m^{(1)}(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$ be the matrices defined in equation 7, equation 9, and equation 10, respectively. Moreover, for every $u \in \{0, 1\}^k$, let*

$$B_u = \{v \in \mathbb{R}^d : \langle v, x_j \rangle > 0 \text{ if } u_j = 1, \langle v, x_j \rangle \le 0 \text{ if } u_j = 0, j = 1, \ldots, k\},$$

and for every $i = 1, \ldots, d$, let e_i denote the d-dimensional vector satisfying

$$e_{ij} = 1 \text{ for } j = i, \quad e_{ij} = 0 \text{ for } j \neq i.$$

As $m \to +\infty$,

$$(\tilde{H}_m^{(1)}(W(0), X), \tilde{H}_m^{(2)}(W(0), X)) \xrightarrow{w} (\tilde{H}_1^*(\alpha), \tilde{H}_2^*(\alpha)),$$

where $\tilde{H}_1^*(\alpha)$ and $\tilde{H}_2^*(\alpha)$ are stochastically independent, positive semi-definite random matrices, distributed as $(\alpha/2)$-Stable distributions with spectral measures

$$\Gamma_1^* = C_{\alpha/2} \sum_{u \in \{0,1\}^k} \mathbb{P}(w_i^{(0)}(0) \in B_u) \frac{\delta\left(\frac{[\langle x_j, x_{j'} \rangle u_j u_{j'}]_{j,j'}}{(\sum_{j,j'} \langle x_j, x_{j'} \rangle^2 u_j u_{j'})^{1/2}}\right)}{\left(\sum_{j,j'} \langle x_j, x_{j'} \rangle^2 u_j u_{j'}\right)^{-\alpha/4}}, \tag{11}$$

and

$$\Gamma_2^* = C_{\alpha/2} \sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} \frac{\delta\left(\frac{[x_{ji} u_j x_{j'i} u_{j'}]_{j,j'}}{\sum_j x_{ji}^2 u_j}\right)}{\left(\sum_j x_{ji}^2 u_j\right)^{-\alpha/2}}, \tag{12}$$

respectively, where $C_{\alpha/2}$ is a constant defined in equation 2. Furthermore, as $m \to \infty$,

$$\tilde{H}_m(W(0), X) \xrightarrow{w} \tilde{H}^*(X, X; \alpha),$$

where $\tilde{H}^*(X, X; \alpha)$ is a positive semi-definite random matrix, distributed according to an $(\alpha/2)$-Stable distribution with spectral measure of the form $\Gamma^* = \Gamma_1^* + \Gamma_2^*$.

See Appendix A.2 for the proof of Theorem 4.1. It turns out that the probability distributions of the random matrices $\tilde{H}_1^*(\alpha)$ and $\tilde{H}_2^*(\alpha)$ are absolutely continuous in suitable subspaces of the space of symmetric and positive semi-definite matrices. In turn, this fact implies that the minimum eigenvalues of $\tilde{H}_m^{(1)}(W(0), X)$ and of $\tilde{H}_m^{(2)}(W(0), X)$ are bounded away from zero, uniformly in m, for m sufficiently large, with arbitrarily high probability. This is precisely the statement of the next theorem.

Theorem 4.2. *Under the assumptions of Theorem 4.1, for every $\delta > 0$ there exist strictly positive numbers λ_0, λ_1 and λ_2 such that, for m sufficiently large,*

$$\lambda_{min}(\tilde{H}_m^{(i)}(W(0), X)) > \lambda_i \qquad i = 1, 2,$$

and

$$\lambda_{min}(\tilde{H}_m(W(0), X)) > \lambda_0.$$

with probability at least $1 - \delta$.

See Appendix A.3 for the proof of Theorem 4.2. In the setting of Gaussian ReLU-NN, Du et al. (2019) showed that if $\eta_m = 1$, then: i) the kernel $H_m(W(0), X)$ converges in probability, as $m \to +\infty$, to the NTK $H^*(X, X)$, which is a deterministic kernel; ii) the least eigenvalue of $H^*(X, X)$ is bounded from below by a positive constant λ_0. See also the works of Jacot et al. (2018), Arora et al. (2019) and Lee et al. (2019), and references therein, for the study of large-width training dynamics of Gaussian ReLU NNs, as well as the work of Yang (2019) for some extensions to more general NN's architectures. Theorem 4.1 and Theorem 4.2 extend the main results of Du et al. (2019) to the α-Stable setting, for $\alpha \in (0, 2)$, showing that: i) the rescaled kernel $(\log m)^{2/\alpha} H_m(W(0), X)$ converges in distribution, as $m \to +\infty$, to the α-Stable NTK $\tilde{H}^*(X, X; \alpha)$, which is $(\alpha/2)$-Stable (almost surely) positive definite random kernel; ii) during training $t > 0$, for every $\delta > 0$ the least eigenvalue of the kernel $\tilde{H}_m(W(t), X; \alpha)$ remains bounded away from zero, for m sufficiently large, with probability $1 - \delta$. The randomness of the α-Stable NTK provides a critical difference between the α-Stable setting and the Gaussian setting, showing that in the α-Stable setting the randomness of the NN at initialization does not vanish in the large-width regime of the training for the NN's.

4.2 Large-width training dynamics of α-Stable ReLU-NNs

We conclude our analysis by exploiting Theorem 4.1 and Theorem 4.2 to study the large-width training dynamics of α-Stable NNs. In particular, the next theorem shows that, if m is sufficiently large, then with high probability the minumum eigenvalue of the random matrix $\tilde{H}_m(W(t), X)$ remains bounded away from zero. This property is critical in order to establish the rate of convergence of the training.

Theorem 4.3. *For any $k \geq 1$ let the collection of NN's inputs $x_1, \ldots, x_k$ be linearly independent, and such that $\|x_j\| = 1$. Let $\gamma \in (0, 1)$ and $c > 0$ be fixed numbers. Further, let $\tilde{H}_m(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$ be the random matrices defined as in equation 7 and equation 10, respectively. For every $\delta > 0$ the following properties hold true for every $t \geq 0$, with probability at least $1 - \delta$, for m sufficiently large:*

(i) for every $j = 1, \ldots, k$,

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W(t), x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F^2 < c m^{-2\gamma/\alpha};$$

(ii) there exists $\lambda_0 > 0$ such that

$$\|\tilde{H}_m^{(2)}(W(t), X) - \tilde{H}_m^{(2)}(W(0), X)\|_F < \lambda_0 m^{-\gamma/\alpha}$$

and

$$\lambda_{min}(\tilde{H}_m(W(t), X)) > \frac{\lambda_0}{2}.$$

See Appendix A.4 for the proof of Theorem 4.3. Theorem 4.3 is critical to study on the large-width training dynamics of α-Stable ReLU-NNs. From Theorem 4.3, for a fixed $\delta > 0$, let m and $\lambda_0 > 0$ be such that

$$\lambda_{\min}(\tilde{H}_m(W(s), X)) > \frac{\lambda_0}{2}.$$

for every $s \leq t$, on a set $N \in \mathcal{F}$ with $\mathbb{P}[N] > 1 - \delta$. According to such a construction, for any random initialization $W(0)(\omega)$ of the α-Stable ReLU-NN , with $\omega \in N$, the following inequality holds

$$\frac{\mathrm{d}}{\mathrm{d}s}\|Y - \tilde{f}_m(W(s)(\omega), X; \alpha)\|_2^2 \leq -\lambda_0 \|Y - \tilde{f}_m(W(s)(\omega), X; \alpha)\|_2^2,$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}s}\exp(\lambda_0 s)\|Y - \tilde{f}_m(W(s)(\omega), X; \alpha)\|_2^2 \leq 0.$$

Since $\exp(\lambda_0 s)\|Y - \tilde{f}_m(W(s)(\omega), X; \alpha)\|_2^2$ is a decreasing function of $s > 0$, then we can write that

$$\|Y - \tilde{f}_m(W(s)(\omega), X; \alpha)\|_2^2 \leq \exp(-\lambda_0 s)\|Y - \tilde{f}_m(W(0)(\omega), X; \alpha)\|_2^2.$$

The next theorem summarizes the main result of this section, completing our study on the training dynamics.

Theorem 4.4. *For any $k \geq 1$ let the collection of NN's inputs $x_1, \ldots, x_k$ be linearly independent, and such that $\|x_j\| = 1$. Under the dynamics equation 5, if $\eta_m = (\log m)^{2/\alpha}$ then for every $\delta > 0$ there exists $\lambda_0 > 0$ such that, for m sufficiently large and any $t > 0$, with probability at least $1 - \delta$ it holds true that*

$$\|Y - \tilde{f}_m(W(t), X; \alpha)\|_2^2 \leq \exp(-\lambda_0 t)\|Y - \tilde{f}_m(W(0), X; \alpha)\|_2^2.$$

5 Discussion

In this paper, we investigated large-width properties of α-Stable ReLU-NNs, focusing on the popular questions of large-width asymptotics and training dynamics of the NN. With regards to large-width asymptotics, we showed that, as the NN's width goes to infinity, a rescaled α-Stable ReLU-NN converges weakly to an α-Stable process. As a novelty with respect to the Gaussian setting, it turns out that in the α-Stable setting the choice of the activation function affects the scaling of the NN, that is: to achieve the infinitely wide α-Stable process, the ReLU activation requires an additional logarithmic term in the scaling with respect to sub-linear activations. With regards to large-width training dynamics, we characterized the infinitely wide dynamics in terms of the α-Stable NTK, and we showed that, for a sufficiently large width, the gradient descent achieves zero training error at a linear rate. The randomness of the α-Stable NTK is a further novelty with respect to the Gaussian setting, that is: within the α-Stable setting, the randomness of the NN at initialization does not vanish in the large-width regime of the training. Our work extends the main result of Favaro et al. (2020; 2021) to the popular ReLU activation function, and then presents the first analysis of the large-width training dynamics of NNs in the α-Stable setting, thus generalizing to heavy-tails distributions the main result of Du et al. (2019), as well as some results of Jacot et al. (2018) and Arora et al. (2019). The use of the α-Stable distributions to initialize NNs, in place of the classical Gaussian distributions, brought some interesting phenomena, paving the way to future research.

It remains open to establish a large-width equivalence between training an α-Stable ReLU-NN and performing a kernel regression with the α-Stable NTK. In particular, Jacot et al. (2018) showed that for Gaussian NNs, during training $t > 0$, if m is sufficiently large then the fluctuations of the squared Frobenious norm $\|H_m(W(t), X) - H_m(W(0), X)\|_F^2$ are vanishing. Accordingly, this result suggested to replace $\eta_m H_m(W(t), X)$ with the NTK $H^*(X, X)$ in the dynamics equation 1, and write

$$\frac{\mathrm{d}f^*(t, X)}{\mathrm{d}t} = -(f^*(t, X) - Y)H^*(X, X).$$

This is precisely the dynamics of a kernel regression under gradient flow, for which at $t \to +\infty$ the prediction for a generic test point $x \in \mathbb{R}^d$ is of the form $f^*(x) = YH^*(X, X)^{-1}H^*(X, x)^T$. In particular, Arora et al. (2019) have proved that the prediction of the Gaussian NN $\tilde{f}_m(W(t), x)$ at $t \to +\infty$, for m sufficiently large, is equivalent to the kernel regression prediction $f^*(x)$. Within the α-Stable setting, it is not clear whether the fluctuations of $\tilde{H}_m(W(t), X) = \tilde{H}_m^{(1)}(W(t), X) + \tilde{H}_m^{(2)}(W(t), X)$ during the training vanish, as $m \to \infty$.

Theorem 4.3 shows that the fluctuations of $\tilde{H}_m^{(2)}(W(t), X)$ vanish, as $m \to \infty$. Such a result is based on the fact that for every $\delta > 0$ it holds that

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F^2 < cm^{-2\gamma/\alpha},$$

for every $j = 1, \ldots, k$, and for every W such that $\|W - W(0)\|_F \leq (\log m)^{2/\alpha}$, with probability at least $1 - \delta$, if m is sufficiently large. See Lemma A.8 for details. The same property is not true if the partial derivatives with respect to w are replaced by the partial derivatives with respect to $w^{(0)}$. Therefore, it is not clear whether the fluctuations of $\tilde{H}_m^{(1)}(W(t), X)$ during training also vanish, as $m \to \infty$.

Another interesting avenue for future research would be to extend our results of Section 3 and Section 4 to the more general setting of deep α-Stable NNs, with $D \geq 2$ being the depth. Let us consider the following setting: i) for $d, k \geq 1$ let X be the $d \times k$ NN's input, with $x_j = (x_{j1}, \ldots, x_{jd})^T$ being the j-th input (column vector); ii) for $D, m \geq 1$ and $n \geq 1$ let: i) $(W^{(1)}, \ldots, W^{(D)})$ be the NN's weights such that $W^{(1)} = (w_{1,1}^{(1)}, \ldots, w_{m,d}^{(1)})$ and $W^{(l)} = (w_{1,1}^{(l)}, \ldots, w_{m,m}^{(1)})$ for $2 \leq l \leq D$, where the $w_{i,j}^{(l)}$'s are i.i.d. as an α-Stable distribution with scale $\sigma > 0$, e.g. we can assume $\sigma = 1$. Then,

$$f_i^{(1)}(X; \alpha) = \sum_{j=1}^{d} w_{i,j}^{(1)} x_j$$

and

$$f_{i,m}^{(l)}(X; \alpha) = \sum_{j=1}^{m} w_{i,j}^{(l)} f_j^{(l-1)}(X, m) I(f_j^{(l-1)}(X, m) > 0)$$

with $f_{i,m}^{(1)}(X; \alpha) := f_i^{(1)}(X; \alpha)$, is a deep α-Stable ReLU-NN of depth D and width m. Under the assumption that the NN's width grows sequentially over the NN's layers, i.e. $m \to +\infty$ one layer at a time, it is easy to extend Theorem 3.1 to $f_{i,m}^{(l)}(X; \alpha)$. Under the same assumption on the growth of m, we expect the NTK analysis of deep α-Stable ReLU-NNs to follow along lines similar to that we have developed for the α-Stable ReLU-NN, though computations may be more involved. A more challenging task would to extend our results to deep α-Stable ReLU-NNs under the assumptions that the NN's width grows jointly over the NN's layers, i.e. $m \to +\infty$ simultaneously over the layers.

References

Arora, S., Du, S.S., Hu, W., Li, Z., Salakhutdinov, R.R. and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*.

Basteri, A. and Trevisan, D. (2022). Quantitative Gaussian approximation of randomly initialized deep neural networks *Preprint arXiv:2203.07379*.

Bordino, A., Favaro, S. and Fortini (2022). Infinite-wide limits for Stable deep neural networks: sublinear, linear and super-linear activation functions. *Transactions of Machine Learning Research* to appear.

Bracale, D., Favaro, S., Fortini and Peluchetti, S. (2021). Large-width functional asymptotics for deep Gaussian neural networks. In *International Conference on Learning representations*.

Der, R. and Lee, D. (2006). Beyond Gaussian processes: on the distributions of infinite networks. In *Advances in Neural Information Processing Systems*.

Simon S. Du, S.S., Zhai, X., and Poczos, B. and Singh, A. (2019). Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *International Conference on Learning Representations*.

Eldan, R., Mikulincer, D. and Schramm, T. (2021). Non-asymptotic approximations of neural networks by Gaussian processes. In *Conference on Learning Theory*.

Favaro, S., Fortini, S. and Peluchetti, S. (2020). Stable behaviour of infinitely wide deep neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Favaro, S., Fortini, S. and Peluchetti, S. (2020). Deep Stable neural networks: large-width asymptotics and convergence rates. *Bernoulli*, to appear.

Fortuin, V., Garriga-Alonso, A., Wenzel, F., Ratsch, G, Turner, R.E., van der Wilk, M. and Aitchison, L. (2020). Bayesian neural network priors revisited. In *Advances in Neural Information Processing Systems*.

Garriga-Alonso, A., Rasmussen, C.E. and Aitchison, L. (2018). Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representation*.

Jacot, A., Gabriel, F, and Hongler, C. (2018). Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Klukowski, A. (2021).Rate of convergence of polynomial networks to Gaussian processes *Preprint arXiv: 2111.03175*.

Lee, H., Ayed, F., Jung, P., Lee, J., Yang, H., Caron, F. (2022). Deep neural networks with dependent weights: Gaussian process mixture limit, heavy tails, sparsity and compressibility. *Preprint arXiv:2205.08187*.

Lee, J., Sohldickstein, J., Pennington, J., Novak, R., Schoenholz, S. and Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representation*.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*.

Matthews, A.G., Rowland, M., Hron, J., Turner, R.E. and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

Neal, R.M. (1996). *Bayesian learning for neural networks*. Springer.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Hron, J., Abolafia, D., Pennington, J. and Sohldickstein, J. (2018). Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representation*.

Samoradnitsky, G. and Taqqu, M.S (1994). *Stable non-Gaussian random processes: stochastic models with infinite variance*. Chapman and Hall/CRC.

Uchaikin, V.V. and Zolotarev, V.M. (2011). *Chance and stability: stable distributions and their applications*. Walter de Gruyter.

Williams, C.K. (1997). Computing with infinite networks. In *Advances in Neural Information Processing Systems*.

Yang, G. (2019). Tensor programs II: neural tangent kernel for any architecture. *Preprint arXiv:2006.14548*.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *Preprint: arXiv:1902.04760*.

Yang, G. (2019). Tensor programs I: wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In *Advances in Neural Information Processing Systems*.

Yang, G. and Hu, E.J. (2021). Tensor programs IV: feature learning in infinite-width neural networks. In *International Conference on Machine Learning*.

Yang, G. and Littwin, E. (2021). Tensor programs IIb: architectural universality of neural tangent kernel training dynamics. In *International Conference on Machine Learning*.

A

A.1 Proof of Theorem 3.1

To simplify the notation, we set in this section: $w := w(0)$, $w^{(0)} := w^{(0)}(0)$, and $W := W(0)$. First, we will prove that $[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j$ belongs to the domain of attraction of an α-stable law with spectral measure

$$\Gamma_1 = C_\alpha \mathbb{E}_{u\sim\Gamma_0}\left(\|[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j\|^\alpha \delta\left(\frac{[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j}{\|[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j\|}\right)\right),$$

where Γ_0 is the spectral measure of $w_i^{(0)}$. For this, it is sufficient to show that

$$\begin{aligned} &r^\alpha \mathbb{P}\left(\frac{[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j}{\|[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j\|} \in B, \|[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j\| > r\right) \\ &\quad \to C_\alpha \Gamma_1(B), \end{aligned}$$

for every Borel set B of $\mathbb{S}^{k-1}$ such that $\Gamma_1(\partial B) = 0$ (see Appendix B). Let $T : \mathbb{S}^{k-1} \mapsto [0,1]^k$ and $C : \mathbb{R}^k \setminus \{0\} \to \mathbb{S}^{k-1}$ be defined as $T(u) = [\langle u, x_j\rangle I(\langle u, x_j\rangle > 0]_j$ and $C(v) = v/\|v\|$, respectively. Fix a Borel set B of $\mathbb{S}^{k-1}$ such that $\Gamma_1(\partial B) = 0$. This condition implies that

$$\begin{aligned} &\Gamma_0\left(\left\{u \in \mathbb{S}^{k-1} : \|T(u)\| \neq 0, T(u) \in C^{-1}(\partial B)\right\}\right) \\ &\quad = \Gamma_0\left(\left\{u \in \mathbb{S}^{k-1} : \|T(u)\| \neq 0, \frac{T(u)}{\|T(u)\|} \in \partial B\right\}\right) = 0. \end{aligned}$$

Hence

$$\begin{aligned} &\Gamma_0\left(T^{-1}\left(\left\{z \in [0,1]^k : \|z\| \neq 0, z \in \partial C^{-1}(B)\right\}\right)\right) \\ &\quad = \Gamma_0\left(T^{-1}\left(\left\{z \in [0,1]^k : \|z\| \neq 0, z \in C^{-1}(\partial B)\right\}\right)\right) = 0. \end{aligned}$$

Now, let $Z = T(w_i^{(0)}/\|w_i^{(0)}\|) I(\|w_i^{(0)}\| \neq 0)$. We can write that

$$\begin{aligned} &r^\alpha \mathbb{P}\left(\frac{[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j}{\|[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j\|} \in B, \|[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j\| > r\right) \\ &\quad = r^\alpha \mathbb{P}\left(\|Z\| \neq 0, \frac{Z}{\|Z\|} \in B, \|w_i^{(0)}\|\|Z\| > r\right) \\ &\quad = \int_{C^{-1}(B)\cap[0,1]^k} r^\alpha \mathbb{P}(\|w_i^{(0)}\| > r\|z\|^{-1}, Z \in dz) \\ &\quad = \int_{C^{-1}(B)\cap[0,1]^k} \|z\|^\alpha (r\|z\|^{-1})^\alpha \mathbb{P}(\|w_i^{(0)}\| > r\|z\|^{-1}, \frac{w_i^{(0)}}{\|w_i^{(0)}\|} \in T^{-1}(dz)). \end{aligned}$$

Since $\Gamma_0\left(T^{-1}\left(\left\{z \in [0,1]^k : z \neq 0, z \in \partial(C^{-1}(B))\right\}\right)\right) = 0$, then the points of discontinuity of the function $\|z\|^\alpha I(C^{-1}(B))(z)$ have zero $\Gamma_0(T^{-1}(\cdot))$-measure. It follows that

$$\begin{aligned} &\int_{C^{-1}(B)\cap[0,1]^k} \|z\|^\alpha (r\|z\|^{-1})^\alpha \mathbb{P}(\|w_i^{(0)}\| > r\|z\|^{-1}, w_i^{(0)} \in T^{-1}(dz)) \\ &\quad \to C_\alpha \int_{C^{-1}(B)\cap[0,1]^k} \|z\|^\alpha \Gamma_0(T^{-1}(dz)) \\ &\quad = C_\alpha \int_{\mathbb{S}^{k-1}} I(u \in B)\left(\frac{T(u)}{\|T(u)\|}\right) \|T(u)\|^\alpha \Gamma_0(du) \\ &\quad = C_\alpha \Gamma_1(B), \end{aligned}$$

as $r \to \infty$, which completes the proof that $[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j$ belongs to the domain of attraction of an α-stable law with spectral measure Γ_1. Then, for every k-dimensional vector s,

$$\frac{1}{m^{1/\alpha}} \sum_{i=1}^{m} \sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0),$$

as a sequence of random variables in m, converges in distribution, as $m \to +\infty$, to a random variable with α-stable distribution and characteristic function

$$\exp\Big(-|t|^\alpha \mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j \langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^\alpha\big)\Big).$$

Thus, the distribution of $\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)$ belongs to the domain of attraction of an α-stable law. In particular, this implies that as $m \to +\infty$

$$\begin{aligned} r^\alpha \mathbb{P}\Big(|\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)| > r\Big) \\ \to C_\alpha \mathbb{E}_{u\sim\Gamma_0}\Big(|\sum_{j=1}^{k} s_j \langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^\alpha\Big). \end{aligned}$$

We now study the tail behaviour of $\big|w_i \sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)\big|$. By Cline (1986, Section 5),

$$\mathbb{P}\Big(|w_i| \ |\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)| > e^t\Big) = \overline{F * G}(t),$$

where

$$\overline{F}(t) = \mathbb{P}\Big(|w_i| > e^t\Big), \qquad \overline{G}(t) = \mathbb{P}\Big(|\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)| > e^t\Big).$$

We now prove that F and G satisfy the assumptions of Cline (1986, Theorem 4) with $\beta = \gamma = 0$. The distribution functions F and G have exponential tails with rate α. Indeed, for all real u,

$$\lim_{t\to\infty} \frac{\overline{F}(t-u)}{\overline{F}(t)} = \lim_{t\to\infty} \frac{\mathbb{P}(|w_i| > e^{t-u})}{\mathbb{P}(|w_i| > e^t)} = \frac{e^{-\alpha(t-u)}}{e^{-\alpha t}} = e^{\alpha u}.$$

Analogously for G. Moreover the functions $b(t) = e^{\alpha t}\overline{F}(t)$ and $c(t) = e^{\alpha t}\overline{G}(t)$ are regularly varying with exponent zero: for all $y > 0$,

$$\lim_{t\to\infty} \frac{b(yt)}{b(t)} = \lim_{t\to\infty} \frac{e^{\alpha y t}\mathbb{P}(|w_i| > e^{yt})}{e^{\alpha t}\mathbb{P}(|w_i| > e^t)} = \lim_{t\to\infty} \frac{e^{\alpha y t} e^{-\alpha y t}}{e^{\alpha t} e^{-\alpha t}} = 1 = y^0.$$

The same property holds for $c(t)$. By Cline (1986, Theorem 4 (v)), as $t \to \infty$,

$$\begin{aligned} \mathbb{P}\Big(|w_i| \ |\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)| > e^t\Big) = \overline{F * G}(t) \\ \sim C_\alpha^2 \mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j \langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^\alpha\big)\alpha t e^{-\alpha t}, \end{aligned}$$

as $t \to \infty$. Thus, for $r \to \infty$,

$$r^{\alpha}\mathbb{P}\Big(|w_i| \ |\sum_{j=1}^{k} s_j \langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)| > r\Big)$$

$$\sim C_{\alpha}^2 \mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^{\alpha}\big)\alpha \log r.$$

Let $\tilde{L}(r) = C_{\alpha}^2\mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^{\alpha}\big)\alpha\log r$. Since the distribution of $w_i\sum_{j=1}^{k} s_j\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)$ is symmetric, then we can write that

$$\frac{1}{a_m}\sum_{i=1}^{m} w_i \sum_{j=1}^{k} s_j\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0),$$

as a sequence of random variables in m, converges in distribution, as $m \to +\infty$, to a random variable with symmetric α-stable law with scale 1 provided $(a_m)_{m\geq 1}$ satisfies

$$\frac{m\tilde{L}(a_m)}{a_m^{\alpha}} \to C_{\alpha}$$

as $m \to \infty$. The condition is satisfied if

$$a_m = \left(C_{\alpha}\mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^{\alpha}\big) m\log m\right)^{1/\alpha}.$$

It follows that

$$\frac{1}{(m\log m)^{1/\alpha}}\sum_{i=1}^{m} w_i\sum_{j=1}^{k} s_j\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0),$$

as a sequence of random variables in m, converges in distribution, as $m \to +\infty$, to a random variable with symmetric α-stable distribution with scale of the form

$$\left(C_{\alpha}\mathbb{E}_{u\sim\Gamma_0}\big(|\sum_{j=1}^{k} s_j\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)|^{\alpha}\big)\right)^{1/\alpha}.$$

Since this holds for every vector s, then

$$\frac{1}{(m\log m)^{1/\alpha}}\sum_{i=1}^{m} w_i[\langle w_i^{(0)}, x_j\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)]_j,$$

as a sequence of random variables in m, converges in distribution, as $m \to +\infty$, to a random vector with symmetric α-stable law with the spectral measure

$$\Gamma_X = \frac{1}{2}C_{\alpha}\mathbb{E}_{u\sim\Gamma_0}\Bigg(\|[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j\|^{\alpha}$$
$$\delta\Big(\frac{[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j}{\|[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j\|}\Big) + \delta\Big(-\frac{[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j}{\|[\langle u, x_j\rangle I(\langle u, x_j\rangle > 0)]_j\|}\Big)\Bigg).$$

Since $\Gamma_0 = \frac{1}{2}\sum_{i=1}^{d}(\delta(e_i) + \delta(-e_i))$, where $e_{ij} = 1$ if $j = i$ and 0 otherwise, then

$$\Gamma_X = \frac{C_{\alpha}}{4}\sum_{i=1}^{d}\Bigg(\|[x_{ji}I(x_{ji} > 0)]_j\|^{\alpha}\left(\delta\big(\frac{[x_{ji}I(x_{ji} > 0)]_j}{\|[x_{ji}I(x_{ji} > 0)]_j\|}\big) + \delta\big(-\frac{[x_{ji}I(x_{ji} > 0)]_j}{\|[x_{ji}I(x_{ji} > 0)]_j\|}\big)\right)$$
$$+\|[x_{ji}I(x_{ji} < 0)]_j\|^{\alpha}\left(\delta\big(\frac{[x_{ji}I(x_{ji} < 0)]_j}{\|[x_{ji}I(x_{ji} < 0)]_j\|}\big) + \delta\big(-\frac{[x_{ji}I(x_{ji} < 0)]_j}{\|[x_{ji}I(x_{ji} < 0)]_j\|}\big)\right)\Bigg).$$

A.2 Proof of Theorem 4.1

To simplify the notation, we set in this section: $w := w(0)$, $w^{(0)} := w^{(0)}(0)$, $W := W(0)$, $\tilde{H}_m^{(1)} := \tilde{H}_m^{(1)}(W(0), X)$ and $\tilde{H}_m^{(2)} := \tilde{H}_m^{(2)}(W(0), X)$, with $\tilde{H}_m^{(1)}(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$ defined in equation 9 and equation 10. The proof of Theorem 4.1 is split into several steps.

Lemma A.1. *If $m \to +\infty$ then*

$$\tilde{H}_m^{(1)} \xrightarrow{w} \tilde{H}_1^*(\alpha),$$

where $\tilde{H}_1^*(\alpha)$ is an $(\alpha/2)$-Stable positive semi-definite random matrix with spectral measure

$$\Gamma_1^* = C_{\alpha/2} \sum_{u \in \{0,1\}^k} \mathbb{P}(w_i^{(0)} \in B_u)(\sum_{j,j'} \langle x_j, x_{j'} \rangle^2 u_j u_{j'})^{\alpha/4} \delta\left(\frac{[\langle x_j, x_{j'} \rangle u_j u_{j'}]_{j,j'}}{(\sum_{j,j'} \langle x_j, x_{j'} \rangle^2 u_j u_{j'})^{1/2}}\right),$$

where, for every $u \in \{0,1\}^k$, $B_u = \{v \in \mathbb{R}^d : \langle v, x_j \rangle > 0$ if $u_j = 1, \langle v, x_j \rangle \leq 0$ if $u_j = 0, j = 1, \ldots, k\}$, and $C_{\alpha/2}$ is the constant defined in Equation equation 2.

Proof. Since $\tilde{H}_m^{(1)}$ is symmetric, is is sufficient to show that, for every k-dimensional vector s,

$$s^T \tilde{H}_m^{(1)} s \xrightarrow{w} s^T \tilde{H}_1^*(\alpha) s.$$

We first prove that the functions defined, for $t \in (-\infty, +\infty)$, by $\overline{F}(t) = \mathbb{P}\left(w_i^2 > e^t\right)$, and

$$\begin{aligned}
\overline{G}(t) &= \mathbb{P}\left(\sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'} \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) I(\langle w_i^{(0)}, x_{j'} \rangle > 0) > e^t\right) \\
&= \mathbb{P}\left(\|\sum_{j=1}^{k} s_j x_j I(\langle w_i^{(0)}, x_j \rangle > 0)\|^2 > e^t\right)
\end{aligned}$$

satisfy the assumptions of Cline (1986, Lemma 1). Indeed, F has exponentail tails with rate $\alpha/2$, since by the properties of the stable law,

$$\lim_{t \to \infty} \frac{\overline{F}(t-u)}{\overline{F}(t)} = \lim_{t \to \infty} \frac{\mathbb{P}(|w_i| > e^{(t-u)/2})}{\mathbb{P}(|w_i| > e^{t/2})} = e^{\alpha u/2}.$$

Moreover, for any γ,

$$m_G(\gamma) = \int_0^\infty e^{\gamma u} G(du) = \mathbb{E}\big(\|\sum_{j=1}^{k} s_j x_j I(\langle w_i^{(0)}, x_j \rangle > 0)\|^\gamma\big) < \infty.$$

By Cline (1986), Section 5 and Lemma 1, as $t \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(w_i^2 \sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'} \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) I(\langle w_i^{(0)}, x_{j'} \rangle > 0) > e^t\right) \\
&= \overline{F * G}(t) \sim m_G(\alpha/2) \overline{F}(t) \\
&\sim C_{\alpha/2} (e^t)^{-\alpha/2} \mathbb{E}\Big(\big(\sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'} \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) I(\langle w_i^{(0)}, x_{j'} \rangle > 0)\big)^{\alpha/2}\Big).
\end{aligned}$$

By the properties of the stable law,

$$s^T \tilde{H}_m^{(1)} s = \frac{1}{m^{2/\alpha}} \sum_{i=1}^{m} w_i^2 \sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'} \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) I(\langle w_i^{(0)}, x_{j'} \rangle > 0)$$

converges weakly, as $m \to \infty$, to a totally skewed to the right, $\alpha/2$-stable random variable, with scale parameter $\mathbb{E}\big(\big|\sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)\big|^{\alpha/2}\big)^{2/\alpha}$. Hence, for every $t \in \mathbb{R}$, as $m \to \infty$,

$$\begin{aligned}
&\mathbb{E}\bigg(\exp(its^T \tilde{H}_m^{(1)} s)\bigg)\\
&\to \exp\bigg(-|t|^{\alpha/2}\mathbb{E}\big(\big|\sum_{j,j'=1}^{k} s_j s_{j'} \langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)\big|^{\alpha/2}\big)\big(1 - i \text{ sign } u \tan(\pi\alpha/4)\big)\bigg)\\
&= \exp\bigg(-\int_{\mathbb{S}^{k^2-1}} \big|\sum_{j,j'} t s_j s_{j'} v_{j,j'}\big|^{\alpha/2}\big(1 - i \text{ sign}\big(t\sum_{j,j'} s_j s_{j'} v_{j,j'}\big) \tan(\pi\alpha/4)\Gamma_1^*(dv)\bigg),
\end{aligned}$$

where

$$\begin{aligned}
\Gamma_1^* = C_{\alpha/2}\mathbb{E}\Bigg(&\|[\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F^{\alpha/2}\\
&\cdot \delta\bigg(\frac{[\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}}{\|[\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F}\bigg)\Bigg).
\end{aligned}$$

It follows that, as $m \to +\infty$,

$$\tilde{H}_m^{(1)} \xrightarrow{\text{w}} \tilde{H}_1^*(\alpha),$$

where $\tilde{H}_1^*(\alpha)$ is an $(\alpha/2)$-Stable random matrix with spectral measure Γ_1^* of the form

$$\Gamma_1^* \qquad = C_{\alpha/2} \sum_{u \in \{0,1\}^k} \mathbb{P}(w_i^{(0)} \in B_u)\big(\sum_{j,j'} \langle x_j, x_{j'}\rangle^2 u_j u_{j'}\big)^{\alpha/4} \delta\left(\frac{[\langle x_j, x_{j'}\rangle u_j u_{j'}]_{j,j'}}{(\sum_{j,j'} \langle x_j, x_{j'}\rangle^2 u_j u_{j'})^{1/2}}\right).$$

We will now prove that $\tilde{H}_1^*(\alpha)$ is positive semi-definite. By definition, $\tilde{H}_m^{(1)}(\omega)$ is positive semi-definite for every ω and every m. By Portmanteau Theorem, for every vector $u \in \mathbb{S}^{k-1}$,

$$\mathbb{P}\left(u^T \tilde{H}_1^*(\alpha) u \geq 0\right) \geq \limsup_m \mathbb{P}\left(u^T \tilde{H}_m^{(1)}\ u \geq 0\right) = 1.$$

Let $\mathcal{A}$ be a countable dense subset of $\mathbb{S}^{k-1}$. Then, with probability one, $a^T \tilde{H}_1^*(\alpha) a \geq 0$ for every $a \in \mathcal{A}$. By continuity, this implies that the same property holds true with probability one for every $u \in \mathbb{S}^{k-1}$, which proves that $\tilde{H}_1^*(\alpha)$ is almost surely positive semi-definite. By eventually modifying $\tilde{H}_1^*(\alpha)$ on a null set, we obtain a positive semi-definite random matrix. □

Lemma A.2. *If $m \to +\infty$ then*

$$\tilde{H}_m^{(2)} \xrightarrow{w} \tilde{H}_2^*(\alpha),$$

where $\tilde{H}_2^*(\alpha)$ is an $(\alpha/2)$-Stable positive semi-definite random matrix with spectral measure

$$\Gamma_2^* = C_{\alpha/2} \sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} \big(\sum_j x_{ji}^2 u_j\big)^{\alpha/2} \delta\left(\frac{[x_{ji} u_j x_{j'i} u_{j'}]_{j,j'}}{\sum_j x_{ji}^2 u_j}\right),$$

where $B_u = \{v \in \mathbb{R}^d : \langle v, x_j\rangle > 0$ if $u_j = 1, \langle v, x_j\rangle \leq 0$ if $u_j = 0, j = 1, \dots, k\}$, e_i is a d-dimensional vector satisfying $e_{ij} = 1$ if $j = i$, and $e_{ij} = 0$ if $j \neq i$ $(i, j = 1, \dots, d)$, and $C_{\alpha/2}$ is the constant defined in Equation equation 2.

Proof. By the properties of the multivariate stable distribution (see Appendix B), it is sufficient to show that

$$\begin{aligned}&\mathbb{P}\left(\frac{\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}}{\|\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}\|_F} \in \cdot,\right.\\
&\qquad\left.\|\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}\|_F > r\right)\\
&\qquad\sim C_{\alpha/2} r^{-\alpha/2}\Gamma_2^*(\cdot),\end{aligned}$$

as $r \to +\infty$. We can write that

$$\begin{aligned}&\mathbb{P}\left(\frac{\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}}{\|\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}\|_F} \in \cdot,\right.\\
&\qquad\left.\|\left[\langle w_1^{(0)}, x_j\rangle\langle w_1^{(0)}, x_{j'}\rangle I(\langle w_1^{(0)}, x_j\rangle > 0) I(\langle w_1^{(0)}, x_{j'}\rangle > 0)\right]_{j,j'}\|_F > r\right)\\
&= \sum_{u\in\{0,1\}^k}\mathbb{P}\left(\frac{\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}}{\|\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}\|_F} \in \cdot,\right.\\
&\qquad\left.\|\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}\|_F > r, w_1^{(0)} \in B_u\right).\end{aligned}$$

For every $u \in \{0,1\}^k$, let X_u be the $d \times k$ matrix, defined as

$$X_u = [x_{ji}u_j]_{j=1,\dots,k,i=1,\dots,d}.$$

Then we can write that

$$\begin{aligned}&\mathbb{P}\left(\frac{\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}}{\|\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}\|_F} \in \cdot,\right.\\
&\qquad\left.\|\left[\langle w_1^{(0)}, u_j x_j\rangle\langle w_1^{(0)}, u_{j'} x_{j'}\rangle\right]_{j,j'}\|_F > r, w_1^{(0)} \in B_u\right)\\
&= \mathbb{P}\left(\frac{X_u^T w_1^{(0)}(w_1^{(0)})^T X_u}{(\mathrm{tr}(X_u^T(w_1^{(0)})^T w_1^{(0)} X_u X_u^T (w_1^{(0)})^T w_1^{(0)} X_u))^{1/2}} \in \cdot,\right.\\
&\qquad\left.\mathrm{tr}(X_u^T(w_1^{(0)})^T w_1^{(0)} X_u X_u^T (w_1^{(0)})^T w_1^{(0)} X_u) > r^2, w_1^{(0)} \in B_u\right)\\
&= \mathbb{P}\left(\frac{X_u^T(w_1^{(0)})^T w_1^{(0)} X_u}{w_1^{(0)} X_u X_u^T (w_1^{(0)})^T} \in \cdot, w_1^{(0)} X_u X_u^T (w_1^{(0)})^T > r, w_1^{(0)} \in B_u\right).\end{aligned}$$

Notice that the maximum eigenvalue of the matrix $X_u X_u^T$ is smaller than or equal to k, since the norm of each column of X_u is smaller than or equal to one. Then $w_1^{(0)} X_u X_u^T (w_1^{(0)})^T > r$ implies that $\|w_1^{(0)}\| > (r/k)^{1/2}$. We can therefore write that

$$\begin{aligned}&\mathbb{P}\left(\frac{X_u^T(w_1^{(0)})^T w_1^{(0)} X_u}{w_1^{(0)} X_u X_u^T (w_1^{(0)})^T} \in \cdot, w_1^{(0)} X_u X_u^T (w_1^{(0)})^T > r, w_1^{(0)} \in B_u\right)\\
&= \mathbb{P}\left(\frac{X_u^T(w_1^{(0)})^T w_1^{(0)} X_u}{w_1^{(0)} X_u X_u^T (w_1^{(0)})^T} \in \cdot, w_1^{(0)} X_u X_u^T (w_1^{(0)})^T > r, \|w_1^{(0)}\| > (r/k)^{1/2}, w_1^{(0)} \in B_u\right).\end{aligned}$$

Since B_u is a cone and the spectral measure of $w_1^{(0)}$ is given by $\sum_i(\delta(e_i)+\delta(-e_i))$, by the properties of the multivariate stable distribution, we can write that

$$\mathbb{P}\left(\frac{X_u^T(w_1^{(0)})^T w_1^{(0)} X_u}{w_1^{(0)} X_u X_u^T (w_1^{(0)})^T} \in \cdot, w_1^{(0)} X_u X_u^T (w_1^{(0)})^T > r, \|w_1^{(0)}\| > (r/k)^{1/2}, w_1^{(0)} \in B_u\right)$$
$$\sim C_{\alpha/2} r^{-\alpha/2} \sum_{\{i:\{e_1,-e_i\}\cap B_u \neq \emptyset\}} (\sum_{j=1}^k x_{ji}^2 u_j)^{\alpha/2} \delta\left(\frac{[x_{ji}x_{j'i}u_j u_{j'}]_{j,j'}}{\sum_j x_{ji}^2 u_j}\right),$$

as $r \to +\infty$. The proof that $\tilde{H}_2^*(\alpha)$ is positive semi-definite can be done by following the same line of reasoning as in the proof of Lemma A.1. $\square$

Lemma A.3. *As $m \to +\infty$, the probability distribution of $(\tilde{H}_m^{(1)}, \tilde{H}_m^{(1)})$ converges weakly to the law of independent stable random matrices, with spectral measures Γ_1^* and Γ_2^* as in equation 11 and equation 12, respectively.*

Proof. Since $\tilde{H}_m^{(1)}$ and $H_m^{(2)}$ converge marginally to $\alpha/2$-stable random matrices, by the properties of the multivariate stable distributions it is sufficient to show that they converge to stochastically independent random matrices. By Theorem B.1, we know that

$$n\mathbb{P}\Bigg(\|[w_i^2\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha},$$
$$\|[\langle x_j, w_i^{(0)}\rangle\langle x_{j'}, w_i^{(0)}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha}\Bigg)$$

and

$$n\mathbb{P}\Bigg(\|[w_i^2\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha}\Bigg)$$

converge to finite limits, as $n \to \infty$. Hence, again by Theorem B.1, it is sufficient to show that

$$\lim_{n\to\infty} n\mathbb{P}\Bigg(\|[w_i^2\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha},$$
$$\|[\langle x_j, w_i^{(0)}\rangle\langle x_{j'}, w_i^{(0)}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha}\Bigg) = 0,$$

which ensures that the Lévy measure of the limit infinitely divisible distribution of $(\tilde{H}_m^{(1)}, \tilde{H}_m^{(2)})$ is the sum of a measure ν_1 concentrated on the space spanned by the first k^2 coordinates and a measure ν_2 on the space spanned by the last k^2 coordinates. We can write that

$$
\begin{aligned}
& n\mathbb{P}\Bigg(\|[w_i^2 \langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha}, \\
& \qquad \|[\langle x_j, w_i^{(0)}\rangle \langle x_{j'}, w_i^{(0)}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0) I(\langle w_i^{(0)}, x_{j'}\rangle > 0)]_{j,j'}\|_F > n^{2/\alpha} \Bigg) \\
& = n \sum_{u \in \{0,1\}^k} \mathbb{P}(w_i^{(0)} \in B_u) \\
& \mathbb{P}\Bigg(\|[w_i^2 \langle x_j, x_{j'}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha}, \|[\langle x_j, w_i^{(0)}\rangle \langle x_{j'}, w_i^{(0)}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha} \mid w_i^{(0)} \in B_u \Bigg) \\
& = n \sum_{u \in \{0,1\}^k} \mathbb{P}(w_i^{(0)} \in B_u) \mathbb{P}\Bigg(\|[\langle x_j, w_i^{(0)}\rangle \langle x_{j'}, w_i^{(0)}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha} \mid w_i^{(0)} \in B_u \Bigg) \\
& \qquad \mathbb{P}\Bigg(\|[w_i^2 \langle x_j, x_{j'}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha} \Bigg) \\
& = \sum_{u \in \{0,1\}^k} n\mathbb{P}\Bigg(\|[\langle x_j, w_i^{(0)}\rangle \langle x_{j'}, w_i^{(0)}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha}, w_i^{(0)} \in B_u \Bigg) \\
& \qquad \mathbb{P}\Bigg(\|[w_i^2 \langle x_j, x_{j'}\rangle u_j u_{j'}]_{j,j'}\|_F > n^{2/\alpha} \Bigg) \to 0,
\end{aligned}
$$

as $n \to \infty$. □

Now, we are in the position of proving Theorem 4.1. By Lemma A.1, Lemma A.1, Lemma A.3, and the properties of stable distributions, $\tilde{H}_m(W(0), X)$ converges in distribution to a positive semi-definite random matrix, with $(\alpha/2)$-stable distribution, and spectral measure $\Gamma_1^* + \Gamma_2^*$. This completes the proof of Theorem 4.1.

A.3 Proof of Theorem 4.2

To simplify the notation, we set in this section: $w := w(0)$, $w^{(0)} := w^{(0)}(0)$, $W := W(0)$, $\tilde{H}_m^{(1)} := \tilde{H}_m^{(1)}(W(0), X)$ and $\tilde{H}_m^{(2)} := \tilde{H}_m^{(2)}(W(0), X)$, with $\tilde{H}_m^{(1)}(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$ defined in equation 9 and equation 10.

From equation 8, $\tilde{H}_m(W(0), X))$ is the sum of two positive semi-definite random matrices, $\tilde{H}_m^{(1)}$ and $\tilde{H}_m^{(2)}$. The following results show that for every $\delta > 0$, there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that, for m sufficiently large, with probability at least $1 - \delta$

$$
\lambda_{\min}(\tilde{H}_m^{(i)}) > \lambda_i.
$$

with the large-width behaviour of $\tilde{H}_m^{(i)}$ being characterized in Lemma A.1 and Lemma A.2, through an $(\alpha/2)$-Stable limiting random matrix $\tilde{H}_i^*(\alpha)$ with spectral measure Γ_i^* of the form equation 11 and equation 12. To prove that the minumum eigenvales of $\tilde{H}_m^{(1)}$ and $\tilde{H}_m^{(2)}$ are bounded away from zero, we first need to inspect the characteristics of the distributions of $\tilde{H}_1^*(\alpha)$ and of $\tilde{H}_2^*(\alpha)$. This is the content of Lemma A.4 and of Lemma A.6. Then, the results concerning the minumum eigenvalues of $\tilde{H}_m^{(1)}$ and $\tilde{H}_m^{(2)}$ are given in Lemma A.5 and Lemma A.7.

Lemma A.4. *Under the assumptions of Theorem 4.4, the distribution of the random matrix $\tilde{H}_1^*(\alpha)$ is absolutely continuous in the subspace of the symmetric positive semi-definite matrices with zero entries in the positions (j, j') such that $\langle x_j, x_{j'}\rangle = 0$, with $j, j' \in \{1, \ldots, k\}$, with the topology of Frobenius norm.*

Proof. From Nolan (2010), it is sufficient to show that

$$
\inf_{s \in \mathbb{S}_0^{k^2-1}} \int |\langle s, u\rangle|^{\alpha/2} \Gamma_1^*(\mathrm{d}u) \neq 0,
$$

where Γ_1^* is the spectral measure equation 11, $\mathbb{S}_0^{k^2-1}$ is the unit sphere in the space of the $k \times k$ symmetric matrices such that $s_{j,j'} = 0$ if $\langle x_j, x_{j'}\rangle = 0$, with the Frobenius metric. Now, since

$$\int |\langle s, u\rangle|^{\alpha/2}\Gamma_1^*(\mathrm{d}u)$$
$$= C_{\alpha/2}\mathbb{E}\left(|\sum_{j,j'} s_{j,j'}\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)I(\langle w_i^{(0)}, x_{j'}\rangle > 0)|^{\alpha/2}\right)$$

is a continuous function of s that takes value in a compact set, then the minimum is attained. Thus it is sufficient to show that for every $s \in \mathbb{S}_0^{k^2-1}$,

$$\mathbb{E}\left(|\sum_{j,j'} s_{j,j'}\langle x_j, x_{j'}\rangle I\langle w_i^{(0)}, x_j\rangle > 0)I(\langle w_i^{(0)}, x_{j'}\rangle > 0)|^{\alpha/2}\right) \neq 0.$$

For every j and every $u_j \in \{0,1\}$, let $A_j^{u_j}$ be the event $(\langle w_i^{(0)}, x_j\rangle > 0)$ if $u_j = 1$ and its complement if $u_j = 0$. Then

$$\mathbb{E}\left(|\sum_{j,j'} s_{j,j'}\langle x_j, x_{j'}\rangle I(\langle w_i^{(0)}, x_j\rangle > 0)I(\langle w_i^{(0)}, x_{j'}\rangle > 0)|^{\alpha/2}\right)$$
$$= \sum_{u_1,\dots,u_k} \mathbb{P}(A_1^{u_1} \cap \dots \cap A_k^{u_k})|\sum_{j,j'} u_j u_{j'} s_{j,j'}\langle x_j, x_{j'}\rangle|^{\alpha/2}.$$

Since $x_1, \dots, x_k$ are linearly independent, then for every $u_1, \dots, u_k$, $\mathbb{P}(A_1^{u_1} \cap \dots, A_k^{u_k}) > 0$. To prove it, assume, without loss of generality, that $u_i = 1$ for every i. Since $x_1, \dots, x_k$ are linearly independent, then we can complete the matrix $X = [x_1\ \dots\ x_k]$ by adding $k-d$ columns in such a way that the completed matrix $\tilde{X}$ is non-singular. For every d-dimensional vector v such that $v_1 > 0, \dots, v_k > 0$ there exists a vector u such that $u = (\tilde{X}^T)^{-1}v$. Thus,

$$\{u \in \mathbb{R}^d : \langle u, x_1\rangle > 0, \dots, \langle u, x_k\rangle > 0\} = \{(\tilde{X}^T)^{-1}v : v_1 > 0, \dots, v_k > 0\}$$

is an open non-empty set. Since $w_i^{(0)}$ has independent and identically distributed components, with stable distribution, then

$$\mathbb{P}\left(w_i^{(0)} \in \{(\tilde{X})^{-1}v : v_1 > 0, \dots, v_k > 0\}\right) > 0.$$

This concludes the proof that $\mathbb{P}(A_1^{u_1} \cap \dots, A_k^{u_k}) > 0$ for every $(u_1, \dots, u_k) \in \{0,1\}^k\}$. It follows that $\int |\langle s, u\rangle|^{\alpha/2}\Gamma_1^*(du)$ is zero if and only if, for every $(u_1, \dots, u_k) \in \{0,1\}^k$, it holds

$$\sum_{j,j'} u_j, u_{j'}\langle x_j, x_{j'}\rangle s_{j,j'} = 0.$$

The only solution of the above system of equations in the space of symmetric matrices s such that $s_{j,j'} = 0$ if $\langle x_j, x_{j'}\rangle = 0$ is $s = 0$, which is not consistent with $\|s\|_F = 1$. □

We observe that the space of the symmetric positive semi-definite matrices with zeros in the entries (j, j') such that $\langle x_j, x_{j'}\rangle = 0$ contains all the matrices with non-zero diagonal element since $\langle x_j, x_j\rangle = 1 \neq 0$ for every index j.

Lemma A.5. *Under the assumptions of Theorem 4.4, for every $\delta > 0$ there exists $\lambda_1 > 0$ such that with probability at least $1 - \delta$*

$$\lambda_{min}(\tilde{H}_1^*(\alpha)) > \lambda_1.$$

Proof. Since the distribution of $\tilde{H}_1^*(\alpha)$ is absolutely continuous in the space of symmetric positive semi-definite matrices with zero entries in the positions j, j' such that $\langle x_, x_{j'} \rangle = 0$, and since this space contains all the symmetric positive semi-definite matrices with non-zero diagonal entries, then we can write that $\mathbb{P}(\det(\tilde{H}_1^*(\alpha)) = 0) = 0$. Moreover, since $\tilde{H}_1^*(\alpha)$ is positive semi-definite, then $\mathbb{P}(\lambda_{\min}(\tilde{H}_1^*(\alpha)) > 0) = 1$. Thus, for every $\delta > 0$, the exists $\lambda_1 > 0$ such that $\mathbb{P}(\lambda_{\min}(\tilde{H}_1^*(\alpha)) > \lambda_1) > 1 - \delta$. $\square$

Lemma A.6. *Under the assumptions of Theorem 4.4, the distribution of the random matrix $\tilde{H}_2^*(\alpha)$ is absolutely continuous in the subspace of the symmetric positive semi-definite matrices, with the topology of Frobenius norm.*

Proof. From Nolan (2010), it is sufficient to show that

$$\inf_{s \in \mathbb{S}^{k^2-1}} \int |\langle s, u \rangle|^{\alpha/2} \Gamma_2^*(\mathrm{d}u) \neq 0,$$

where Γ_2^* is the spectral measure equation 12, $\mathbb{S}^{k^2-1}$ is the unit sphere in the space of the $k \times k$ symmetric positive semi-definite matrices, with the Frobenius norm. For every $u \in \{0,1\}^k$, let $B_u = \{v \in \mathbb{R}^d : \langle v, x_j \rangle > 0 \text{ if } u_j = 1, \langle v, x_j \rangle \leq 0 \text{ if } u_j = 0\}$. Moreover, for every $i = 1, \dots, k$, let e_i be a d-dimensional random vector satisfying $e_{ij} = 1$ for $j = i$ and $e_{ij} = 0$ for $j \neq i$. Finally, let $C_{\alpha/2}$ be the constant defined in Equation equation 2. Then

$$\int |\langle s, u \rangle|^{\alpha/2} \Gamma_2^*(\mathrm{d}u) = C_{\alpha/2} |\sum_{j,j'} s_{j,j'} \sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} x_{ji} u_j x_{j'i} u_{j'}|^{\alpha/2}.$$

Since $\sum_{j,j'} s_{j,j'} \sum_{u \in \mathcal{U}} \sum_E z_{u,i} x_{ji} u_j x_{j'i} u_{j'}$ is continuous as a function of s and s takes values in a compact set, then the minimum is attained. Thus it is sufficient to show that for every $s \in \mathbb{S}^{k^2-1}$,

$$\sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} \sum_{j,j'} s_{j,j'} x_{ji} u_j x_{j'i} u_{j'} \neq 0.$$

Since $\|s\|_F = 1$, then s is not the null matrix. Hence there exist $c > 0$, a vector a with $\|a\| = 1$ and a positive semi-definite, symmetric matrix s' such that

$$s = caa^T + s'.$$

Since $B_u \cap B_{u'} = \emptyset$, when $u \neq u'$, then, for every $i = 1, \dots, d$ and $j = 1, \dots, k$, there exists one and only one $u \in \{0,1\}^k$ such that $u_j = 1$ and $\{e_i, -e_i\} \cap B_u \neq \emptyset$. Then we can write that

$$\begin{aligned}
&\sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} \sum_{j,j'} s_{j,j'} x_{ji} u_j x_{j'i} u_{j'} \\
&\geq c \sum_{u \in \{0,1\}^k} \sum_{\{i : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} (\sum_j a_j x_{ji} u_j)^2 \\
&= \sum_{i=1}^d \left((\sum_{j=1}^k a_j x_{ji})^2 \sum_{\{u : \{e_i, -e_i\} \cap B_u \neq \emptyset\}} u_j \right) \\
&= \sum_{i=1}^d (\sum_{j=1}^k a_j x_{ji})^2,
\end{aligned}$$

which is strictly positive, since the x_j are linearly independent, and $\|a\| = 1$. This concludes the proof. $\square$

Lemma A.7. *Under the assumptions of Theorem 4.4, for every $\delta > 0$ there exists $\lambda_2 > 0$ such that with probability at least $1 - \delta$*

$$\lambda_{min}(\tilde{H}_2^*(\alpha)) > \lambda_2.$$

Proof. Since the distribution of $\tilde{H}_2^*(\alpha)$ is absolutely continuous in the space of symmetric positive semi-definite matrices then we can write that $\mathbb{P}(\det(\tilde{H}_2^*(\alpha)) = 0) = 0$. Moreover, since $\tilde{H}_2^*(\alpha)$ is positive semi-definite, then $\mathbb{P}(\lambda_{\min}(\tilde{H}_2^*(\alpha)) > 0) = 1$. Thus, for every $\delta > 0$, the exists $\lambda_2 > 0$ such that $\mathbb{P}(\lambda_{\min}(\tilde{H}_2^*(\alpha)) > \lambda_2) > 1 - \delta$. □

Now, we are in the position of proving Theorem 4.2. Let $\delta > 0$ be a fixed number. By Lemmas A.5 and A.7, there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that, for $i = 1, 2$, $\mathbb{P}(\lambda_{\min}(\tilde{H}_i^*(\alpha)) > \lambda_i) \geq 1 - \delta/2$. Since the minimum eigenvalue map is continuous with respect to Frobenius norm then, by Portmanteau theorem, for $i = 1, 2$,

$$\liminf_m \mathbb{P}(\lambda_{\min}(\tilde{H}_m^{(i)}(W(0), X)) > \lambda_i) \geq \mathbb{P}(\lambda_{\min}(\tilde{H}_i^*(\alpha)) > \lambda_i) \geq 1 - \delta/2.$$

Let $\lambda_0 = \lambda_1 + \lambda_2$. Since the minimum eigenvalue of a sum of symmetric, positive semi-definite matrices is greater than or equal to the sum of the eigenvalues of the two matrices (see Horn and Johnson (1985) Theorem 4.3.1), then we can write that

$$\begin{aligned}
&\liminf_m \mathbb{P}(\lambda_{\min}(\tilde{H}_m(W(0), X)) > \lambda_0) \\
&\quad \geq \liminf_m \mathbb{P}(\lambda_{\min}(\tilde{H}_m^{(1)}(W(0), X)) + \lambda_{\min}(\tilde{H}_m^{(2)}(W(0), X)) > \lambda_0) \\
&\quad \geq \liminf_m \mathbb{P}(\cap_{i=1,2}(\lambda_{\min}(\tilde{H}_m^{(i)}(W(0), X)) > \lambda_i)) \\
&\quad \geq 1 - \limsup_m \left(\sum_{i=1}^{2} \mathbb{P}(\lambda_{\min}(\tilde{H}_m^{(i)}(W(0), X)) \leq \lambda_i) \right) \\
&\quad \geq 1 - \delta,
\end{aligned}$$

thus completing the proof of Theorem 4.2.

A.4 Proof of Theorem 4.3

Before proving Theorem 4.3, we give some preliminary results.

Lemma A.8. *Let $\gamma \in (0, 1)$ and $c > 0$ be fixed numbers. For every $\delta > 0$ the following property holds true, for m sufficiently large, with probability at least $1 - \delta$:*

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F^2 < cm^{-2\gamma/\alpha},$$

for every W such that $||W - W(0)||_F \leq (\log m)^{2/\alpha}$ and every NN's input x_j, with $j = 1, \ldots, k$.

Proof. For a fixed $W(0)$, let W be such that $\|W - W(0)\|_F \leq (\log m)^{2/\alpha}$. Then it holds $\|w^{(0)} - w^{(0)}(0)\|_F^2 \leq \|W - W(0)\|_F^2 \leq (\log m)^{4/\alpha}$. Accordingly, we can write the following

$$\begin{aligned}
&(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F^2 \\
&\leq \frac{1}{m^{2/\alpha}} \sum_{i=1}^{m} \left(\langle w_i^{(0)}, x_j \rangle I(\langle w_i^{(0)}, x_j \rangle > 0) - \langle w_i^{(0)}(0), x_j \rangle I(\langle w_i^{(0)}(0), x_j \rangle > 0) \right)^2 \\
&\leq \frac{2}{m^{2/\alpha}} \sum_{i=1}^{m} \left(\langle w_i^{(0)}, x_j \rangle - \langle w_i^{(0)}(0), x_j \rangle \right)^2 I(\langle w_i^{(0)}, x_j \rangle > 0) \\
&\quad + \frac{2}{m^{2/\alpha}} \sum_{i=1}^{m} \langle w_i^{(0)}(0), x_j \rangle^2 \left(I(\langle w_i^{(0)}, x_j \rangle > 0) - I(\langle w_i^{(0)}(0), x_j \rangle > 0) \right)^2.
\end{aligned}$$

We will bound the two terms of the sum separately. First, we define $r_i = |\langle w_i^{(0)} - w_i^{(0)}(0), x_j \rangle|$ for $i = 1, \dots, m$. Then, we can write that

$$\sum_{i=1}^m r_i^2 \le \sum_{i=1}^m \|w_i^{(0)} - w_i^{(0)}(0)\|^2 \cdot \|x_j\|^2 \le \|w^{(0)} - w^{(0)}(0)\|_F^2 \le (\log m)^{4/\alpha}.$$

Since $\gamma < 1$,

$$\begin{aligned} &\frac{2}{m^{2/\alpha}} \sum_{i=1}^m \left(\langle w_i^{(0)}, x_j \rangle - \langle w_i^{(0)}(0), x_j \rangle \right)^2 I(\langle w_i^{(0)}, x_j \rangle > 0) \\ &\quad \le 2m^{-2/\alpha} (\log m)^{4/\alpha} < \frac{c}{4} m^{-2\gamma/\alpha}, \end{aligned}$$

for m sufficiently large. In order to bound the second term, we observe that the following set

$$\{w^{(0)}(0) : \exists w^{(0)} s.t. |\langle w_i^{(0)} - w_i^{(0)}(0), x_j \rangle| = r_i,\ I(\langle w^{(0)}, x_j \rangle > 0) \neq I(\langle w^{(0)}(0), x_j \rangle > 0)\}$$

is included in the set $\{w_i^{(0)}(0) : |\langle w_i^{(0)}(0), x_j \rangle| \le r_i\}$. Therefore, we can write that

$$\begin{aligned} &\sup_{\sum_i r_i^2 \le \log m} \sup_{|w_i^{(0)} - w_i^{(0)}(0)| \le r_i} \frac{2}{m^{2/\alpha}} \sum_{i=1}^m \langle w_i^{(0)}(0), x_j \rangle^2 \left(I(\langle w_i^{(0)}, x_j \rangle > 0) - I(\langle w_i^{(0)}(0), x_j \rangle > 0) \right)^2 \\ &\le \sup_{\sum_i r_i^2 \le \log m} \sup_{|w_i^{(0)} - w_i^{(0)}(0)| \le r_i} \frac{2}{m^{2/\alpha}} \sum_{i=1}^m \langle w_i^{(0)}(0), x_j \rangle^2 I(\langle w_i^{(0)}(0), x_j \rangle < r_i) \\ &\le \sup_{\sum_i r_i^2 \le \log m} \sup_{|w_i^{(0)} - w_i^{(0)}(0)| \le r_i} \frac{2}{m^{2/\alpha}} \sum_{i=1}^m r_i^2 \\ &\le \frac{1}{m^{2/\alpha}} (\log m)^{4/\alpha} < \frac{c}{4} m^{-2\gamma/\alpha}, \end{aligned}$$

for m sufficiently large. ◻

Lemma A.9. *For every $\delta > 0$ there exist $\lambda > 0$ such that the following two properties hold true, for m sufficiently large, with a probability at least $1 - \delta$:*

i)

$$\|\tilde{H}_m^{(2)}(W, X) - \tilde{H}_m^{(2)}(W(0), X)\|_F < \lambda m^{-\gamma/\alpha};$$

ii)

$$\lambda_{min}(\tilde{H}_m(W, X)) > \frac{\lambda}{2};$$

for every W such that $\|W - W(0)\|_F \le (\log m)^{2/\alpha}$.

Proof. By Lemma A.7, for every $\delta > 0$ there exists λ such that

$$\lambda_{\min}(\tilde{H}_2^*(\alpha)) > \lambda$$

with probability at least $1-\delta/2$. For every vector W, we can write that

$$
\begin{aligned}
&|\tilde{H}_m^{(2)}(W,X)[i,j]-\tilde{H}_m^{(2)}(W(0),X)[i,j]| \\
&\quad= (\log m)^{2/\alpha}\left|\left\langle \frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha), \frac{\partial \tilde{f}_m}{\partial w}(W,x_j;\alpha)\right\rangle - \left\langle \frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha), \frac{\partial \tilde{f}_m}{\partial w}(W(0),x_j;\alpha)\right\rangle\right| \\
&\quad\le (\log m)^{2/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)\right\|_F\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_j;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_j;\alpha)\right\|_F \\
&\qquad+ (\log m)^{2/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_j;\alpha)\right\|_F\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F \\
&\quad\le (\log m)^{2/\alpha}\left(\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F+\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)\right\|_F\right) \\
&\qquad\times\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_j;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_j;\alpha)\right\|_F \\
&\qquad+ (\log m)^{2/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_j;\alpha)\right\|_F\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F.
\end{aligned}
$$

For every $i=1,\dots,k$,

$$
\begin{aligned}
(\log m)^{2/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F^2 &= \frac{1}{m^{2/\alpha}}\sum_{i=1}^m \langle w_i^{(0)}(0),x_i\rangle^2 I(|\langle w_i^{(0)}(0),x_i\rangle|>0) \\
&\le \frac{1}{m^{2/\alpha}}\sum_{i=1}^m \langle w_i^{(0)}(0),x_i\rangle^2,
\end{aligned}
$$

which converges in distribution, as $m\to\infty$. Thus there exist $M>0$ and m_0 such that for every $m\ge m_0$ and every $i=1,\dots,k$,

$$
\mathbb{P}\left((\log m)^{1/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F>M\right)<\frac{\delta}{8k^2}.
$$

By Lemma A.8, for m sufficiently large, with probability at least $1-\delta/(4k^2)$

$$
(\log m)^{1/\alpha}\left(\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F+\left\|\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)\right\|_F\right)<2M
$$

whenever $\|W-W(0)\|_F<(\log m)^{2/\alpha}$. Lemma A.8 also implies that, for every $\gamma\in(0,1)$, and $i=1,\dots,k$, with probability at least $1-\delta/(8k^2)$

$$
(\log m)^{1/\alpha}\left\|\frac{\partial \tilde{f}_m}{\partial w}(W,x_i;\alpha)-\frac{\partial \tilde{f}_m}{\partial w}(W(0),x_i;\alpha)\right\|_F<\frac{\lambda}{4Mk^2}m^{-\gamma/\alpha}
$$

whenever $\|W-W(0)\|_F^2<(\log m)^{4/\alpha}$, provided m is sufficiently large,. Thus, with probability at least $1-\delta$, if m is sufficiently large

$$
\max_{i,j}|\tilde{H}_m^{(2)}(W,X)[i,j]-\tilde{H}_m^{(2)}(W(0),X)[i,j]|<\frac{\lambda}{k^2}m^{-\gamma/\alpha},
$$

whenever $\|W-W(0)\|_F<(\log m)^{2/\alpha}$. Thus

$$
\begin{aligned}
&\|\tilde{H}_m^{(2)}(W,X)-\tilde{H}_m^{(2)}(W(0),X)\|_2 \\
&\quad\le\|\tilde{H}_m^{(2)}(W,X)-\tilde{H}_m^{(2)}(W(0),X)\|_F<\lambda m^{-\gamma/\alpha}<\frac{\lambda}{2},
\end{aligned}
$$

whenever $\|W-W(0)\|_F<(\log m)^{2/\alpha}$, provided m is sufficiently large. The last inequality and Lemma A.6 imply that, with probability at least $1-\delta$, if m is sufficiently large, then

$$
\|\tilde{H}_m^{(2)}(W,X)\|_2>\lambda/2,
$$

for every W such that $\|W - W(0)\|_F < (\log m)^{2/\alpha}$. Since $\tilde{H}_m(W, X)$ is the sum of two positive semi-definite matrices $\tilde{H}_m^{(1)}(W, X)$ and $\tilde{H}_m^{(2)}(W, X)$, then

$$\|\tilde{H}_m(W, X)\|_2 \geq \|\tilde{H}_m^{(2)}(W, X)\|_2 > \lambda/2,$$

for every W such that $\|W - W(0)\|_F < (\log m)^{2/\alpha}$, if m is sufficiently large. $\square$

Lemma A.10. *For every $\delta > 0$ the following property holds true, for m sufficiently large, with probabillity at least $1 - \delta$: there exists $M > 0$ such that*

$$(\log m)^{1/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0), x_j; \alpha) \right\|_F < M,$$

for every $j = 1, \dots, k$, and for every W such that $\|W - W(0)\|_F \leq (\log m)^{2/\alpha}$.

Proof. Let us define $r_i = |\langle w_i^{(0)} - w_i^{(0)}(0), x_j \rangle|$ for $i = 1, \dots, m$. Now, since $\|x_j\| = 1$ by assumption, for $j = 1, \dots, k$, then we can write

$$\sum_i r_i^2 \leq \|x_j\|^2 \cdot \|w_i^{(0)} - w^{(0)}(0)\|_F^2 \leq \|W - W(0)\|_F^2 \leq (\log m)^{4/\alpha}.$$

It holds

$$\begin{aligned}
&(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0), x_j; \alpha) \right\|_F^2 \\
&\leq \frac{1}{m^{2/\alpha}} \sum_{i=1}^m \Big(w_i I(\langle w_i^{(0)}, x_j \rangle > 0) - w_i(0) I(\langle w_i^{(0)}(0), x_j \rangle > 0) \Big)^2 \\
&\leq \frac{2}{m^{2/\alpha}} \sum_{i=1}^m (w_i - w_i(0))^2 I(\langle w_i^{(0)}, x_j \rangle > 0) \\
&\quad + \frac{2}{m^{2/\alpha}} \sum_{i=1}^m w_i(0)^2 |I(\langle w_i^{(0)}, x_j \rangle > 0) - I(\langle w_i^{(0)}(0), x_j \rangle > 0)|.
\end{aligned}$$

We will bound the two terms separately. First,

$$\begin{aligned}
&\frac{2}{m^{2/\alpha}} \sum_{i=1}^m (w_i - w_i(0))^2 I(\langle w_i^{(0)}, x_j \rangle > 0) \\
&\leq \frac{1}{m^{2/\alpha}} \sum_{i=1}^m (w_i - w_i(0))^2 \\
&\leq \frac{2}{m^{2(1-\gamma)/\alpha}} \|w - w(0)\|_F^2 \\
&\leq \frac{2}{m^{2/\alpha}} (\log m)^{4/\alpha} < \frac{c}{4} m^{-2\gamma/\alpha},
\end{aligned}$$

if m is sufficiently large. To bound the second term, we can write that

$$\begin{aligned}
&\frac{2}{m^{2/\alpha}} \sum_{i=1}^m w_i(0)^2 |I(\langle w_i^{(0)}, x_j \rangle > 0) - I(\langle w_i^{(0)}(0), x_j \rangle > 0)| \\
&\leq \frac{2}{m^{2/\alpha}} \sum_{i=1}^m w_i(0)^2,
\end{aligned}$$

which converges in distribution to a stable random variable, as $m \to \infty$. Hence there exists M_1 such that, with probability at least $1 - \delta/4$,

$$\frac{2}{m^{2/\alpha}} \sum_{i=1}^m (w_i - w_i(0))^2 I(\langle w_i^{(0)}, x_j \rangle > 0) < \frac{M_1^2}{2k^2}$$

and

$$\frac{2}{m^{2/\alpha}} \sum_{i=1}^{m} w_i(0)^2 |I(\langle w_i^{(0)}, x_j \rangle > 0) - I(\langle w_i^{(0)}(0), x_j \rangle > 0)| < \frac{M_1^2}{2k^2},$$

for m sufficiently large, which entail

$$(\log m)^{1/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0)(\omega), x_j; \alpha) \right\|_F < \frac{M_1}{k}.$$

On the other hand, there exist $N_3 \in \mathcal{F}$ and M_2 with $P(N_3) > 1 - \delta/4$ such that, for every $\omega \in N_3$ and for m sufficiently large,

$$\|\tilde{f}_m(W(0)(\omega), X; \alpha) - Y\|_F < M_2,$$

and

$$\max_{1 \leq i \leq k} \left\| \frac{\partial}{\partial W} \tilde{f}_m(W(0)(\omega), x_i; \alpha) \right\|_F < M_2 (\log m)^{-1/\alpha}.$$

The above inequalities follow from the convergence in distribution of $\tilde{f}_m(W(0), x_i; \alpha)$ and of

$$(\log m)^{2/\alpha} \left\| \frac{\partial}{\partial W} \tilde{f}_m(W(0), x_i; \alpha) \right\|_F^2 = \tilde{H}(W(0), X; \alpha)[i, i] \quad (i = 1, \dots, k),$$

as $m \to \infty$. □

Lemma A.11. *Let $\gamma \in (0, 1)$ and $c > 0$ be fixed numbers. For every $\delta > 0$ the following property holds true, for m sufficiently large, with probability at least $1 - \delta$:*

$$\|W(t) - W(0)\|_F < (\log m)^{2/\alpha}.$$

if

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W(s), x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F^2 \leq c m^{-2\gamma/\alpha}$$

for every NN's input x_j, with $j = 1, \dots, k$, and for every $s \leq t$.

Proof. By Lemmas A.8 and A.9, there exists $N_1 \in \mathcal{F}$ with probability at least $1 - \delta/2$ such that, for every $\omega \in N_1$,

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0)(\omega), x_j; \alpha) \right\|_F^2 < c m^{-2\gamma/\alpha},$$

for arbitrarily fixed $c >$ and $\gamma \in (0, 1/2)$, and

$$\lambda_{\min}(\tilde{H}_m(W, X)) > \frac{\lambda}{2},$$

for some $\lambda > 0$, for every W such that $\|W - W(0)(\omega)\|_F \leq (\log m)^{2/\alpha}$ and every $j = 1, \dots, k$, provided m is sufficiently large. Moreover, by Lemma A.10, there exist, for m sufficiently large, $M_1 > 0$ and N_2 with $\mathbb{P}(N_2) > 1 - \delta$, such that

$$(\log m)^{1/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W, x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0)(\omega), x_j; \alpha) \right\|_F < \frac{M_1}{k},$$

for every $j = 1, \dots, k$, and for every W such that $\|W - W(0)(\omega)\|_F \leq (\log m)^{2/\alpha}$. We will prove, by contradiction, that for every $\omega \in N_1 \cap N_2 \cap N_3$, $\|W(t) - W(0)\|_F < (\log m)^{2/\alpha}$ for every $t > 0$. In the following we will write $W(s)$ in the place of $W(s)(\omega)$ and always assume that ω belongs to $N_1 \cap N_2 \cap N_3$. Suppose that there exists t such that $\|W(t) - W(0)\|_F \geq (\log m)^{2/\alpha}$, and let

$$t_0 = \text{argmin}_{t \geq 0} \{t : \|W(t) - W(0)\|_F \geq (\log m)^{2/\alpha}\}.$$

Since $\|W(s) - W(0)\|_F \le (\log m)^{2/\alpha}$ for every $s \le t_0$, then, for every $s \le t_0$,

$$
\begin{aligned}
&\lambda_{\min}(\tilde{H}_m(W(s), X)) > \frac{\lambda}{2},\\
&\left\| \frac{\partial \tilde{f}_m}{\partial w}(W(s), x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_j; \alpha) \right\|_F < c m^{-\gamma/\alpha} (\log m)^{-1/\alpha},\\
&\left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(s), x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0)(\omega), x_j; \alpha) \right\|_F < \frac{M_1}{k} (\log m)^{-1/\alpha} \quad (j = 1, \dots, k),\\
&\|\tilde{f}_m(W(0)(\omega), X; \alpha) - Y\|_F < M_2,\\
&\max_{1 \le i \le k} \left\| \frac{\partial}{\partial W} \tilde{f}_m(W(0)(\omega), x_i; \alpha) \right\|_F < M_2 (\log m)^{-1/\alpha}.
\end{aligned}
$$

Let us now consider the gradient descent dynamic, with continuous learning rate $\eta = (\log m)^{2/\alpha}$:

$$
\begin{aligned}
\frac{\mathrm{d}W(s)}{\mathrm{d}s} &= -(\log m)^{2/\alpha} \nabla_W \frac{1}{2} \sum_{i=1}^k \left(\tilde{f}_m(W(s), x_i; \alpha) - y_i \right)^2\\
&= -(\log m)^{2/\alpha} \sum_{i=1}^k \left(\tilde{f}_m(W(s), x_i) - y_i \right) \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_i; \alpha).
\end{aligned}
$$

This expression allows to write

$$
\begin{aligned}
&\|W(t_0) - W(0)\|_F\\
&\quad \le \left\| \int_0^{t_0} \frac{\mathrm{d}}{\mathrm{d}s} W(s) \mathrm{d}s \right\|_F\\
&\quad \le (\log m)^{2/\alpha} \left\| \int_0^{t_0} \sum_{i=1}^k (\tilde{f}_m(W(s), x_i; \alpha) - y_i) \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_i; \alpha) \mathrm{d}s \right\|_F\\
&\quad \le (\log m)^{2/\alpha} \max_{0 \le s \le t_0} \sum_{i=1}^k \left\| \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_i; \alpha) \right\|_F \int_0^{t_0} \|\tilde{f}_m(W(s), X; \alpha) - Y\| ds.
\end{aligned}
$$

To bound the term $\|\tilde{f}_m(W(s), X; \alpha) - Y\|$ we will exploit the dynamics of the NN output

$$
\begin{aligned}
\frac{\mathrm{d}\tilde{f}_m(W(s), X; \alpha)}{\mathrm{d}s} &= \frac{\partial \tilde{f}_m}{\partial W}(W(s), X; \alpha) \frac{\mathrm{d}W^T(s)}{\mathrm{d}s}\\
&= -(\log m)^{2/\alpha} (\tilde{f}_m(W(s), X; \alpha) - Y) H_m(W(s), X)\\
&= -(\tilde{f}_m(W(s), X; \alpha) - Y) \tilde{H}_m(W(s), X),
\end{aligned}
$$

that gives

$$
\frac{\mathrm{d}}{\mathrm{d}s} \|\tilde{f}_m(W(s), X; \alpha) - Y\|_2^2 = -2 \left(\tilde{f}_m(W(s), X; \alpha) - Y \right) \tilde{H}_m(W(s), X) \left(\tilde{f}_m(W(s), X; \alpha) - Y \right)^T.
$$

Since $\lambda_{\min}(\tilde{H}_m(W(s), X)) > \lambda/2$ for every $s \le t_0$, then

$$
\frac{\mathrm{d}}{\mathrm{d}s} \|\tilde{f}_m(W(s), X; \alpha) - Y\|_2^2 \le -\lambda \|\tilde{f}_m(W(s), X; \alpha) - Y\|_2^2,
$$

which implies that

$$
\frac{\mathrm{d}}{\mathrm{d}s} \left(\exp(\lambda s) \|\tilde{f}_m(W(s), X; \alpha) - Y\|_2^2 \right) \le 0.
$$

It follows that $\exp(\lambda s) \|\tilde{f}_m(W(s), X; \alpha) - Y\|_2^2$ is a decreasing function of s, and therefore

$$
\|\tilde{f}_m(W(s), X; \alpha) - Y\|_2 \le \exp(-\lambda/2) \|\tilde{f}_m(W(0), X; \alpha) - Y\|_2,
$$

for every $s \leq t_0$. Substituting in the integral, we can write that

$$
\begin{aligned}
&\|W(t_0) - W(0)\|_F \\
&\quad \leq (\log m)^{2/\alpha} \max_{0 \leq s \leq t_0} \sum_{i=1}^{k} \left\| \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_i; \alpha) \right\|_F \int_0^{t_0} \exp(-\lambda s/2) \mathrm{d}s \cdot \|\tilde{f}_m(W(0), X; \alpha) - Y\| \\
&\quad \leq \frac{2(\log m)^{2/\alpha}}{\lambda} \max_{0 \leq s \leq t_0} \sum_{i=1}^{k} \left(\left\| \frac{\partial \tilde{f}_m}{\partial W}(W(0), x_i; \alpha) \right\|_F + \left\| \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_i; \alpha) - \frac{\partial \tilde{f}_m}{\partial W}(W(0), x_i; \alpha) \right\|_F \right) \\
&\qquad \times \|\tilde{f}_m(W(0), X; \alpha) - Y\| \\
&\quad \leq \frac{2(\log m)^{2/\alpha}}{\lambda} \max_{0 \leq s \leq t_0} \sum_{i=1}^{k} \left(\left\| \frac{\partial \tilde{f}_m}{\partial W}(W(0), x_i; \alpha) \right\|_F + \left\| \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(s), x_i; \alpha) - \frac{\partial \tilde{f}_m}{\partial w^{(0)}}(W(0), x_i; \alpha) \right\|_F \right. \\
&\qquad \left. + \left\| \frac{\partial \tilde{f}_m}{\partial w}(W(s), x_i; \alpha) - \frac{\partial \tilde{f}_m}{\partial w}(W(0), x_i; \alpha) \right\|_F \right) \|\tilde{f}_m(W(0), X; \alpha) - Y\| \\
&\leq \quad \frac{2(\log m)^{1/\alpha}}{\lambda} \left(M_2 + M_1 + kcm^{-\gamma/\alpha} \right) M_2,
\end{aligned}
$$

which, for m large, contradicts $\|W(t_0) - W(0)\|_F \geq (\log m)^{2/\alpha}$. □

Now, we are in the position of proving Theorem 4.3. Let $m \in \mathbb{N}$ and $N \in \mathcal{F}$ be such that $\mathbb{P}(N) > 1 - \delta$ and the properties mentioned in Lemma A.8, Lemma A.9, Lemma A.10 and Lemma A.11 hold true for every $\omega \in N$. Therefore, by means of Lemma A.8 and of Lemma A.9, it is sufficient to show that

$$\|W(t) - W(0)\|_F^2(\omega) < (\log m)^{2/\alpha}$$

for every $t > 0$ and $\omega \in N$. By contradiction, suppose that there exists, for some $\omega \in N$, $t_0(\omega)$ finite with

$$t_0(\omega) := \inf_{t \geq 0} \{t : \|W(t) - W(0)\|_F(\omega) \geq (\log m)^{2/\alpha}\}.$$

Since $W(t)(\omega)$ is a continuous function of t, then $\|W(t_0(\omega)) - W(0)\|_F^2(\omega) = (\log m)^{2/\alpha}$. Then, by Lemma A.8,

$$(\log m)^{2/\alpha} \left\| \frac{\partial \tilde{f}_m}{\partial W}(W(s), x_j; \alpha) - \frac{\partial \tilde{f}_m}{\partial W}(W(0), x_j; \alpha) \right\|_F^2 (\omega) < cm^{-2\gamma/\alpha},$$

for every $s \leq t_0$ and every j. Therefore, by Lemma A.11 it holds true that $\|W(t_0(\omega)) - W(0)\|_F(\omega) < (\log m)^{2/\alpha}$, which contradicts the definition of t_0. This completes the proof of Theorem 4.3.

B

The distribution of a random vector ξ is said to be infinitely divisible if, for every n, there exist some i.i.d. random vectors $\xi_{n1}, \ldots, \xi_{nn}$ such that $\sum_k \xi_{nk} \stackrel{d}{=} \xi$. A k-dimensional random vector ξ is infinitely divisible if and only if its characteristic function admits the representation $e^{\psi(u)}$, where

$$\psi(u) = iu^T b - \frac{1}{2} u^T a u + \int \left(e^{iu^T x} - 1 - iu^T x I(||x|| \leq 1) \right) \nu(dx) \tag{13}$$

where ν is a measure on $\mathbb{R}^k \setminus \{0\}$ satisfying $\int (||x||^2 \wedge 1) \nu(dx) < \infty$, a is a $k \times k$ positive semi-definite, symmetric matrix and b is a vector. The measure ν is called the Lévy measure of ξ and (a, b, ν) are called the characteristics of the infinitely divisible distribution. We will write $\xi \sim i.d.(a, b, \nu)$. Other kinds of truncation can be used for the term $iu^T x$. This affects only the vector of centering constants b. An i.i.d. array of random vectors is a collection of random vectors $\{\xi_{nj}, j \leq m_n, n \geq 1\}$ such that, for every n,

$\xi_{n1}, \ldots, \xi_{nm_n}$ are i.i.d. The class of infinitely divisible distributions coincides with the class of limits of sums of i.i.d. arrays (Kallenberg, 2002, Theorem 13.12).

To state a general criterion of convergence, we first introduce some notations. Let $\xi \sim i.d.(a, b, \nu)$. Define, for each $h > 0$,

$$a^{(h)} = a + \int_{||x||<h} xx^T \nu(dx),$$

$$b^{(h)} = b - \int_{h<||x||\leq 1} x\nu(dx),$$

where $\int_{h<||x||\leq 1} = -\int_{1<||x||\leq h}$ if $h > 1$. Denote by $\xrightarrow{v}$ vague convergence, that is convergence of measures with respect to the topology induced by bounded, measurable functions with compact support. Moreover, let $\overline{\mathbb{R}^k}$ be the one-point compactification of $\mathbb{R}^k$. The following criterion for convergence holds (Kallenberg, 2002, Corollary 13.16).

Theorem B.1. *Consider in $\mathbb{R}^k$ an i.i.d. array $(\xi_{nj})_{j=1,\ldots,m_n,n\geq 1}$ and let ξ be i.d.(a, b, ν). Let $h > 0$ be such that $\nu(||x|| = h) = 0$. Then $\sum_j \xi_{nj} \xrightarrow{d} \xi$ if and only if the following conditions hold:*

(i) $m_n \mathbb{P}(\xi_{n1} \in \cdot) \xrightarrow{v} \nu(\cdot)$ *on* $\overline{\mathbb{R}^k} \setminus \{0\}$

(ii) $m_n \mathbb{E}(\xi_{n1}\xi_{n1}^T I(||\xi_{n1}|| < h)) \to a^{(h)}$

(iii) $m_n \mathbb{E}(\xi_{n1} I(||\xi_{n1}|| < h)) \to b^{(h)}$

Inside the class of infinitely divisible distribution, we can distinguish the subclass of stable distributions. A k-dimensional random vector ξ has stable distribution if, for every independent random vectors ξ_1 and ξ_2 with $\xi_1 \stackrel{d}{=} \xi_2 \stackrel{d}{=} \xi$ and every $a, b \in \mathbb{R}$, there exists $c \in \mathbb{R}$ and $d \in \mathbb{R}^k$ such that $a\xi_1 + b\xi_2 \stackrel{d}{=} c\xi + d$. This is equivalent to the condition: for every $n \geq 1$,

$$\xi_1 + \cdots + \xi_n \stackrel{d}{=} n^{1/\alpha}\xi + d_n \tag{14}$$

where $\alpha \in (0, 2]$, $\xi_1, \ldots, \xi_n$ are i.i.d. copies of ξ and d_n is a vector. The random vector ξ is said to be strictly stable if equation 14 holds with $d_n = 0$. A stable vector ξ is strictly stable if and only if all its components are strictly stable. The coefficient α is called the index of stability of ξ and the law of ξ is called α-stable. A stable vector ξ is symmetric stable if $\mathbb{P}(\xi \in A) = \mathbb{P}(-\xi \in A)$ for every Borel set A. A symmetric stable vector is strictly stable. The class of stable distributions coincides with the class of limit laws of sequences $((\sum_{k=1}^n X_k - b_n)/a_n)$, where (X_n) are i.i.d. random variables.

A stable distribution is infinitely divisible. Thus its characteristic function admits the Lévy representation equation 13. If $\alpha = 2$, then the Lévy measure is the null measure and, therefore, the stable distribution coincides with the multivariate normal distribution with covariance matrix a and mean vector b. If $\alpha < 2$, then $a = 0$ (the zero matrix) and the α-stability implies that there exists a measure σ on the unit sphere $\mathbb{S}^{k-1}$ such that $\nu(dx) = r^{-(\alpha+1)} dr\sigma(ds)$, where $r = ||x||$ and $s = x/||x||$. Substituting in equation 13, we obtain

$$\psi(u) = iu^T b + \int_S \int_0^\infty \left(e^{iru^T s} - 1 - iru^T sI(r \leq 1)\right) \frac{1}{r^{1+\alpha}} dr\sigma(ds)$$

For $\alpha < 1$, the centering $iru^T sI(r \leq 1)$ is not needed, since the function (of r) is integrable, and we can write

$$\psi(u) = iu^T b' + \int_S \int_0^\infty \left(e^{iru^T s} - 1\right) \frac{1}{r^{1+\alpha}} dr\sigma(ds),$$

for some vector b'. After evaluating the inner integrals as in Feller (1968, Example XVII.3), we obtain

$$\psi(u) = iu^T b' - \int_S |u^T s|^\alpha \Gamma(1 - \alpha) \left(\cos(\pi\alpha/2) - i \operatorname{sign}(u^T s) \sin(\pi\alpha/2)\right) \sigma(ds)$$

$$= iu^T b' - \int_S |u^T s|^\alpha \left(1 - i\,\mathrm{sign}(u^T s)\tan(\pi\alpha/2)\right)\Gamma(1-\alpha)\cos(\pi\alpha/2)\sigma(ds).$$

For $\alpha > 1$, using the centering $iru^T s$, we can write

$$\psi(u) = iu^T b'' + \int_S \int_0^\infty \left(e^{iru^T s} - 1 - iru^T s\right)\frac{1}{r^{1+\alpha}} dr\sigma(ds),$$

for some b''. After evaluating the inner integrals as in Feller (1968, Example XVII.3), we obtain

$$\psi(u) = iu^T b'' + \int_S |u^T s|^\alpha \frac{\Gamma(2-\alpha)}{\alpha - 1}\left(\cos(\pi\alpha/2) - i\,\mathrm{sign}(u^T s)\sin(\pi\alpha/2)\right)\sigma(ds)$$

$$= iu^T b'' - \int_S |u^T s^\alpha \left(1 - i\,\mathrm{sign}(u^T s)\tan(\pi\alpha/2)\right)\frac{\Gamma(2-\alpha)}{1-\alpha}\cos(\pi\alpha/2)\sigma(ds).$$

Since, for $\alpha < 1$, $\Gamma(2-\alpha) = (1-\alpha)\Gamma(1-\alpha)$, we can encompass the above results in one equation, and write, for $\alpha \neq 1$,

$$\psi(u) = iu^T b''' - \int_S |u^T s|^\alpha \left(1 - i\,\mathrm{sign}(u^T s)\tan(\pi\alpha/2)\right)\frac{\Gamma(2-\alpha)}{1-\alpha}\cos(\pi\alpha/2)\sigma(ds),$$

for some b'''. Finally, for $\alpha = 1$, using the centering $ir\sin ru^T s$, we can write

$$\psi(u) = iu^T b'''' + \int_S \int_0^\infty \left(e^{iru^T s} - 1 - ir\sin ru^T s\right)\frac{1}{r^2} dr\sigma(ds),$$

for some b''''. Evaluating the inner integral as in Feller (1968, Example XVII.3), we obtain

$$\psi(u) = iu^T b'''' - \int_S |u^T s|\left(\frac{\pi}{2} + i\mathrm{sign}(u^T s)\log|u^T s|\right)\sigma(ds)$$

$$= iu^T b'''' - \int_S |u^T s|\left(1 + i\frac{2}{\pi}\mathrm{sign}(u^T s)\log|u^T s|\right)\frac{\pi}{2}\sigma(ds).$$

Considering the spectral representation $e^{\psi(u)}$ of the multivariate stable characteristic function

$$\psi(u) = \begin{cases} -\int_S |u^T s|^\alpha \left(1 - i\ \mathrm{sign}(u^T s)\tan(\pi\alpha/2)\right)\Gamma(ds) + iu^T\mu^{(0)} & \alpha \neq 1 \\ \\ -\int_S |u^T s|\left(1 + i\frac{2}{\pi}\ \mathrm{sign}(u^T s)\log|u^T s|\right)\Gamma(ds) + iu^T\mu^{(0)} & \alpha = 1, \end{cases}$$

we can establish the following relationship between the Lévy measure ν and the spectral measure Γ:

$$\nu(dx) = C_\alpha \frac{1}{r^{\alpha+1}}\Gamma(ds),$$

where $r = ||x||$, $s = x/||x||$ and

$$C_\alpha = \begin{cases} \dfrac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} & \alpha \neq 1 \\ \\ 2/\pi & \alpha = 1 \end{cases}$$

A Stable random vector ξ is strictly stable if and only if

$$\begin{cases} \mu^{(0)} = 0 & \alpha \neq 1 \\ \int_S s_j \Gamma(ds) = 0 \text{ for every j} & \alpha = 1. \end{cases}$$

(see e.g. Samoradnitsky and Taqqu (1994, Theorem 2.4.1)). By Theorem B.1, the spectral measure Γ of a symmetric stable random vector ξ satisfies

$$\lim_{n\to\infty} n\mathbb{P}\left(||\xi|| > n^{1/\alpha}x, \frac{\xi}{||\xi||} \in A\right) = C_\alpha x^{-\alpha}\Gamma(A) \tag{15}$$

for every Borel set A of S such that $\Gamma(\partial A) = 0$. Moreover, the distribution of a random vector ξ belongs to the domain of attraction of the $\mathrm{St}_k(\alpha, \Gamma)$ distribution, with $\alpha \in (0,2)$ and Γ simmetric finite measure on $\mathbb{S}^{k-1}$, if and only if equation 15 holds (see e.g. Davydov et al. (2008, Theorem 4.3)).

References

Cline, D.B.H. (1986). Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields* **72**, 525–557.

Davydov, Y., Molchanov, I. and Zuyev, S. (2008). Strictly stable distributions on convex cones. *Electronic Journal of Probability* **13**, 259 – 321.

Feller, W. (1968). *An introduction to probability theory and its applications.* Wiley.

Horn, R.A. and Johnson, C.R. (1985). *Matrix Analysis.* Cambridge University Press.

Kallenberg, O. (2002). *Foundations of modern probability.* Springer.

Nolan, J.P. (2010). Metrics for multivariate stable distributions. *Banach Center Publications* **90**, 83–102.