INFINITE-WIDTH LIMIT OF DEEP LINEAR NEURAL NETWORKS

LÉNAÏC CHIZAT, MARIA COLOMBO, XAVIER FERNÁNDEZ-REAL, AND ALESSIO FIGALLI

Abstract. This paper studies the infinite-width limit of deep linear neural networks initialized with random parameters. We obtain that, when the number of parameters diverges, the training dynamics converge (in a precise sense) to the dynamics obtained from a gradient descent on an infinitely wide deterministic linear neural network. Moreover, even if the weights remain random, we get their precise law along the training dynamics, and prove a quantitative convergence result of the linear predictor in terms of the number of parameters.

We finally study the continuous-time limit obtained for infinitely wide linear neural networks and show that the linear predictors of the neural network converge at an exponential rate to the minimal ℓ_2-norm minimizer of the risk.

1. Introduction

The description of the training dynamics of (artificial) neural networks (NNs) in the infinite-width limit, has in recent years shed light on several aspects of deep learning theory, such as (i) the existence of well-posed limits, which suggests to interpret practical large scale models as approximations of those limits, (ii) the importance of the choice of scalings/parametrization[1] when passing to the limit — since several well-behaved but fundamentally different limits can be obtained, and (iii) the characterization of the long-term behavior of the dynamics — such as global convergence or algorithmic regularization — which in turn helps understanding the learning abilities of neural networks.

These aspects are rather well understood for two-layer neural networks, but the theory is lacunary for deeper NNs. A description of the infinite-width dynamics is available for the Neural Tangent (NTP) and Integrable (IP) parameterizations (discussed below), but both limits exhibit a form of degeneracy such as a lack of feature learning. In [45], the Maximal Update Parameterization (μP) — which is in a sense intermediate between (NTP) and (IP) in terms of scale — was introduced

2020 *Mathematics Subject Classification.* 68T07, 35Q49.

∗ Authors are listed in alphabetical order.

M. C. and X. F. were supported by the SNF grant 200021_182565 and by the Swiss State Secretariat for Education Research and Innovation (SERI) under contract number MB22.00034. X. F. was furthermore supported by the SNF grant PZ00P2_208930, and by the AEI project PID2021-125021NA-I00 (Spain). A. F. was supported by the European Research Council (ERC) under grant agreement No 721675 "Regularity and Stability in Partial Differential Equations (RSPDE)" and by the Lagrange Mathematics and Computation Research Center.

[1]That is, the choice, as a function of the width, of the variance of the random initialization and of the learning rates for each layer.

and shown to preserve feature learning in the limit for certain architectures, such as fully-connected NNs, which suggests that μP is a natural case of study. However, the theoretical understanding of this limit is so far very limited, because this limit involves large random matrices in an intricate way. In particular, the following fundamental questions are still open:

(i) Is the infinite-width limit of Gradient Descent (GD) a GD trajectory in some infinite-dimensional space?
(ii) Does it admit a well-posed continuous-time limit?
(iii) Does it converge to minimizers? And when several minimizers exist, can we characterize which particular solution it selects?

In this paper, we study the infinite-width limit of deep *linear*[2] NNs under μP, and we answer positively to all these questions. Throughout the paper, we focus on the three-layer case, although our tools and analysis could be extended to more layers (the main conceptual gap happens when going from two to three layers). The last section shows, without technical details, how our three-layer results read in the case of deep neural networks with an arbitrary number of layers. Our analysis of linear NN is intended as a step towards understanding the dynamics in the general non-linear case, for which the three questions above are still unresolved.

1.1. **Related work and other limits.** The first analysis of wide NNs may be traced back to [34, 7]. The dynamics of wide NNs were first studied in [28, 17, 14, 2] for the NTP (or related linear dynamics) and in [35, 33, 11, 39, 41] for non-linear dynamics in two-layer NNs under μP, which is known as *mean-field parameterization* in this case (see Remark 2.1 for a description of these various parameterizations). The importance of the choice of parameterization when passing to the limit was first highlighted in [12, 32] and systematically studied in [23, 45]. Parameterizations akin to μP were previously empirically studied in [21] as a natural extension of the two-layer mean-field parameterization and a fix to the degeneracy of IP using large initial learning rates was proposed in [26].

Our work has strong connections to [45], which shows, essentially, that all the random vectors that are generated when running a finite number of GD steps on a (non-linear) NN converge jointly in law to a family of objects characterized by an abstract algorithm. Because of the intricate dependency that arises between random matrices and random vectors, this limit "algorithm" is, unfortunately, more complex than its finite-width counterpart and hard to study beyond a few GD steps (in particular, it is non Markovian, i.e. the state of the infinite width system at time t is not enough to determine the state at time $t+1$). One of our contributions is, for the particular case of linear NNs, to exhibit a simple and theoretically tractable structure in this limit. From a technical viewpoint, [45] relies on the technique of Gaussian conditioning, which originated in the field of statistical physics to describe TAP equations [8, 9], while we use the *method of moments* which is another classical technique of random matrix theory that allows to easily obtain universality (i.e., our results apply for non-Gaussian initializations as well; note that the universality of the technique of [45] was proved recently in [24]) and rates that are quantitative in the

[2]Linear NNs are NNs without nonlinear maps between layers. Although they are linear in the input data, we note that these models are non-linear in their parameters.

width. The random matrix statements of our work (in particular Proposition 3.3) have thus their counterpart in the language of [45]; by proposing an independent proof with different techniques, our purpose is to make the analysis self-contained as well as to shed a different light on the objects appearing in the limit. See also [27] for more links between random matrix theory and NN theory.

Another closely related work is [10], which introduces a closed system of equations describing the dynamics of infinite width (non-linear) NNs. These equations can be written in continuous-time as well, thus giving an answer the question (ii). The aforementioned work, however, which relies on tools from dynamical mean-field theory, is derived at a formal level with no explicit control of the error terms. For the linear case, [10] writes a more specific system of equations that describes the same dynamics as ours, but our descriptions are of different nature: as in [45], the system derived in [10] is non Markovian. In contrast, our description is a gradient flow dynamics — thus Markovian — and all the complexity that arises from the correlated random matrices is encoded in the initial state of the dynamics. The simplicity of our limit system allows us in particular to study the large time behavior of the dynamics to answer question (iii).

Finally, there is a rich literature on the training dynamics of linear NNs. Some works show that the optimization landscape is benign [6, 19, 15] (the latter studies the NTP and thus a dynamics that becomes linear in the large width case), other works study settings where the dynamics display a "saddle to saddle" behavior [31, 29, 22, 40], and finally, a line of works studies the implicit bias of gradient descent [30, 4]; that is, which solution is chosen when the problem is underdetermined. A common assumption in this literature is that the matrices are *balanced* at initialization, that is, $W_\ell W_\ell^\top = W_{\ell+1}^\top W_{\ell+1}$ where W_ℓ and $W_{\ell+1}$ are the initial matrices in two consecutive layers of the NN. This assumption leads to simple dynamics [3], which remain tractable even in the infinite depth limit [13]. While the assumption covers the case of orthogonal initialization, it is not satisfied for standard i.i.d. initializations, where $W_\ell W_\ell^\top$ and $W_{\ell+1}^\top W_{\ell+1}$ are independent random matrices. The simplification that results from the "balancedness" assumption can be seen in our analysis from the fact that in Proposition 3.3, most of the terms in the right hand side would vanish, leading to a much simpler recursion.

Our analysis in the last section borrows ideas from [30] which shows a min-ℓ_2 implicit bias for linear NNs with the logistic loss. Note that linear NN do not always exhibit this type of implicit bias: there are subtle results for architectures that are not fully connected [25, 38, 44, 37].

1.2. **Organization of the paper.** In Section 2, we present our main results and illustrate them with numerical experiments. Section 3 studies the structure of iterated products of large random matrices with random vectors, and we show that they can be expanded in a basis of random vectors. These objects are the building blocks of the GD iterations and these results are exploited in Section 4, which contains the proof of the infinite-width limit. In Section 5 we study properties of the limit system. Finally, in Section 6 we describe the analogous results for multi-layer NNs.

2. Presentation of the main results

This section presents a rigorous discussion of linear neural networks under μP of width m, in the limit as $m \to \infty$. In the case of two-layer neural networks, the analogous problem has been qualitatively understood in [35, 33, 11, 39, 41, 43] (see also the reviews [18, 20, 5]). The first striking special feature of the two layers case is that there is a natural choice of the parametrization — which mathematically is represented by a suitable factor of m in front of the output weights — that allows the parameters to remain nondegenerate and deterministic in the limit $m \to \infty$. Under this parametrization, two-layer neural networks can be interpreted as a Wasserstein gradient flow for the weights (also in the limit), and hence the problem as $m \to \infty$ is also a solution of a Wasserstein gradient flow (and in particular it can be written as a family of parabolic equations).

For neural networks of more than two layers, several aspects of the previous analysis change. Firstly, as discussed in the introduction, it is not possible to find a natural parametrization (that is, a consistent rescaling of the three layers of weights) such that one expects them to remain nondegenerate or deterministic in the limit $m \to \infty$. In fact, as we will also see a posteriori, with the right choice of parametrization outlined in subsections 2.1 and 2.2 below, the evolution of the entries of the intermediate layer is negligible with respect to their initialization size, but these small variations change significantly the output. Due to this issue with parametrizations, it is essential in our analysis to consider randomized initial data, and to expect such random effect to survive in our limit system with some averaging effects.

Our limit system is expressed in a basis of independent, identically distributed gaussian random variables. In turn, its coefficients are obtained by solving an infinitely wide linear neural network, which in the continuous-time limit can be represented as an explicit collection of ODEs.

2.1. Setting.

Let $\tilde{h}^m$ be a single output three-layer linear neural network with input $x \in \mathbb{R}^d$, width $m \in \mathbb{N}$, and weights $\tilde{\boldsymbol{U}}^m \in \mathbb{R}^{m\times d}$, $\tilde{\boldsymbol{W}}^m \in \mathbb{R}^{m\times m}$, and $\tilde{\boldsymbol{V}}^m \in \mathbb{R}^m$:

$$y = \tilde{h}^m(x, \tilde{\boldsymbol{U}}^m, \tilde{\boldsymbol{W}}^m, \tilde{\boldsymbol{V}}^m) = \sum_{i=1}^m \tilde{\boldsymbol{V}}_i^m \sum_{j=1}^m \tilde{\boldsymbol{W}}_{ij}^m \sum_{\ell=1}^d \tilde{\boldsymbol{U}}_{j\ell}^m x_\ell = \langle \tilde{\boldsymbol{V}}^m, \tilde{\boldsymbol{W}}^m \tilde{\boldsymbol{U}}^m x \rangle.$$

Given a smooth loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we study the behavior of Gradient Descent (GD) starting from a random initialization on the expected loss $\tilde{F}$ defined as

$$\tilde{F}^m(\tilde{\boldsymbol{U}}^m, \tilde{\boldsymbol{W}}^m, \tilde{\boldsymbol{V}}^m) := \int_{\mathbb{R}^d\times\mathbb{R}} \mathcal{L}(\tilde{h}^m(x), y)\, d\rho(x, y).$$

where $\rho \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R})$ is a probability distribution that represents the input/output data. Specifically, we consider the sequence initialized as

$$\tilde{U}_{j\ell}^m(0) \sim \mathcal{N}(0, 1), \qquad \tilde{W}_{ij}^m(0) \sim \mathcal{N}\left(0, \frac{1}{m}\right), \qquad \tilde{V}_i^m(0) \sim \mathcal{N}\left(0, \frac{1}{m^2}\right), \tag{2.1}$$

and, with a step-size/learning rate τ, defined recursively as

$$\begin{cases} \tilde{\boldsymbol{U}}^m(\kappa+1) = \tilde{\boldsymbol{U}}^m(\kappa) - \tau m \int \mathcal{L}'(\tilde{h}^m_{\kappa,\tau}(x), y) \nabla_{\tilde{\boldsymbol{U}}^m} \tilde{h}^m_{\kappa,\tau}(x) d\rho_\kappa(x,y), \\ \tilde{\boldsymbol{W}}^m(\kappa+1) = \tilde{\boldsymbol{W}}^m(\kappa) - \tau \int \mathcal{L}'(\tilde{h}^m_{\kappa,\tau}(x), y) \nabla_{\tilde{\boldsymbol{W}}^m} \tilde{h}^m_{\kappa,\tau}(x) d\rho_\kappa(x,y), \\ \tilde{\boldsymbol{V}}^m(\kappa+1) = \tilde{\boldsymbol{V}}^m(\kappa) - \tau m^{-1} \int \mathcal{L}'(\tilde{h}^m_{\kappa,\tau}(x), y) \nabla_{\tilde{\boldsymbol{V}}^m} \tilde{h}^m_{\kappa,\tau}(x) d\rho_\kappa(x,y). \end{cases}$$

where for notational convenience we are denoting

$$\tilde{h}^m_{\kappa,\tau}(x) = \tilde{h}^m(x, \tilde{\boldsymbol{U}}^m(\kappa), \tilde{\boldsymbol{W}}^m(\kappa), \tilde{\boldsymbol{V}}^m(\kappa))$$
$$\nabla_\bullet \tilde{h}^m_{\kappa,\tau}(x) = (\nabla_\bullet \tilde{h}^m)(x, \tilde{\boldsymbol{U}}^m(\kappa), \tilde{\boldsymbol{W}}^m(\kappa), \tilde{\boldsymbol{V}}^m(\kappa)),$$

and $\mathcal{L}'$ denotes the derivative of the loss function with respect to the first argument. For the sake of generality, we are also considering ρ_κ depending on κ, so that (mini-batch) stochastic gradient descent (SGD) is covered by our analysis[3]. Our only assumption is that these probability measures have uniformly bounded second moments in the first variable:

$$\sup_\kappa \int |x|^2 \rho_\kappa(x,y) < +\infty. \tag{2.2}$$

The factors in red (m and m^{-1}) are layer-wise learning rates introduced so that each layer contributes equally to the variations of the predictor in the limit, as the theory will verify.

The randomness of the initialization — and in particular the large random matrix $\tilde{\boldsymbol{W}}(0)$ — play a key role in our analysis. The choice of scalings is motivated as follows:

- The scaling of $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{W}}$ is chosen so that $\tilde{\boldsymbol{U}}^m(0)x$ and $\tilde{\boldsymbol{W}}^m(0)\tilde{\boldsymbol{U}}^m(0)x$ have a nonzero variance that does not depend on m for large m (by the CLT);
- The scaling of $\tilde{\boldsymbol{V}}$ is of order $1/m$ in order to avoid the lazy training phenomenon [12], that leads to a linear dynamics described in [28].

Remark 2.1. This choice of scale for initialization is referred to as *Maximal Update Parametrization* (μP) in [45], where it is shown to lead to feature-learning for each layer[4]. In the introduction, we mentioned NTP, which corresponds to the scales (2.1) but with $\tilde{V}_i(0) \sim \mathcal{N}(0, 1/m)$; and IP which corresponds to (2.1) but with $\tilde{W}_{ij}(0) \sim N(c, 1/m^2)$ which is degenerate unless one chooses $c \neq 0$ or time-dependent learning rates [26].

Computing the gradient using the chain rule, we get the following recursion

$$\begin{cases} \tilde{\boldsymbol{U}}^m(\kappa+1) = \tilde{\boldsymbol{U}}^m(\kappa) - \tau m \tilde{\boldsymbol{W}}^m(\kappa)^\top \tilde{\boldsymbol{V}}^m(\kappa) (\tilde{\boldsymbol{\xi}}^m_\kappa)^\top, \\ \tilde{\boldsymbol{W}}^m(\kappa+1) = \tilde{\boldsymbol{W}}^m(\kappa) - \tau \tilde{\boldsymbol{V}}^m(\kappa) (\tilde{\boldsymbol{\xi}}^m_\kappa)^\top \tilde{\boldsymbol{U}}^m(\kappa)^\top, \\ \tilde{\boldsymbol{V}}^m(\kappa+1) = \tilde{\boldsymbol{V}}^m(\kappa) - \tau m^{-1} \tilde{\boldsymbol{W}}^m(\kappa) \tilde{\boldsymbol{U}}^m(\kappa) \tilde{\boldsymbol{\xi}}^m_\kappa. \end{cases}$$

[3]For instance, mini-batch SGD is obtained by defining ρ_κ as the (random) empirical distribution of samples chosen in the mini-batch at time step κ.

[4]In our context, there is no feature learning *per se* since the predictor is linear, but we will see that the dynamics remains non-linear in the parameters in the limit (in contrast to NTP).

where we have denoted $\tilde{\boldsymbol{\xi}}_\kappa^m := \int \mathcal{L}'(\tilde{h}_{\kappa,\tau}^m(x), y) x \, d\rho_\kappa(x,y) \in \mathbb{R}^d$,

2.2. **Scale-free parameterization.** In the theory, it will appear convenient to deal with objects with a scale that is independent of m. To this end, we let

$$\boldsymbol{Z}^m := \sqrt{m}\tilde{\boldsymbol{W}}^m(0)$$

(which is a $m \times m$ matrix with independent $\mathcal{N}(0,1)$ entries) and we define:

$$\begin{aligned} \boldsymbol{U}^m(\kappa) &:= \tilde{\boldsymbol{U}}^m(\kappa), \\ \boldsymbol{W}^m(\kappa) &:= m\big(\tilde{\boldsymbol{W}}^m(\kappa) - \tilde{\boldsymbol{W}}^m(0)\big) = m\tilde{\boldsymbol{W}}^m(\kappa) - \sqrt{m}\boldsymbol{Z}^m, \\ \boldsymbol{V}^m(\kappa) &:= m\tilde{\boldsymbol{V}}^m(\kappa) \end{aligned} \tag{2.3}$$

where the scaling factors are adjusted so that these matrices/vectors have entries of order 1, as the theory will verify. By definition, $\boldsymbol{U}^m(0)$ and $\boldsymbol{V}^m(0)$ are random arrays with entries $\mathcal{N}(0,1)$ and $\boldsymbol{W}^m(0)$ is the zero matrix:

$$U_{j\ell}^m(0) \sim \mathcal{N}(0,1), \qquad W_{ij}^m(0) = 0, \qquad V_i^m(0) \sim \mathcal{N}(0,1). \tag{2.4}$$

The neural network in these new variables becomes

$$y = h^m(x, \boldsymbol{U}^m, \boldsymbol{W}^m, \boldsymbol{V}^m) = \left\langle \frac{1}{m}\boldsymbol{V}^m, \left(\frac{1}{\sqrt{m}}\boldsymbol{Z}^m + \frac{1}{m}\boldsymbol{W}^m\right)\boldsymbol{U}^m x \right\rangle.$$

The evolution of $(\boldsymbol{U}^m(\kappa), \boldsymbol{W}^m(\kappa), \boldsymbol{V}^m(\kappa))_{\kappa\in\mathbb{N}}$ can be also interpreted as GD (with layer-wise learning rates) on the objective function

$$F^m(\boldsymbol{U}^m, \boldsymbol{W}^m, \boldsymbol{V}^m) := \int_{\mathbb{R}^d\times\mathbb{R}} \mathcal{L}\left(h^m(x, \boldsymbol{U}^m, \boldsymbol{W}^m, \boldsymbol{V}^m), y\right) \, d\rho(x,y).$$

We do not explicitly include $\boldsymbol{Z}$ in the variables as it is fixed during the training (i.e., we interpret F^m as a random function). All in all, we have

$$\begin{cases} \boldsymbol{U}^m(\kappa+1) = \boldsymbol{U}^m(\kappa) - \tau\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^m + \frac{1}{m}\boldsymbol{W}^m(\kappa)\right]^\top \boldsymbol{V}^m(\kappa)(\boldsymbol{\xi}_{\kappa,\tau}^m)^\top, \\ \boldsymbol{W}^m(\kappa+1) = \boldsymbol{W}^m(\kappa) - \tau\boldsymbol{V}^m(\kappa)(\boldsymbol{\xi}_{\kappa,\tau}^m)^\top(\boldsymbol{U}^m(\kappa))^\top, \\ \boldsymbol{V}^m(\kappa+1) = \boldsymbol{V}^m(\kappa) - \tau\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^m + \frac{1}{m}\boldsymbol{W}^m(\kappa)\right]\boldsymbol{U}^m(\kappa)\boldsymbol{\xi}_{\kappa,\tau}^m, \end{cases} \tag{2.5}$$

where we have denoted $\boldsymbol{\xi}_{\kappa,\tau}^m = \int x \, \mathcal{L}'(h_{\kappa,\tau}^m(x), y) d\rho_\kappa(x,y) \in \mathbb{R}^d$ as above, with $h_{\kappa,\tau}^m(x) = h^m(x, \boldsymbol{U}^m(\kappa), \boldsymbol{W}^m(\kappa), \boldsymbol{V}^m(\kappa))$.

2.3. **Limit dynamics.** Our main result is that, when $m \to \infty$, the training dynamics converge, in a sense detailed below, to some dynamics which are obtained by running the same gradient-based algorithm (i.e., GD or SGD) on an infinitely wide three-layer linear neural network

$$\chi(x, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{G}) = \boldsymbol{B}^\top(\boldsymbol{\Lambda} + \boldsymbol{G})\boldsymbol{A}x, \tag{2.6}$$

where the variables

$$\boldsymbol{A} \in \ell^2(\mathbb{N}\times\{1,\dots,d\}) \subset \mathbb{R}^{\infty\times d}, \quad \boldsymbol{G} \in \ell^2(\mathbb{N}\times\mathbb{N}) \subset \mathbb{R}^{\infty\times\infty}, \quad \boldsymbol{B} \in \ell^2(\mathbb{N}) \subset \mathbb{R}^\infty$$

are initialized with

$$\boldsymbol{A}(0) = \begin{pmatrix} \mathrm{Id}_d \\ \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times 1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_1(0) \\ \vdots \\ \boldsymbol{A}_d(0) \\ \boldsymbol{A}_{d+1}(0) \\ \boldsymbol{A}_{d+2}(0) \\ \vdots \end{pmatrix} \in \mathbb{R}^{\infty\times d}, \quad \boldsymbol{B}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \in \mathbb{R}^{\infty\times 1}, \tag{2.7}$$

where $\boldsymbol{A}_i(0) \in \mathbb{R}^{1\times d}$ for $i \in \mathbb{N}$ and

$$(\boldsymbol{G})_{ij}(0) = 0 \quad \forall (i,j) \in \mathbb{N}^2. \tag{2.8}$$

Also $\boldsymbol{\Lambda}$ is fixed (not trained) and represents the initialization of the intermediate layer. It is given by

$$\boldsymbol{\Lambda} = \begin{pmatrix} \overbrace{0 & \dots & 0}^{\mathrm{d}} & 1 & 0 & 0 & \dots \\ 1 & 0 & \dots & 0 & 1 & 0 & \ddots \\ 0 & 1 & 0 & \dots & 0 & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \in \mathbb{R}^{\infty\times\infty}, \tag{2.9}$$

i.e., $\boldsymbol{\Lambda} = (\Lambda_{ij})_{ij}$ where

$$\Lambda_{ij} = \begin{cases} 1 & \text{if } i + d = j \text{ or } j + 1 = i, \\ 0 & \text{otherwise.} \end{cases}$$

The dynamics are therefore given by the following recursion

$$\begin{cases} \boldsymbol{A}(\kappa+1) &=& \boldsymbol{A}(\kappa) - \tau[\boldsymbol{\Lambda} + \boldsymbol{G}(\kappa)]^\top \boldsymbol{B}(\kappa)\boldsymbol{\xi}_{\kappa,\tau}^\top, \\ \boldsymbol{G}(\kappa+1) &=& \boldsymbol{G}(\kappa) - \tau\boldsymbol{B}(\kappa)\boldsymbol{\xi}_{\kappa,\tau}^\top(\boldsymbol{A}(\kappa))^\top, \\ \boldsymbol{B}(\kappa+1) &=& \boldsymbol{B}(\kappa) - \tau[\boldsymbol{\Lambda} + \boldsymbol{G}(\kappa)]\boldsymbol{A}(\kappa)\boldsymbol{\xi}_{\kappa,\tau}, \end{cases} \tag{2.10}$$

with

$$\chi_{\kappa,\tau}(x) = \chi(x, \boldsymbol{A}(\kappa), \boldsymbol{G}(\kappa), \boldsymbol{B}(\kappa)) \quad \text{and} \quad \boldsymbol{\xi}_{\kappa,\tau} = \int x\mathcal{L}'(\chi_{\kappa,\tau}(x), y)d\rho_\kappa(x,y) \in \mathbb{R}^d.$$

When $\rho_\kappa = \rho$ for all $\kappa \in \mathbb{N}$, this recursion is exactly the GD on the (deterministic) objective function $\mathcal{E}$ defined by

$$\mathcal{E}(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{B}) = \int \mathcal{L}(\boldsymbol{B}^\top(\boldsymbol{\Lambda} + \boldsymbol{G})\boldsymbol{A}x, y)d\rho(x,y). \tag{2.11}$$

2.4. **Main statements.** Let us consider two families of independent infinite Gaussian vectors

$$(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \dots) \qquad \text{and} \qquad (\tilde{\boldsymbol{\Gamma}}_1, \tilde{\boldsymbol{\Gamma}}_2, \dots), \tag{2.12}$$

where the entries of $\boldsymbol{\Gamma}_k, \tilde{\boldsymbol{\Gamma}}_k \in \mathbb{R}^{\mathbb{N}}$ are all independent $\mathcal{N}(0,1)$ random vectors. We define

$$\begin{cases} \boldsymbol{U}^\infty(\kappa) = \sum_{i\geq 1} \boldsymbol{\Gamma}_i \boldsymbol{A}_i(\kappa), \\ \boldsymbol{W}^\infty(\kappa) = \sum_{i,j\geq 1} \tilde{\boldsymbol{\Gamma}}_i \boldsymbol{\Gamma}_j{}^\top G_{ij}(\kappa), \\ \boldsymbol{V}^\infty(\kappa) = \sum_{i\geq 1} \tilde{\boldsymbol{\Gamma}}_i B_i(\kappa). \end{cases} \tag{2.13}$$

We shall prove the convergence in distribution, as $m \to \infty$, of the finite dimensional time-discretized dynamics to the infinite one (see Definition 3.1 for the precise definition of convergence that we use). The proof of convergence will rely on the method of moments: we will prove that the moments of our random variables converge to the ones of the limit as $m \to \infty$, and this implies convergence in distribution. Our main theorem is the following:

Theorem 2.2 (Infinite-width limit)**.** *Let $\tau > 0$ be fixed, let $\mathcal{L}$ be such that $\mathcal{L}''$ is bounded, and let us suppose that* (2.2) *holds.*

Let $(\boldsymbol{U}^m(\kappa), \boldsymbol{W}^m(\kappa), \boldsymbol{V}^m(\kappa))_{\kappa\in\mathbb{N}}$ be the solution to (2.5) *with initialization* (2.4)*, and let $(\boldsymbol{U}^\infty(\kappa), \boldsymbol{W}^\infty(\kappa), \boldsymbol{V}^\infty(\kappa))_{\kappa\in\mathbb{N}}$ be given by* (2.13) *(see* (2.7)-(2.8)-(2.10)*). Then, for any stopping time $\kappa_* \in \mathbb{N}$,*

$$\begin{array}{c} \big((\boldsymbol{U}^m(0), \boldsymbol{W}^m(0), \boldsymbol{V}^m(0)), \dots, (\boldsymbol{U}^m(\kappa_*), \boldsymbol{W}^m(\kappa_*), \boldsymbol{V}^m(\kappa_*))\big) \\ \downarrow \mathrm{d}. \\ \big((\boldsymbol{U}^\infty(0), \boldsymbol{W}^\infty(0), \boldsymbol{V}^\infty(0)), \dots, (\boldsymbol{U}^\infty(\kappa_*), \boldsymbol{W}^\infty(\kappa_*), \boldsymbol{V}^\infty(\kappa_*))\big) \end{array}$$

as $m \to \infty$. Moreover, the vectors in $\mathbb{R}^d$ that represent the linear predictors of the neural network,

$$\begin{aligned} \lambda^m(\kappa) &= \boldsymbol{U}^m(\kappa)^\top (m^{-1/2}\boldsymbol{Z}^m + m^{-1}\boldsymbol{W}^m(\kappa))^\top (m^{-1}\boldsymbol{V}^m(\kappa)) \\ \lambda^\infty(\kappa) &= \boldsymbol{A}(\kappa)^\top (\boldsymbol{\Lambda} + \boldsymbol{G}(\kappa))^\top \boldsymbol{B}(\kappa), \end{aligned}$$

satisfy $\lambda^m(\kappa) \overset{a.s.}{\to} \lambda^\infty(\kappa)$ for every $\kappa \in \mathbb{N}$ (with the quantitative estimate (2.14) *below).*

We can make the following remarks :

(i) Since $(\boldsymbol{U}^\infty_j(\kappa), \boldsymbol{W}^\infty_{i,j}(\kappa), \boldsymbol{V}^\infty_i(\kappa))_{\kappa=1}^{\kappa^*}$ is a *separately exchangeable* $\mathbb{R}^{3\kappa^*}$-valued random array, the dependency structure between its entries that we obtain in Theorem 2.2 is consistent, as it should, with the Aldous-Hover representation of infinite exchangeable arrays [1, Thm. 1.4], which is a generalization of De Finetti's theorem. See [36] for a study of gradient flows with a similar dependency structure.

(ii) A perhaps counter-intuitive consequence of this theorem is that, even if this parametrization μP preserves *feature-learning* in the limit, the evolution of the entries of the intermediate layer $\tilde{\boldsymbol{W}}^m_{i,j}(\kappa) - \tilde{\boldsymbol{W}}^m_{i,j}(0)$ (of order $1/m$) is negligible in front of their magnitude at initialization $\tilde{\boldsymbol{W}}^m_{i,j}(0)$ (of order $1/\sqrt{m}$). Still, these small variations collectively create a significant variation of the output.

(iii) In the proof of this theorem, the convergence of the predictor is quantified as

$$\mathbb{E}\left[\|\lambda^m(\kappa) - \lambda^\infty(\kappa)\|^2\right] \leq C_{\varepsilon,\kappa} m^{-1+\varepsilon}, \tag{2.14}$$

for any $\varepsilon > 0$ and for some $C_{\varepsilon,\kappa}$ depending on $\varepsilon > 0$ and κ, but independent of m. As can be seen from numerical experiments (see Figure 2.1-(C)) this convergence is expected to be (almost) optimal, which is also consistent with the fact that it comes from a Central Limit Theorem.

(iv) Our proofs are based on universality properties and only use that $\boldsymbol{Z}^m$ has i.i.d. subgaussian entries with zero mean and unit variance. In particular, the previous statement is also true for these more general initializations of the $\tilde{\boldsymbol{W}}^m$ weights.

(v) If we want to take more general subgaussian initializations $\boldsymbol{U}^m(0)$ and $\boldsymbol{V}^m(0)$ we can also do it, provided that in the previous statement (more precisely, in (2.12)) we change $\boldsymbol{\Gamma}_1$ and $\tilde{\boldsymbol{\Gamma}}_1$ by $\boldsymbol{U}^\infty(0)$ and $\boldsymbol{V}^\infty(0)$; see Figure 2.3.

Our second statement studies the behavior of the limit model, which is an infinitely wide linear neural network with a particular *deterministic* initialization. For the sake of simplicity, we consider the continuous-time limit $\tau \to 0$ of the dynamics, that is, the gradient flow of the functional F^∞ and the corresponding linear predictor $(\lambda^\infty(t))_{t\geq 0}$.

Theorem 2.3. *Consider the square loss $\mathcal{L}(\hat{y}, y) = \frac{1}{2}|y - \hat{y}|^2$ and assume that ρ has finite second moments. Then $\lambda^\infty(t)$ converges at an exponential rate to the minimal ℓ_2-norm minimizer of the risk $\lambda \mapsto \frac{1}{2}\int |\lambda^\top x - y|^2 d\rho(x, y)$.*

Note that this *implicit bias* towards min-ℓ_2 norm solutions is not a particularly impressive property as such, since just the basic gradient flow on the square-loss initialized from 0, i.e.,

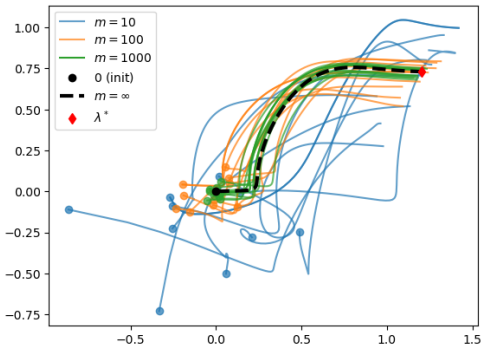
$$\lambda_{\mathrm{gf}}(0) = 0 \qquad\qquad \lambda'_{\mathrm{gf}}(t) = -\int x(\lambda_{\mathrm{gf}}(t)^\top x - y) d\rho(x, y), \tag{2.15}$$

satisfies the same statement (notice, however, that our dynamics are truly non-linear, see Figure 2.2). This result is mostly intended to highlight the fact that our characterization of the infinite-width dynamics is precise enough to obtain such properties.

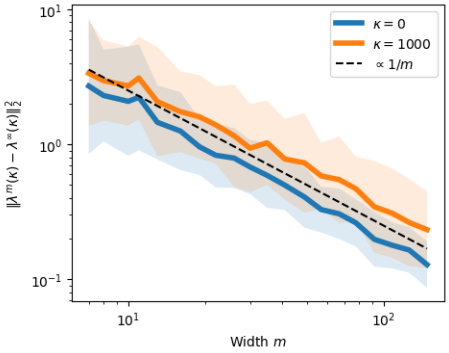
2.5. **Numerical illustrations.** We consider GD for the finite-width and infinite-width models, with input dimension $d = 10$, the square loss, a data distribution given by $x \sim \mathcal{N}(0, \mathrm{Id}_d)$ and $y = x^\top \lambda^*$ for some $\lambda^* \in \mathbb{R}^d$ that is randomly drawn from $\mathcal{N}(0, \mathrm{Id}_d)$. The code to reproduce the experiments is available online[5].

Figure 2.1 illustrates the convergence to the limit model as the width $m \to \infty$, with a step-size $\tau = 0.2$. In (A), we show the path of $(\lambda^m(\kappa))_{\kappa\geq 0}$ projected on two first coordinates of $\mathbb{R}^d$. We observe that, as the width increases, they follow a trajectory approaching that of the limit $(\lambda^\infty(\kappa))_{\kappa\geq 0}$, which starts at $\lambda^\infty(0) = 0$ and converges to the min-ℓ_2 norm predictor λ^* shown as a red diamond, and computed via the pseudo-inverse formula. In (B) we represent the rate of convergence in m of the predictor as a function of the width, at both initialisation and large time. As it

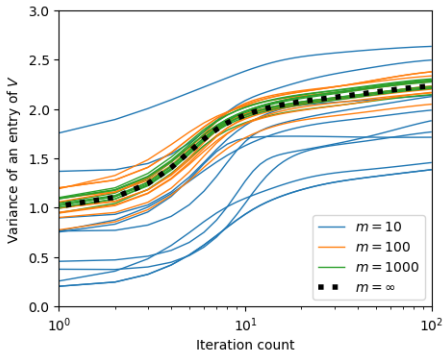
[5] `https://github.com/lchizat/2022-wide-linear-NN`



(A) Convergence in predictor space



(B) Rate of convergence of the predictor



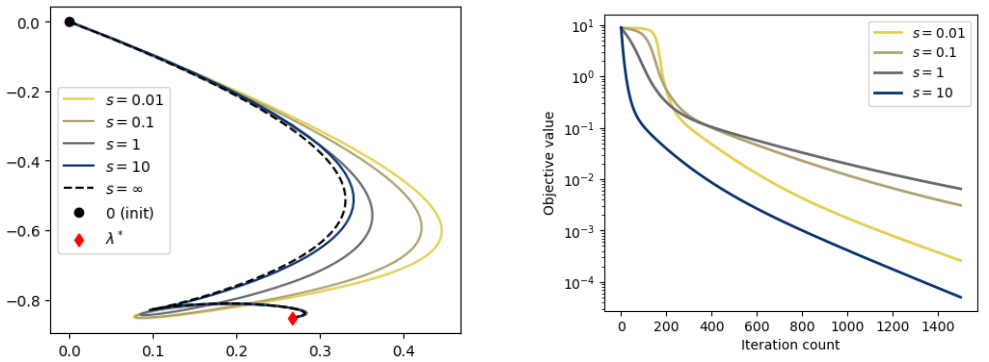
(C) Convergence in parameter space

FIGURE 2.1. Convergence to the limit model (A) Trajectory of the predictor $\lambda^m(t)$, projection on the two first coordinates (bullets represent $\lambda^m(0)$). (B) Rate of convergence of the predictor as a function of the width m, at initialization $\kappa = 0$ and large time $\kappa = 1000$ (shaded area represent standard deviation over 50 repetitions). (C) Evolution of the average square of an entry of $\boldsymbol{V}^m$ and in the limit.

can be seen, it corresponds to (2.14) with $\varepsilon = 0$. Finally, in (C), we represent the mean square of the entries of $\boldsymbol{V}^m(\kappa)$, computed as $v_\kappa = \frac{1}{m}\sum_{i=1}^m \boldsymbol{V}_j(\kappa)^2$ (which is also a proxy for the variance of $\boldsymbol{V}_j^m(\kappa)$ for $1 \le j \le m$ since the entries of $\boldsymbol{V}^m(\kappa)$ are asymptotically independent) and its limit which is $\|\boldsymbol{B}(\kappa)\|_2^2$ by (2.13). This is just a simple example of a statistics described by our limit model.

In Figure 2.2, we take a small step-size to approximate the gradient flow $\tau = 0.001$ and explore the behavior of the limit model. It can be seen from the GD equations that at κ steps, only the first $d \cdot \kappa$ rows of $\boldsymbol{A}(\kappa)$ and of $\boldsymbol{B}(\kappa)$ are non-zero. Thus the infinite model can be trained exactly for a bounded number of steps[6]. We also introduce a fixed scale parameter $s > 0$ that multiplies the predictor, which is equivalent to scaling the standard deviation of the initialization by $s^{1/3}$ at each

[6]We also noticed that truncating the limit model (2.6) below this size introduces an error that decays exponentially in the width, instead of the $m^{-1/2}$ rate for the randomly initialized model.



(A) Trajectory of GD (projection) (B) Evolution of the objective function

FIGURE 2.2. Behavior of the limit model and effect of the scale parameter s. (left) Projection of $\lambda^m(t)$ on the two first coordinates. (right) Evolution of the loss.

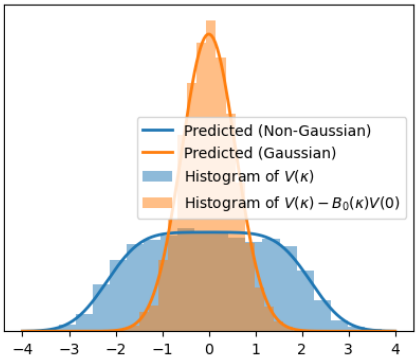
layer. By [12] and since $\lambda^\infty(0) = 0$, we know that as $s \to \infty$, the dynamics converges to the linear dynamics (2.15). This illustration confirms that the dynamics of λ^∞ is non-linear (unless $s \to \infty$), although it has the same endpoints at $t = 0$ and $t = \infty$ as the linear dynamics. For small scales, $s \ll 1$, we observe on the right plot that the objective function starts with a plateau; this is reflected by our convergence analysis in Proposition 5.5, which is a two-phase analysis: a first phase to escape from the initialization (which is close to a stationary point when $s \ll 1$) and a second phase with exponential convergence. We note that the convergence speeds in this plot are not directly comparable because we did not attempt to find the best step-size τ for various values of s.

Finally, in Figure 2.3 (A) we plot the distribution of the weights at large times with non-Gaussian initialization (in blue). As discussed above (remark (v)) the weights are never Gaussian in this case (not even in the large time limit) since in general the first coefficient in the basis (e.g. $\boldsymbol{B}_1(\kappa)$) does not necessarily vanish at $\kappa = \infty$. However, the analysis described in Theorem 2.2 still works and the non-gaussianity of the weights is only due to the interference of this first element of the basis: the other elements are still Gaussian. In the figure, this can be seen by subtracting the first element from the distribution of weights, where we recover a Gaussian profile (in orange). In any case, we still expect a rate of convergence to the minimizer given by the rate of the Central Limit Theorem (Figure 2.3 (B)).

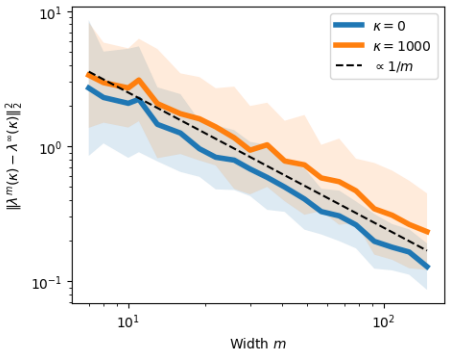
3. An independent family of Gaussian vectors

3.1. **Notation.** We denote vectors and matrices with bolded symbols (except for $x \in \mathbb{R}^d$), and scalars with plain symbols. Given an element $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ with $m, n \in \mathbb{N} \cup \{\infty\}$, we denote

$$\|\boldsymbol{M}\|^2 := \sum_{i=1}^{m}\sum_{j=1}^{n} |M_{ij}|^2.$$



(A) Distribution of $\boldsymbol{V}(\kappa = 20)$.



(B) Rate of convergence of the predictor

FIGURE 2.3. Illustration for a non-Gaussian initialization (centered uniform distribution with the same variance as in the Gaussian case). (A) The distribution of parameters is non-Gaussian at all times, but its exact shape can be computed using the limit model (see remark (v) after Theorem 2.2) (B) The convergence of the predictor to the limit model happens at the same rate as in the Gaussian case.

More generally, for any $p \geq 1$, we denote

$$\|\boldsymbol{M}\|_p^p := \sum_{i=1}^{m} \sum_{j=1}^{n} |M_{ij}|^p.$$

When $\boldsymbol{M}$ is a random array in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we still denote $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ where now it is implicitly evaluated at an element of the sample space $\omega \in \Omega$. In particular, $\|\boldsymbol{M}\|^2 = \|\boldsymbol{M}(\omega)\|^2$ where the evaluation will be implicit whenever there is no ambiguity. It must not be confused with $\mathbb{E}[\|\boldsymbol{M}\|^2]$, which is given by

$$\mathbb{E}[\|\boldsymbol{M}\|^2] = \int_\Omega \|\boldsymbol{M}(\omega)\|^2 d\mathbb{P}(\omega).$$

Finally, let us give the following definition on the convergence in law/distribution for arrays of increasing size:

Definition 3.1. Given a family of random vectors $(\boldsymbol{X}_i^m)_{i \in \mathbb{N}}$ with $\boldsymbol{X}_i^m \in \mathbb{R}^m$, we say that they converge in distribution to a family of infinite-dimensional random vectors $(\boldsymbol{X}_i^\infty)_{i \in \mathbb{N}}$ with $\boldsymbol{X}_i^\infty \in \mathbb{R}^\infty$ if for every fixed $M, N \in \mathbb{N}$, the family $((\boldsymbol{X}_i^m)_{1..N})_{1 \leq i \leq M}$ converges in distribution to $((\boldsymbol{X}_i^\infty)_{1..N})_{1 \leq i \leq M}$ (where we have denoted, for $\boldsymbol{X} \in \mathbb{R}^m$, $\boldsymbol{X}_{1..N}$ its first N components; which is always well defined, for m large enough).

3.2. **The Gaussian bases.** For the sake of readability, we first construct the independent family that will act as a basis of our evolution in the unidimensional input case. We refer to Section 3.5 below for the statements in the multi-dimensional input case, where the proofs are essentially the same.

Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be two infinite random vectors,

$$\boldsymbol{U} := \begin{pmatrix} U_1 \\ U_2 \\ \vdots \end{pmatrix}, \qquad \boldsymbol{V} := \begin{pmatrix} V_1 \\ V_2 \\ \vdots \end{pmatrix}, \tag{3.1}$$

with entries $(U_i)_{i\in\mathbb{N}}$ and $(V_i)_{i\in\mathbb{N}}$ that are independent random variables with $U_i, V_i \sim \mathcal{N}(0,1)$ for $i \in \mathbb{N}$.

Let $\boldsymbol{Z}$ be an infinite random matrix,

$$\boldsymbol{Z} := \begin{pmatrix} Z_{11} & Z_{12} & \dots \\ Z_{21} & Z_{22} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix},$$

whose entries $(Z_{ij})_{i,j\in\mathbb{N}}$ are independent random variables $Z_{ij} \sim \mathcal{N}(0,1)$ for all $i, j \in \mathbb{N}$, and also independent from $(U_i)_{i\in\mathbb{N}}$ and $(V_i)_{i\in\mathbb{N}}$.

Let us denote by $\boldsymbol{U}^m$ the restriction of $\boldsymbol{U}$ to the first m entries, $\boldsymbol{U}^m \in \mathbb{R}^m$, with $U_i^m = U_i$ for $1 \le i \le m$ (respectively $\boldsymbol{V}^m$). Similarly, we denote by $\boldsymbol{Z}^m$ the restriction of $\boldsymbol{Z}$ to the first $m \times m$ entries, $\boldsymbol{Z}^m \in \mathbb{R}^{m\times m}$, with $Z_{ij}^m = Z_{ij}$ for $1 \le i, j \le m$.

Let us denote by $\boldsymbol{\Xi}_i^m(k)$ the set of bipartite (directed) k-chains between two equal sets of indices $I_0 = I_1 = \{1, \dots, m\}$ that start in I_1 and end at i, where $i \in I_0$ if k is odd, and $i \in I_1$ if k is even. Namely,

$$\begin{aligned}\boldsymbol{\Xi}_i^m(k) := \{\, &((i_2,i_1),(i_2,i_3),(i_4,i_3),(i_4,i_5),\dots,(i_{k-1},i_k),(i,i_k)) : \\ &i, i_{2j} \in I_0 \ \text{ for } 1 \le j \le (k-1)/2, \quad i_{2j-1} \in I_1 \ \text{ for } 1 \le j \le (k+1)/2\}\end{aligned}$$

if $k \in \mathbb{N}$ is odd, and

$$\begin{aligned}\boldsymbol{\Xi}_i^m(k) := \{\, &((i_2,i_1),(i_2,i_3),(i_4,i_3),(i_4,i_5),\dots,(i_k,i_{k-1}),(i_k,i)) : \\ &i_{2j} \in I_0 \ \text{ for } 1 \le j \le k/2, \quad i, i_{2j-1} \in I_1 \ \text{ for } 1 \le j \le k/2\}\end{aligned}$$

if $k \in \mathbb{N}$ is even. In particular, an element $\Xi \in \boldsymbol{\Xi}_i^m(k)$ is of the form $\Xi = (\Xi_1, \dots, \Xi_k)$ where $\Xi_\ell = ((\Xi_\ell)_1, (\Xi_\ell)_2)$ with $(\Xi_\ell)_1 \in I_0$ and $(\Xi_\ell)_2 \in I_1$, for $1 \le \ell \le k$,

$$\Xi_\ell = (i_\ell, i_{\ell+1}) \quad \text{if } \ell \text{ is even}, \qquad \Xi_\ell = (i_{\ell+1}, i_\ell) \quad \text{if } \ell \text{ is odd}, \qquad \text{and } i_{k+1} = i.$$

We think of each Ξ_ℓ for $1 \le \ell \le k$ as possible indices of a matrix $m \times m$. In this way, if $\Xi \in \boldsymbol{\Xi}_i^m(k)$ we denote $\Xi^\top := (\Xi_1^\top, \dots, \Xi_K^\top)$ where $\Xi_\ell^\top := ((\Xi_\ell)_2, (\Xi_\ell)_1)$ (that is, transposing every matrix; or alternatively, reflecting the bipartite chain).

We can use the previous definitions to compute iterative multiplications of $(\boldsymbol{Z}^m)^\top$ and $\boldsymbol{Z}^m$ against $\boldsymbol{U}^m$. That is, (using $(\Xi_1)_2 = i_1$ in the notation above)

$$\begin{aligned}\left(\boldsymbol{Z}^m[(\boldsymbol{Z}^m)^\top(\boldsymbol{Z}^m)]^{\frac{k-1}{2}}\boldsymbol{U}^m\right)_i &= \big(\overbrace{\boldsymbol{Z}^m(\boldsymbol{Z}^m)^\top \dots \boldsymbol{Z}^m}^{k}\boldsymbol{U}^m\big)_i \\ &= \sum_{i_1,\dots,i_k=1}^{m} Z_{i,i_k}\dots Z_{i_2,i_3}Z_{i_2,i_1}U_{i_1} = \sum_{\Xi\in\boldsymbol{\Xi}_i^m(k)}\left(\prod_{\ell=1}^{k} Z_{\Xi_\ell}\right)U_{(\Xi_1)_2}\end{aligned} \tag{3.2}$$

if $k \in \mathbb{N}$ is odd, and

$$\begin{aligned}\left([(\boldsymbol{Z}^m)^\top(\boldsymbol{Z}^m)]^{\frac{k}{2}}\boldsymbol{U}^m\right)_i &= \big(\overbrace{(\boldsymbol{Z}^m)^\top \boldsymbol{Z}^m \dots \boldsymbol{Z}^m}^{k}\boldsymbol{U}^m\big)_i \\ &= \sum_{i_1,\dots,i_k=1}^{m} Z_{i_k,i}\dots Z_{i_2,i_3}Z_{i_2,i_1}U_{i_1} = \sum_{\Xi\in\boldsymbol{\Xi}_i^m(k)}\left(\prod_{\ell=1}^{k} Z_{\Xi_\ell}\right)U_{(\Xi_1)_2}\end{aligned} \tag{3.3}$$

if $k \in \mathbb{N}$ is even.

Let us denote by $\#v(\Xi)$ with $\Xi \in \boldsymbol{\Xi}_i^m(k)$ the number of vertices seen by the k-chain Ξ, namely,[7]

$$\begin{aligned}\#v(\Xi) &= |\{i_2, i_4, \dots, i_{k-1}, i\}| + |\{i_1, i_3, \dots, i_k\}|, \\ &\quad \text{where } \Xi = ((i_2,i_1),(i_2,i_3),\dots,(i_{k-1},i_k),(i,i_k)) \in \boldsymbol{\Xi}_i^m(k)\end{aligned}$$

if $k \in \mathbb{N}$ is odd, and

$$\begin{aligned}\#v(\Xi) &= |\{i_2, i_4, \dots, i_k\}| + |\{i_1, i_3, \dots, i_{k-1}, i\}|, \\ &\quad \text{where } \Xi = ((i_2,i_1),(i_2,i_3),\dots,(i_k,i_{k-1}),(i_k,i)) \in \boldsymbol{\Xi}_i^m(k)\end{aligned}$$

if $k \in \mathbb{N}$ is even.

Let us define $\tilde{\boldsymbol{\Xi}}_i^m(k)$ to be the subset of $\boldsymbol{\Xi}_i^m(k)$ with chains that contain no loops (alternatively, chains that visit each vertex at most once),

$$\tilde{\boldsymbol{\Xi}}_i^m(k) := \{\Xi \in \boldsymbol{\Xi}_i^m(k) : \#v(\Xi) = k+1\} \subset \boldsymbol{\Xi}_i^m(k).$$

Finally, we define (cf. (3.2)-(3.3)),

$$\begin{aligned}J_{k,i}^m &:= \frac{1}{m^{k/2}}\sum_{\Xi\in\tilde{\boldsymbol{\Xi}}_i^m(k)}\left(\prod_{\ell=1}^{k} Z_{\Xi_\ell}\right)U_{(\Xi_1)_2}, \\ K_{k,i}^m &= \frac{1}{m^{k/2}}\sum_{\Xi\in\tilde{\boldsymbol{\Xi}}_i^m(k)}\left(\prod_{\ell=1}^{k} Z_{\Xi_\ell^\top}\right)V_{(\Xi_1)_2},\end{aligned} \tag{3.4}$$

namely, we consider the product in (3.2)-(3.3) (both against $\boldsymbol{U}$ and $\boldsymbol{V}$) but we keep only those elements of the sum that have no loops (and we rescale by the appropriate size, where the size-preserving objects are $m^{-1/2}\boldsymbol{Z}^m$). Notice that in the multiplication against $\boldsymbol{V}$ we are considering matrices $\boldsymbol{Z}^\top$ instead of $\boldsymbol{Z}$. In particular, we could alternatively think of $K_{k,i}^m$ as the (renormalized) sum over loopless chains starting from the set I_0 and ending at i and with length k, where now $i \in I_0$ if k is even and $i \in I_1$ if k is odd (the opposite from before).

We define the vectors

$$\boldsymbol{J}_k^m := \begin{pmatrix} J_{k,1}^m \\ J_{k,2}^m \\ \vdots \\ J_{k,m}^m \end{pmatrix} \quad \text{and} \quad \boldsymbol{K}_k^m := \begin{pmatrix} K_{k,1}^m \\ K_{k,2}^m \\ \vdots \\ K_{k,m}^m \end{pmatrix}, \quad \text{with } J_{k,i}^m \text{ and } K_{k,i}^m \text{ given by (3.4).} \tag{3.5}$$

[7]Here and in the sequel, given a finite set A, we denote by $|A|$ its cardinality.

(We denote $\boldsymbol{J}_0^m = \boldsymbol{U}^m$ and $\boldsymbol{K}_0^m = \boldsymbol{V}^m$.) Let us also define, for $k \in \mathbb{N}$, $\{\boldsymbol{J}_k\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{K}_k\}_{k\in\mathbb{N}}$ families of independent, identically distributed (infinite) random vectors

$$\boldsymbol{J}_k := \begin{pmatrix} J_{k,1} \\ J_{k,2} \\ \vdots \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{K}_k := \begin{pmatrix} K_{k,1} \\ K_{k,2} \\ \vdots \end{pmatrix}. \tag{3.6}$$

with $J_{k,i}, K_{k,i} \sim \mathcal{N}(0,1)$ and all independent between them.

3.3. **Convergence of the family.** Let us now prove that the family of vectors $\{(\boldsymbol{J}_k^m, \boldsymbol{K}_k^m)\}_{k\in\mathbb{N}}$ converges in distribution to the family of i.i.d. Gaussian vectors. We will in fact prove convergence of all the moments.

Theorem 3.2. *The family of vectors $\{(\boldsymbol{J}_k^m, \boldsymbol{K}_k^m)\}_{k\in\mathbb{N}}$ defined in* (3.5) *converges, in distribution, to the family $\{(\boldsymbol{J}_k, \boldsymbol{K}_k)\}_{k\in\mathbb{N}}$ (according to Definition 3.1).*

This statement says that products of the form (3.2)-(3.3) have a simple asymptotic structure *provided that we remove all the chains of indices with loops.* The chains with loops add correlations, which are described in the next proposition on the recursion property for this family.

Proof. We will show that, for each $N_J, N_K \in \mathbb{N}$ and every fixed families of pairs of indices $(k_1, i_1), \dots, (k_{N_J}, i_{N_J})$ and $(\ell_1, j_1), \dots, (\ell_{N_K}, j_{N_K})$, we have

$$(J^m_{k_1,i_1}, \dots J^m_{k_{N_J},i_{N_J}}, K^m_{\ell_1,j_1}, \dots K^m_{\ell_{N_K},j_{N_K}}) \xrightarrow{d.} (J_{k_1,i_1}, \dots J_{k_{N_J},i_{N_J}}, K_{\ell_1,j_1}, \dots K_{\ell_{N_K},j_{N_K}})$$

as $m \to \infty$.

We use the method of moments. Let us fix the indices $(k_1, i_1), \dots, (k_{N_J}, i_{N_J})$ and $(\ell_1, j_1), \dots, (\ell_{N_K}, j_{N_K})$, and the powers $p_1, \dots, p_{N_J}, q_1, \dots, q_{N_K} \in \mathbb{N}$, and let

$$\mathcal{E}_m := \mathbb{E}\left[(J^m_{k_1,i_1})^{p_1} \dots (J^m_{k_{N_J},i_{N_J}})^{p_{N_J}} (K^m_{\ell_1,j_1})^{q_1} \dots (K^m_{\ell_{N_K},j_{N_K}})^{q_{N_K}}\right] \tag{3.7}$$

where we assume that $J^m_{k_1,i_1}, \dots, J^m_{k_{N_J},i_{N_J}}, K^m_{\ell_1,j_1}, \dots, K^m_{\ell_{N_K},j_{N_K}}$ are all different. We will show that

$$\mathcal{E}_m \xrightarrow{m\to\infty} \mu_{p_1} \dots \mu_{p_{N_J}} \mu_{q_1} \dots \mu_{q_{N_K}},$$

where μ_k denotes the k-th plain moments of a normal distribution $\mathcal{N}(0,1)$,

$$\mu_k = \begin{cases} 0 & \text{if } k \in \mathbb{N} \text{ is odd,} \\ (k-1)!! & \text{if } k \in \mathbb{N} \text{ is even,} \end{cases}$$

with $(k-1)!! = (k-1)\cdot(k-3)\cdot\dots\cdot 5\cdot 3\cdot 1$ being the double factorial. This will directly give the desired result.

Recall that each element $J^m_{k_1,i_1}, \dots, J^m_{k_{N_J},i_{N_J}}, K^m_{\ell_1,j_1}, \dots, K^m_{\ell_{N_K},j_{N_K}}$ can be thought of as a sum over bipartite loopless chains between I_0 and I_1, starting at I_1 for $J^m_{k_\alpha,i_\alpha}$, starting at I_0 for $K^m_{k_{\alpha'},i_{\alpha'}}$, and ending at $i_1, \dots, i_{N_J}, j_1, \dots, j_{N_k}$ with length $k_1, \dots, k_{N_J}, \ell_1, \dots, \ell_{N_K}$, respectively; also, each ending vertex belongs to either I_0 or I_1 depending on the parity of the length (k_α or $\ell_{\alpha'}$). We denote by $\#v_{\text{end}}$ the

number of different such ending vertices,

$$\#v_{\text{end}} := \left| \bigcup_{\substack{\alpha=1 \\ k_\alpha \equiv 1 \bmod 2}}^{N_J} \{i_\alpha\} \cup \bigcup_{\substack{\alpha'=1 \\ \ell_{\alpha'} \equiv 0 \bmod 2}}^{N_K} \{j_{\alpha'}\} \right| + \left| \bigcup_{\substack{\alpha=1 \\ k_\alpha \equiv 0 \bmod 2}}^{N_J} \{i_\alpha\} \cup \bigcup_{\substack{\alpha'=1 \\ \ell_{\alpha'} \equiv 1 \bmod 2}}^{N_K} \{j_{\alpha'}\} \right|.$$

Let us define by $\mathfrak{G}$ the family of bipartite graphs (each graph is seen as a disjoint union of edges) connecting the sets of vertices I_0 and I_1 appearing in the expansion of the definition of $\mathcal{E}_m$. Namely, any graph $G \in \mathfrak{G}$ is a set of edges between I_0 and I_1 given by the (disjoint) union of p_1 elements in $\tilde{\Xi}^m_{i_1}(k_1)$, p_2 elements in $\tilde{\Xi}^m_{i_2}(k_2)$, ..., and q_{N_K} elements in $\tilde{\Xi}^m_{j_{N_K}}(\ell_{N_K})$:

$$\begin{aligned} \mathfrak{G} := \Bigg\{ G : G = \bigsqcup_{\alpha=1}^{N_J} \bigsqcup_{\beta=1}^{p_\alpha} \Xi^{\beta,\alpha} \sqcup \bigsqcup_{\alpha'=1}^{N_K} \bigsqcup_{\beta'=1}^{q_{\alpha'}} \Theta^{\beta',\alpha'}, \quad \text{for some} \quad \Xi^{\beta,\alpha} \in \tilde{\Xi}^m_{i_\alpha}(k_\alpha) \\ \text{and} \quad \Theta^{\beta',\alpha'} \in \tilde{\Xi}^m_{j_{\alpha'}}(\ell_{\alpha'}) \quad \text{with} \quad 1 \le \alpha \le N_J, 1 \le \alpha' \le N_K \Bigg\}. \end{aligned} \tag{3.8}$$

Observe that each element $G \in \mathfrak{G}$ contains $\#e(G)$ edges (with multiplicity), where

$$\#e(G) = k_1 p_1 + \cdots + k_{N_J} p_{N_J} + \ell_1 q_1 + \cdots + \ell_{N_K} q_{N_K} =: N, \tag{3.9}$$

which is independent of the element $G \in \mathfrak{G}$ chosen.

Also, given a fixed element

$$\mathfrak{G} \ni G = \bigsqcup_{\alpha=1}^{N_J} \bigsqcup_{\beta=1}^{p_\alpha} \Xi^{\beta,\alpha} \sqcup \bigsqcup_{\alpha'=1}^{N_K} \bigsqcup_{\beta'=1}^{q_{\alpha'}} \Theta^{\beta',\alpha'},$$

we denote by

$$U(G) = \prod_{\alpha=1}^{N_J} \prod_{\beta=1}^{p_\alpha} U_{(\Xi_1^{\beta,\alpha})_2} \qquad \text{and} \qquad V(G) = \prod_{\alpha'=1}^{N_K} \prod_{\beta'=1}^{q_{\alpha'}} V_{(\Theta_1^{\beta',\alpha'})_2}.$$

(Recall that $(\Xi_1^{\beta,\alpha})_2$ and $(\Theta_1^{\beta',\alpha'})_2$ denote the starting vertex of the chains $\Xi^{\beta,\alpha}$ and $\Theta^{\beta',\alpha'}$ respectively.) Then, we can rewrite (3.7) in terms of $\mathfrak{G}$ by expanding the products as

$$\mathcal{E}_m = m^{-\frac{N}{2}} \mathbb{E} \left[\sum_{G \in \mathfrak{G}} \left(\prod_{e \in G} Z_e \right) U(G) V(G) \right].$$

(Recall (3.9).) By denoting $\text{mult}_G(e)$ the multiplicity of an edge e in G, we can define

$$\mathfrak{G}_2 := \{ G \in \mathfrak{G} : \text{mult}_G(e) \ge 2 \quad \text{for all} \quad e \in G \},$$

that is, the subset of $\mathfrak{G}$ whose graphs have edges all with multiplicity 2 or higher. By linearity of the expected value, and the fact that all Z_{ij}, U_i, V_j are independent between them and with average zero, we immediately have that, in fact, we can sum only over $\mathfrak{G}_2$,

$$\mathcal{E}_m = m^{-\frac{N}{2}} \mathbb{E} \left[\sum_{G \in \mathfrak{G}_2} \left(\prod_{e \in G} Z_e \right) U(G) V(G) \right].$$

Let us denote, for any $G \in \mathfrak{G}_2$, $\#v(G)$ the number of different vertices seen by the edges in G. In particular, since each edge appears twice for $G \in \mathfrak{G}_2$, we have that

$$\#v(G) \leq \#v_{\text{end}} + \frac{N}{2}, \tag{3.10}$$

where we are using that the last $\#v_{\text{end}}$ vertices are fixed, that we can add edges (from the end) in such a way that they always see at most one new vertex (since G is connected), and that each edge appears at least twice.

Notice that we have equality in (3.10) only if each edge in G has multiplicity exactly 2 (otherwise, we would be seeing less vertices than the maximum possible; at some point adding one edge on G would neither contribute to a new vertex nor be a first time repetition):

$$\#v(G) = \#v_{\text{end}} + \frac{N}{2} \quad \Rightarrow \quad \text{for all } e \in G, \ \ \text{mult}_G(e) = 2. \tag{3.11}$$

Let us define $\mathfrak{G}_M$ as the subset of graphs in $\mathfrak{G}_2$ that see the maximum number of vertices,

$$\mathfrak{G}_M := \left\{ G \in \mathfrak{G}_2 : \#v(G) = \#v_{\text{end}} + \frac{N}{2} \right\},$$

and let us compute $|\mathfrak{G}_2 \setminus \mathfrak{G}_M|$. The elements in $\mathfrak{G}_2 \setminus \mathfrak{G}_M$ are all bipartite graphs G between I_0 and I_1 with $\#v(G) < \#v_{\text{end}} + \frac{N}{2}$ vertices. Since the last $\#v_{\text{end}}$ vertices are fixed, the number of elements in $\mathfrak{G}_2 \setminus \mathfrak{G}_M$ will be upper bounded by the number of ways to choose the remaining $\#v(G) - \#v_{\text{end}}$ vertices (among $2m$, that is, $(2m - \#v_{\text{end}}) \cdot (2m - \#v_{\text{end}} - 1) \cdot \cdots \cdot (2m - \#v(G) + 1) \leq C m^{\#v(G) - \#v_{\text{end}}}$). We are also using that, for each configuration of vertices, there is a bounded number of possible graphs with such vertices that is independent of m (but may depend on N_J, N_K, etc.). In all, since $\#v(G) - \#v_{\text{end}} \leq \frac{N}{2} - 1$,

$$|\mathfrak{G}_2 \setminus \mathfrak{G}_M| \leq C m^{\frac{N}{2} - 1}$$

for some C independent of m. Using that all the elements Z_{ij}, U_i, V_j, have finite moments, we obtain that

$$\left| \mathcal{E}_m - m^{-\frac{N}{2}} \mathbb{E} \left[\sum_{G \in \mathfrak{G}_M} \left(\prod_{e \in G} Z_e \right) U(G) V(G) \right] \right| \leq \frac{C}{m} \tag{3.12}$$

for some C independent of m.

Let now $G \in \mathfrak{G}_M$ be fixed, a graph with maximal number of vertices, (3.11), with N edges each with multiplicity two. Let us count the edges from the vertices to obtain a further characterization of G:

The last $\#v_{\text{end}}$ vertices are the ending points of the $p + q := p_1 + \cdots + p_{N_J} + q_1 + \cdots + q_{N_K}$ chains, and as such, they are connected to at least $p + q$ edges. From the remaining $\frac{N}{2}$ vertices, let us denote by $\#v_E(G)$ the ones that see exactly two edges (which must be the same edge, repeated twice). Then, $\#v_E(G) \leq \frac{1}{2}(p + q)$. Indeed, since chains have no loops, the same edge cannot be repeated inside a chain, and elements of $v_E(G)$ are necessarily reached by two different chains (and hence they are a starting point for each one). Thus, these starting points see $2\#v_E(G) \leq p + q$ edges (counting with multiplicity).

Finally, the remaining $\#v(G) - \#v_{\text{end}} - \#v_E(G) = \frac{N}{2} - \#v_E(G)$ vertices see at least four edges each one, so that the total amount of edges as seen from the vertices (i.e., the sum over vertices of the number of edges seen by each vertex) is:

$$\begin{aligned} 2N = 2\#e(G) &\geq p + q + 2\#v_E(G) + 4(\#v(G) - \#v_{\text{end}} - \#v_E(G)) \\ &= p + q - 2\#v_E(G) + 2N \geq 2N, \end{aligned}$$

where we are also using that each edge is seen from two vertices. In particular, all the previous inequalities are, in fact, equalities, and there are exactly $p + q$ edges connected to $\#v_{\text{end}}$, exactly $\frac{1}{2}(p + q)$ vertices that are starting points (seeing only two edges each), and all the remaining vertices see four edges (two edges, each with multiplicity two).

At the level of $G \in \mathfrak{G}_M$ this implies that each chain in its definition is repeated exactly identically twice, and that they never share vertices (except for the final ones). In particular, there is an even number of chains ending at each vertex: if $\mathfrak{G}_M \neq \emptyset$ then all $p_1, \dots, p_{N_J}, q_1, \dots, q_{N_K}$ are even. Observe, also, that this implies that

$$\mathbb{E}\left[\left(\prod_{e \in G} Z_e\right) U(G)V(G)\right] = 1 \qquad \text{for all} \qquad G \in \mathfrak{G}_M,$$

where we are using $\mathbb{E}[Z_{ij}^2] = \mathbb{E}[U_i^2] = \mathbb{E}[V_j^2] = 1$.

A short combinatorial argument combined with (3.12) now gives the desired result: we need to count in how many ways we can produce graphs in $G \in \mathfrak{G}_M$ with the definition (3.8) in such a way that each chain is repeated identically twice and they never share non-ending vertices. We choose first the chains, which can be done in $m^{\frac{N}{2}}$ ways at leading order (for each chain we choose the previous vertex starting from the end, so there is always $m - r$ possibilities, where r is a bounded number independent of m; and we do so for each of the N edges, each of which is repeated twice). For each family of p_α chains ending at i_α we now have multiple ways to produce the same graph $G \in \mathfrak{G}_M$: for each p_α (and $q_{\alpha'}$) we need to count the number of ways in which a family of p_α (and $q_{\alpha'}$) elements can be divided into couples, and then do the same for each α and α'.

In all, given a family of $2n$ elements with $n \in \mathbb{N}$, there are $\frac{(2n)!}{2^n n!}$ ways to split it into couples: there are $(2n)!$ ways to arrange them in a line and we now split them in order into couples. Since we can change the order within each pair, and we can change the order of the pairs, we are actually generating each possible configuration $2^n n!$ times. The number of ways to split $2n$ elements into couples is then $\frac{(2n)!}{2^n n!} = (2n-1)!!$.

Thus, given a fixed graph $G \in \mathfrak{G}_M$, we have $\prod_{\alpha'}^{N_J}(p_\alpha - 1)!! \prod_{\alpha'}^{N_K}(q_{\alpha'} - 1)!!$ ways to produce the same graph with the previous constructions. Combined with (3.12) and the fact that there are $m^{\frac{N}{2}}$ possible configurations (at leading order) we have

$$\left|\mathcal{E}_m - \mu_{p_1} \dots \mu_{p_{N_J}} \mu_{q_1} \dots \mu_{q_{N_K}}\right| \leq \frac{C}{m}$$

for some C independent of m. □

3.4. **A recursion property.** We next show a recursion property for the family (3.5) that will be crucial in the following section (recall the notation from subsection 3.1).

Proposition 3.3. *The random vectors $\boldsymbol{J}_k^m$ and $\boldsymbol{K}_k^m$ satisfy,*

$$\begin{aligned}
m^{-\frac{1}{2}}\boldsymbol{Z}^m\boldsymbol{J}_k^m &= \boldsymbol{J}_{k+1}^m + \boldsymbol{J}_{k-1}^m + \boldsymbol{R}_k^m, && \text{if } k\in\mathbb{N} \text{ is even},\\
m^{-\frac{1}{2}}(\boldsymbol{Z}^m)^\top\boldsymbol{J}_k^m &= \boldsymbol{J}_{k+1}^m + \boldsymbol{J}_{k-1}^m + \boldsymbol{R}_k^m, && \text{if } k\in\mathbb{N} \text{ is odd},\\
m^{-\frac{1}{2}}(\boldsymbol{Z}^m)^\top\boldsymbol{K}_k^m &= \boldsymbol{K}_{k+1}^m + \boldsymbol{K}_{k-1}^m + \boldsymbol{S}_k^m, && \text{if } k\in\mathbb{N} \text{ is even},\\
m^{-\frac{1}{2}}\boldsymbol{Z}^m\boldsymbol{K}_k^m &= \boldsymbol{K}_{k+1}^m + \boldsymbol{K}_{k-1}^m + \boldsymbol{S}_k^m, && \text{if } k\in\mathbb{N} \text{ is odd},
\end{aligned}$$

for some random vectors $\boldsymbol{R}_k^m$ and $\boldsymbol{S}_k^m$ with

$$\mathbb{E}\left[\|\boldsymbol{R}_k^m\|_{2p}^{2p}\right] + \mathbb{E}\left[\|\boldsymbol{S}_k^m\|_{2p}^{2p}\right] \leq C_p < +\infty,$$

for any $p\in\mathbb{N}$, and for some C independent of m (but it might depend on k and p).

Proof. Let us do the first equality, the others follow by analogy. Thus, we assume $k\in\mathbb{N}$ is even and we deal with $\boldsymbol{J}_k^m$.

$$\begin{aligned}
(m^{-\frac{1}{2}}\boldsymbol{Z}^m\boldsymbol{J}_k^m)_j &= m^{-\frac{k+1}{2}}\sum_{i=1}^m Z_{j,i}\sum_{\Xi\in\tilde{\boldsymbol{\Xi}}_i^m(k)}\left(\prod_{\ell=1}^k Z_{\Xi_\ell}\right)U_{(\Xi_1)_2}\\
&= m^{-\frac{k+1}{2}}\sum_{\Xi\in\mathring{\boldsymbol{\Xi}}_j^m(k+1)}\left(\prod_{\ell=1}^{k+1} Z_{\Xi_\ell}\right)U_{(\Xi_1)_2},
\end{aligned}$$

where we are denoting

$$\mathring{\boldsymbol{\Xi}}_j^m(k+1) := \left\{\Xi\cup\{(j,i)\} : \Xi\in\tilde{\boldsymbol{\Xi}}_i^m(k) \quad \text{for some} \quad 1\leq i\leq m\right\}.$$

That is, we are taking loopless chains starting in I_1 and with length k, and adding an extra edge towards j at the end. In particular, we can divide:

$$\mathring{\boldsymbol{\Xi}}_j^m(k+1) = \tilde{\boldsymbol{\Xi}}_j^m(k+1)\cup\mathring{\boldsymbol{\Xi}}_{j,*}^m(k+1)\cup\mathring{\boldsymbol{\Xi}}_{j,r}^m(k+1),$$

where

$$\mathring{\boldsymbol{\Xi}}_{j,*}^m(k+1) := \left\{\Xi\in\mathring{\boldsymbol{\Xi}}_j^m(k+1) \quad \text{such that} \quad \Xi_{k+1}=\Xi_k\right\},$$

namely, the added edge was already part of the chain (and since we are adding it to a loopless chain, it must be the last edge); and

$$\mathring{\boldsymbol{\Xi}}_{j,r}^m(k+1) := \mathring{\boldsymbol{\Xi}}_j^m(k+1)\setminus\left(\tilde{\boldsymbol{\Xi}}_j^m(k+1)\cup\mathring{\boldsymbol{\Xi}}_{j,*}^m(k+1)\right)$$

those chains where the extra edge is not adding a new vertex, but is not a repeated edge either. Thus,

$$(m^{-\frac{1}{2}}\boldsymbol{Z}^m\boldsymbol{J}_k^m)_j = J_{k+1,j}^m + A_{k+1,j}^m + B_{k+1,j}^m, \tag{3.13}$$

where, if we denote $v_1(\Xi)$ for $\Xi \in \mathring{\boldsymbol{\Xi}}^m_j(k-1)$ the set of vertices in Ξ from I_1,

$$A^m_{k+1,j} = m^{-\frac{k+1}{2}} \sum_{\Xi \in \mathring{\boldsymbol{\Xi}}^m_{j,*}(k+1)} \left(\prod_{\ell=1}^{k+1} Z_{\Xi_\ell} \right) U_{(\Xi_1)_2}$$

$$= m^{-\frac{k+1}{2}} \sum_{\Xi \in \tilde{\boldsymbol{\Xi}}^m_j(k-1)} \sum_{i \notin v_1(\Xi)} Z^2_{j,i} \left(\prod_{\ell=1}^{k-1} Z_{\Xi_\ell} \right) U_{(\Xi_1)_2},$$

$$B^m_{k+1,j} = m^{-\frac{k+1}{2}} \sum_{\Xi \in \mathring{\boldsymbol{\Xi}}^m_{j,r}(k+1)} \left(\prod_{\ell=1}^{k+1} Z_{\Xi_\ell} \right) U_{(\Xi_1)_2}.$$

Observe now that, on the one hand, using the same arguments as in Theorem 3.2, we can directly compute

$$\mathbb{E}\left[(B^m_{k+1,j})^{2p} \right] \leq \frac{C_p}{m}, \tag{3.14}$$

for any $p \in \mathbb{N}$, and for some C independent of m. (We are using here that in the sum we are only considering elements that do not see the maximal number of vertices.)

On the other hand, we can rewrite

$$A^m_{k+1,j} = J^m_{k-1,j} + D^m_{k+1,j} + \tilde{D}^m_{k+1,j}, \tag{3.15}$$

with

$$D^m_{k+1,j} = m^{-\frac{k+1}{2}} \sum_{\Xi \in \tilde{\boldsymbol{\Xi}}^m_j(k-1)} \sum_{i \notin v_1(\Xi)} (Z^2_{j,i} - 1) \left(\prod_{\ell=1}^{k-1} Z_{\Xi_\ell} \right) U_{(\Xi_1)_2},$$

$$\tilde{D}^m_{k+1,j} = m^{-\frac{k+1}{2}} \sum_{\Xi \in \tilde{\boldsymbol{\Xi}}^m_j(k-1)} |v_1(\Xi)| \left(\prod_{\ell=1}^{k-1} Z_{\Xi_\ell} \right) U_{(\Xi_1)_2}.$$

From the same arguments as in Theorem 3.2 (since $|v_1(\Xi)|$ is bounded independent of m) we get on the one hand that

$$\mathbb{E}\left[(\tilde{D}^m_{k+1,j})^{2p} \right] \leq \frac{C_p}{m^2} \tag{3.16}$$

for any $p \in \mathbb{N}$, and on the other hand, since $Z^2_{j,i} - 1$ has average zero and is independent of all the other elements in each term of the sum (since $i \notin v_1(\Xi)$), the same type of reasoning done in Theorem 3.2 also gives

$$\mathbb{E}\left[(D^m_{k+1,j})^{2p} \right] \leq \frac{C_p}{m} \tag{3.17}$$

for any $p \in \mathbb{N}$.

In all, joining (3.13)-(3.14)-(3.15)-(3.16)-(3.17),

$$(m^{-\frac{1}{2}} \boldsymbol{Z}^m \boldsymbol{J}^m_k)_j = J^m_{k+1,j} + J^m_{k-1,j} + R^m_{k,j},$$

with $\mathbb{E}\left[(R^m_{k,j})^{2p} \right] \leq \frac{C_p}{m}$ for any $p \in \mathbb{N}$, and for some C independent of m. Using the symmetry of the problem, we get the desired result. □

3.5. **Multi-dimensional input.** More generally, we can take $d \in \mathbb{N}$ i.i.d. copies of $\boldsymbol{U}$, denoted $\boldsymbol{U}^{(1)}, \dots \boldsymbol{U}^{(d)}$ (also independent of $\boldsymbol{V}$) and define

$$\boldsymbol{U}^{(1\dots d)} := \begin{pmatrix} \boldsymbol{U}^{(1)} & \dots & \boldsymbol{U}^{(d)} \end{pmatrix} = \begin{pmatrix} U_1^{(1)} & \dots & U_1^{(d)} \\ U_2^{(1)} & \dots & U_2^{(d)} \\ \vdots & \ddots & \vdots \end{pmatrix}. \tag{3.18}$$

Similarly, we denote $\boldsymbol{U}^{(\zeta),m} \in \mathbb{R}^m$ for $1 \le \zeta \le d$, and

$$\boldsymbol{J}_k^{(\zeta),m} := \begin{pmatrix} J_{k,1}^{(\zeta),m} \\ J_{k,2}^{(\zeta),m} \\ \vdots \\ J_{k,m}^{(\zeta),m} \end{pmatrix} \qquad \text{with} \quad 1 \le \zeta \le d, \tag{3.19}$$

where

$$J_{k,i}^{(\zeta),m} := \frac{1}{m^{k/2}} \sum_{\Xi \in \tilde{\boldsymbol{\Xi}}_i^m(k)} \left(\prod_{\ell=1}^{k} Z_{\Xi_\ell} \right) U^{(\zeta)}_{(\Xi_1)_2}, \qquad \text{with} \quad 1 \le \zeta \le d \tag{3.20}$$

(cf. (3.5)). Finally, we also consider, for $k \in \mathbb{N}$, families of independent, identically distributed (infinite) random vectors $\{\boldsymbol{J}_k^{(1)}\}_{k\in\mathbb{N}}$,...,$\{\boldsymbol{J}_k^{(d)}\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{K}_k\}_{k\in\mathbb{N}}$

$$\boldsymbol{J}_k^{(1)} := \begin{pmatrix} J_{k,1}^{(1)} \\ J_{k,2}^{(1)} \\ \vdots \end{pmatrix}, \quad \dots \quad , \boldsymbol{J}_k^{(d)} := \begin{pmatrix} J_{k,1}^{(d)} \\ J_{k,2}^{(d)} \\ \vdots \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{K}_k := \begin{pmatrix} K_{k,1} \\ K_{k,2} \\ \vdots \end{pmatrix}. \tag{3.21}$$

with $J_{k,i}^{(1)}, \dots, J_{k,i}^{(d)}, K_{k,i} \sim \mathcal{N}(0,1)$ and all independent between them.

Then, Theorem 3.2 also holds for this family as well. That is:

Proposition 3.4. *The family of random vectors* $\{(\boldsymbol{J}_k^{(1),m}, \dots, \boldsymbol{J}_k^{(d),m}, \boldsymbol{K}_k^m)\}_{k\in\mathbb{N}}$ *defined by* (3.21) *converges, in distribution, to the family* $\{(\boldsymbol{J}_k^{(1)}, \dots, \boldsymbol{J}_k^{(d)}, \boldsymbol{K}_k)\}_{k\in\mathbb{N}}$ *(see Definition 3.1).*

Proof. We can follow the same ideas as in the proof of Theorem 3.2, by interpreting (3.7) in this context. We get again that each chain must be repeated twice, and we are only interested in configurations that see a maximal amount of vertices. Now, however, chains can be repeated coming from different elements of the family, namely, $J_{k,i}^{(\zeta),m}$ and $J_{k,i}^{(\zeta'),m}$ might share a chain for $\zeta \neq \zeta'$, and still see the maximum number of vertices. Those repetitions, however, do not contribute to the expected value (3.7), since they contain a single term $U_s^{(\zeta)} U_s^{(\zeta')}$ for some $1 \le s \le m$, and $\mathbb{E}[U_s^{(\zeta)} U_s^{(\zeta')}] = 0$ for $\zeta \neq \zeta'$. □

We also see the recurrence in Proposition 3.3:

Proposition 3.5. *The random vectors* $\boldsymbol{J}_k^{(\zeta),m}$ *satisfy,*

$$\begin{aligned} m^{-\frac{1}{2}} \boldsymbol{Z}^m \boldsymbol{J}_k^{(\zeta),m} &= \boldsymbol{J}_{k+1}^{(\zeta),m} + \boldsymbol{J}_{k-1}^{(\zeta),m} + \boldsymbol{R}_k^{(\zeta),m}, && \text{if } k \in \mathbb{N} \text{ is even}, \\ m^{-\frac{1}{2}} (\boldsymbol{Z}^m)^\top \boldsymbol{J}_k^{(\zeta),m} &= \boldsymbol{J}_{k+1}^{(\zeta),m} + \boldsymbol{J}_{k-1}^{(\zeta),m} + \boldsymbol{R}_k^{(\zeta),m}, && \text{if } k \in \mathbb{N} \text{ is odd}, \end{aligned}$$

for some random vectors $\boldsymbol{R}_k^{(\zeta),m}$ *with*

$$\mathbb{E}\left[\|\boldsymbol{R}_k^{(\zeta),m}\|_{2p}^{2p}\right] \leq C_p < +\infty,$$

for any $p \in \mathbb{N}$*, and for any* $1 \leq \zeta \leq m$*, and for some* C *independent of* m *(but it might depend on* k *and* p*).*

Proof. Follows by Proposition 3.3 applied to each $\zeta \in \{1, \dots, d\}$. □

For notational convenience, we will denote

$$\boldsymbol{J}_k^m = (\boldsymbol{J}_k^{(1),m}, \dots, \boldsymbol{J}_k^{(d),m}) \in \mathbb{R}^{m \times d}. \tag{3.22}$$

Proposition 3.6 (Almost-Orthonormality property)**.** *The family of random vectors* $\{(\boldsymbol{J}_k^m, \boldsymbol{K}_k^m)\}_{k \in \mathbb{N}}$ *defined by* (3.21)-(3.22) *satisfies*

$$\mathbb{E}\left[\|\boldsymbol{O}_{JJ}^m(k_1,k_2)\|_{2p}^{2p} + \|\boldsymbol{O}_{JK}^m(k_1,k_2)\|_{2p}^{2p} + \|O_{KK}^m(k_1,k_2)\|_{2p}^{2p}\right] \leq \frac{C_p}{m},$$

for any $p \in \mathbb{N}$*, and for some* C *depending only on* $\max\{k_1, k_2\}$*,* d*, and* p*, where*[8]

$$\begin{aligned}
\mathbb{R}^{d\times d} \ni \boldsymbol{O}_{JJ}^m(k_1,k_2) &= \frac{1}{m}(\boldsymbol{J}_{k_1}^m)^\top \boldsymbol{J}_{k_2}^m - \delta_{k_1,k_2}\mathrm{Id}_d \\
\mathbb{R}^{d\times 1} \ni \boldsymbol{O}_{JK}^m(k_1,k_2) &= \frac{1}{m}(\boldsymbol{J}_{k_1}^m)^\top \boldsymbol{K}_{k_2}^m \\
\mathbb{R} \ni O_{KK}^m(k_1,k_2) &= \frac{1}{m}\boldsymbol{K}_{k_1}^m \cdot \boldsymbol{K}_{k_2}^m - \delta_{k_1,k_2},
\end{aligned}$$

for all $k_1, k_2 \in \{1, \dots, m\}$*.*

Proof. Let us show

$$\mathbb{E}\left[\left(\frac{1}{m}\boldsymbol{K}_k^m \cdot \boldsymbol{K}_k^m - 1\right)^2\right] \leq \frac{C}{m},$$

and the rest follow by analogy. We develop the square to obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{m}\boldsymbol{K}_k^m \cdot \boldsymbol{K}_k^m - 1\right)^2\right] &= \frac{1}{m^2}\sum_{i,j=1}^m \mathbb{E}\left[(K_{k,i}^m)^2 (K_{k,j}^m)^2\right] + 1 - \frac{2}{m}\sum_{i=1}^m \mathbb{E}\left[(K_{k,i}^m)^2\right] \\
&= \frac{m(m-1)}{m^2}\left(\mathbb{E}\left[(K_{k,1}^m)^2\right]\right)^2 + \frac{1}{m}\mathbb{E}\left[(K_{k,1}^m)^4\right] \\
&\quad + 1 - 2\mathbb{E}\left[(K_{k,1}^m)^2\right],
\end{aligned}$$

where we are using the symmetry in the definition of $K_{k,i}$.

From the proof of Theorem 3.2 we have that if $k > 0$

$$\left|\mathbb{E}\left[(K_{k,1}^m)^2\right] - 1\right| \leq \frac{C}{m}, \qquad \text{and} \qquad \left|\mathbb{E}\left[(K_{k,1}^m)^4\right] - 3\right| \leq \frac{C}{m},$$

from which the first result now follows. In general, again using the proof of Theorem 3.2 we have

$$\mathbb{E}\left[\left(\frac{1}{m}\boldsymbol{K}_k^m \cdot \boldsymbol{K}_k^m - 1\right)^{2p}\right] \leq \frac{C_p}{m},$$

for any $p \in \mathbb{N}$, which gives the desired result. □

[8]Here, δ_{k_1,k_2} is the Kronecker delta: $\delta_{k_1,k_2} = 1$ if $k_1 = k_2$, and $\delta_{k_1,k_2} = 0$ if $k_1 \neq k_2$.

4. Proof of the main result

Let us now proceed with the proof of our main result. Before doing so, we show an intermediary lemma on the possible growth of exchangeable vectors after multiplication by a random matrix.

For that, we need the following result on random matrices with Gaussian entries, which can be found, for example, in [42, Theorem 4.4.5].

Theorem 4.1. *Let $\boldsymbol{A}$ be a random $m \times m$ matrix with subgaussian independent entries with mean zero. Then there exists a universal constant $C > 0$ such that, for any $t > 0$,*

$$\|\boldsymbol{A}\|_M := \sup_{x \in \mathbb{S}^{m-1}} \langle \boldsymbol{A}x, x\rangle \leq CK(\sqrt{m} + t)$$

with probability at least $1 - 2e^{-t^2}$. Here $K = \max_{i,j} \|A_{i,j}\|_{\psi_2}$, where $\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left(e^{X^2/t^2}\right) \leq 2\right\}$.

Thanks to Theorem 4.1 we can prove the following, where we recall that $\boldsymbol{Z}$ denotes a random matrix with i.i.d. entries of $\mathcal{N}(0,1)$ (in this case, of size $m \times m$).

Lemma 4.2. *Suppose that $\mathfrak{U} \in \mathbb{R}^{m \times d}$ is a random exchangeable array that satisfies*

$$\mathbb{E}\left[\|\mathfrak{U}\|^2\right] \leq C_\circ m^\alpha, \qquad \text{and} \qquad \mathbb{E}\left[\|\mathfrak{U}\|^\varrho\right] \leq C_\circ m^\beta,$$

for some $\varrho > 2$, $C_\circ \geq 1$, and some $\alpha, \beta > 0$. We define

$$\mathfrak{U}' := \frac{1}{\sqrt{m}} \boldsymbol{Z}\mathfrak{U}.$$

Then, if we let $\delta > 0$ such that $\varrho > 2 + \delta$, we have

$$\mathbb{E}\left[\|\mathfrak{U}'\|^2\right] \leq C_\circ C_\varrho m^\alpha, \qquad \text{and} \qquad \mathbb{E}\left[\|\mathfrak{U}'\|^{\varrho-\delta}\right] \leq C_\circ^{\frac{\varrho-\delta}{\rho}} C_{\varrho,\delta} m^{\beta\frac{\varrho-\delta}{\varrho}},$$

for some constants $C_\varrho, C_{\varrho,\delta} > 0$ independent of m, but that might depend on ϱ, α, and β (and also on δ in the case of $C_{\varrho,\delta}$).

Proof. We implicitly consider the random elements in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define, for any $i \in \mathbb{N} \cup \{0\}$,

$$\Omega_i := \left\{\omega \in \Omega : (i+1)\|\mathfrak{U}(\omega)\|_2^2 \geq \|\mathfrak{U}'(\omega)\|_2^2 \geq i\|\mathfrak{U}(\omega)\|_2^2\right\}.$$

By Theorem 4.1, for some C independent of m and i,

$$\begin{aligned}\mathbb{P}(\Omega_i) &\leq \mathbb{P}\left(\|\boldsymbol{Z}\|_M^2 \geq im\right) \\ &= \mathbb{P}\left(\|\boldsymbol{Z}\|_M \geq CK\left(\sqrt{m} + \left(\sqrt{i}(CK)^{-1} - 1\right)\sqrt{m}\right)\right) \leq Ce^{-cim},\end{aligned} \tag{4.1}$$

for some universal constants C and c. Now observe that, by Hölder's inequality,

$$\begin{aligned}\mathbb{E}\left[\|\mathfrak{U}'\|^2\right] &= \int_\Omega \|\mathfrak{U}'(\omega)\|^2 \, d\mathbb{P}(\omega) \leq \sum_{i \geq 0}(i+1)\int_{\Omega_i} \|\mathfrak{U}(\omega)\|^2 \, d\mathbb{P}(\omega) \\ &\leq \int_{\Omega_0} \|\mathfrak{U}(\omega)\|^2 \, d\mathbb{P}(\omega) + \sum_{i \geq 1}(i+1)\left(\int_{\Omega_i} \|\mathfrak{U}(\omega)\|^\varrho \, d\mathbb{P}(\omega)\right)^{\frac{2}{\varrho}} (\mathbb{P}(\Omega_i))^{\frac{\varrho-2}{\varrho}}.\end{aligned}$$

Using the previous estimate, (4.1),

$$\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^2\right] \leq \mathbb{E}\left[\left\|\mathfrak{U}\right\|^2\right] + C\sum_{i\geq 1}(i+1)\mathbb{E}\left[\|\mathfrak{U}\|^\varrho\right]^{\frac{2}{\varrho}} e^{-cim\frac{\varrho-2}{\varrho}}.$$

From our assumptions, and for some C_ϱ that depends on ϱ,

$$\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^2\right] \leq C_\circ m^\alpha + C(C_\circ m^\beta)^{\frac{2}{\varrho}}\sum_{i\geq 1}(i+1)e^{-cim\frac{\varrho-2}{\varrho}} = C_\circ\left(m^\alpha + C_\varrho m^{\frac{2\beta}{\varrho}} e^{-cm\frac{\varrho-2}{\varrho}}\right),$$

and hence

$$\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^2\right] \leq C_\circ C_\varrho m^\alpha,$$

for some possibly different C_ϱ.

On the other hand, following the same strategy we get:

$$\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^{\varrho-\delta}\right] \leq C\sum_{i\geq 0}(i+1)\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^\varrho\right]^{\frac{\varrho-\delta}{\varrho}} e^{-cim\frac{\delta}{\varrho}}.$$

From the assumptions again,

$$\mathbb{E}\left[\left\|\mathfrak{U}'\right\|^{\varrho-\delta}\right] \leq C_\circ^{\frac{\rho-\delta}{\rho}} C_{\varrho,\delta} m^{\beta\frac{\varrho-\delta}{\varrho}}\sum_{i\geq 0}(i+1)e^{-cim\frac{\delta}{\varrho}} \leq C_\circ^{\frac{\rho-\delta}{\rho}} C_{\varrho,\delta} m^{\beta\frac{\varrho-\delta}{\varrho}},$$

so we get the desired result. □

We can now prove the main result, Theorem 2.2:

Proof of Theorem 2.2. We use the notation from Section 3.2, in particular, the random arrays $(\boldsymbol{J}_k^{(1),m},\dots,\boldsymbol{J}_k^{(d),m})$ and $\boldsymbol{K}_k^m$, defined by (3.1)-(3.4)-(3.5)-(3.18)-(3.19)-(3.20), with $\boldsymbol{U}^{(1\dots d)}$ and $\boldsymbol{V}$ taken to be $\boldsymbol{U}(0)$ and $\boldsymbol{V}(0)$.

We divide the proof into seven steps.

Step 1: The structure. For notational convenience, we drop the subscript $\tau > 0$ and the superscript m, which will be implicit in the following variables; also, we denote by $\boldsymbol{J}_k = (\boldsymbol{J}_k^{(1),m},\dots,\boldsymbol{J}_k^{(d),m}) \in \mathbb{R}^{m\times d}$. We show by induction that we can write

$$\begin{aligned}\boldsymbol{U}(\kappa) &= \mathring{\boldsymbol{U}}(\kappa) + \mathfrak{U}(\kappa)\\ \boldsymbol{W}(\kappa) &= \mathring{\boldsymbol{W}}(\kappa) + \mathfrak{W}(\kappa)\\ \boldsymbol{V}(\kappa) &= \mathring{\boldsymbol{V}}(\kappa) + \mathfrak{V}(\kappa),\end{aligned} \tag{4.2}$$

with

$$\begin{aligned}\mathring{\boldsymbol{U}}(\kappa) &= \sum_{k\geq 0}\left(\boldsymbol{J}_k\boldsymbol{\alpha}_k(\kappa) + \boldsymbol{K}_k\overline{\boldsymbol{\alpha}}_k(\kappa)\right)\\ \mathring{\boldsymbol{W}}(\kappa) &= \sum_{i,j\geq 0}\left(\boldsymbol{J}_i\boldsymbol{\gamma}_{ij}(\kappa)\boldsymbol{J}_j^\top + \boldsymbol{K}_i\overline{\gamma}_{ij}(\kappa)\boldsymbol{K}_j^\top + \boldsymbol{J}_i\hat{\boldsymbol{\gamma}}_{ij}(\kappa)\boldsymbol{K}_j^\top + \boldsymbol{K}_i\hat{\overline{\boldsymbol{\gamma}}}_{ij}(\kappa)\boldsymbol{J}_j^\top\right)\\ \mathring{\boldsymbol{V}}(\kappa) &= \sum_{k\geq 0}\left(\boldsymbol{J}_k\boldsymbol{\beta}_k(\kappa) + \boldsymbol{K}_k\overline{\beta}_k(\kappa)\right),\end{aligned} \tag{4.3}$$

and where

$$\mathfrak{U}(\kappa)\in\mathbb{R}^{m\times d}, \mathfrak{W}(\kappa)\in\mathbb{R}^{m\times m}, \mathfrak{V}(\kappa)\in\mathbb{R}^m \tag{4.4}$$

satisfy

$$\mathbb{E}\left[\|\mathfrak{U}(\kappa)\|^2+\frac{1}{m}\|\mathfrak{W}(\kappa)\|^2+\|\mathfrak{V}(\kappa)\|^2\right]\leq Cm^{\frac{1}{2}}, \tag{4.5}$$

for some C depending on κ but independent of $m\in\mathbb{N}$. Moreover,

$$\begin{array}{ll}
\boldsymbol{\alpha}_k(\kappa)\in\mathbb{R}^{d\times d}, & \boldsymbol{\alpha}_k(\kappa)=0 \quad \text{if } k \text{ is odd},\\
\overline{\boldsymbol{\alpha}}_k(\kappa)\in\mathbb{R}^{1\times d}, & \overline{\boldsymbol{\alpha}}_k(\kappa)=0 \quad \text{if } k \text{ is even},\\
\boldsymbol{\beta}_k(\kappa)\in\mathbb{R}^{d\times 1}, & \boldsymbol{\beta}_k(\kappa)=0 \quad \text{if } k \text{ is even},\\
\overline{\beta}_k(\kappa)\in\mathbb{R}, & \overline{\beta}_k(\kappa)=0 \quad \text{if } k \text{ is odd},\\
\boldsymbol{\gamma}_{ij}(\kappa)\in\mathbb{R}^{d\times d}, & \boldsymbol{\gamma}_{ij}(\kappa)=0 \quad \text{if } i \text{ is even } \underline{\text{or}} \text{ } j \text{ is odd},\\
\overline{\gamma}_{ij}(\kappa)\in\mathbb{R}, & \overline{\gamma}_{ij}(\kappa)=0 \quad \text{if } i \text{ is odd } \underline{\text{or}} \text{ } j \text{ is even},\\
\hat{\boldsymbol{\gamma}}_{ij}(\kappa)\in\mathbb{R}^{d\times 1}, & \hat{\boldsymbol{\gamma}}_{ij}(\kappa)=0 \quad \text{if } i \text{ is even } \underline{\text{or}} \text{ } j \text{ is even},\\
\hat{\mathring{\boldsymbol{\gamma}}}_{ij}(\kappa)\in\mathbb{R}^{1\times d}, & \hat{\mathring{\boldsymbol{\gamma}}}_{ij}(\kappa)=0 \quad \text{if } i \text{ is odd } \underline{\text{or}} \text{ } j \text{ is odd}.
\end{array} \tag{4.6}$$

Step 2: Computing the update. Let us compute $(\boldsymbol{U}(\kappa+1),\boldsymbol{W}(\kappa+1),\boldsymbol{V}(\kappa+1))$ in terms of (4.2)-(4.3), by using (2.5). By the inductive assumption, we will assume that (4.6) holds at time κ.

We compute first $h_\kappa(x)$, by expanding:

$$\boldsymbol{p}(\kappa):=\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}+\frac{1}{m}\boldsymbol{W}(\kappa)\right]\boldsymbol{U}(\kappa).$$

On the one hand, thanks to (4.2)-(4.3)-(4.6) and the recursion property in Proposition 3.5, we have

$$\frac{1}{\sqrt{m}}\boldsymbol{Z}\boldsymbol{U}(\kappa)=\sum_{k\geq 0}\left(\left[\boldsymbol{J}_{k+1}+\boldsymbol{J}_{k-1}\right]\boldsymbol{\alpha}_k(\kappa)+\left[\boldsymbol{K}_{k+1}+\boldsymbol{K}_{k-1}\right]\overline{\boldsymbol{\alpha}}_k(\kappa)\right)+\mathfrak{E}_1(\kappa),$$

where

$$\mathfrak{E}_1(\kappa)=\sum_{k\geq 0}\left(\boldsymbol{R}_k\boldsymbol{\alpha}_k(\kappa)+\boldsymbol{S}_k\overline{\boldsymbol{\alpha}}_k(\kappa)\right)+\frac{1}{\sqrt{m}}\boldsymbol{Z}\mathfrak{U}(\kappa),$$

and where from now on we assume that whenever an index is negative, the corresponding object is identically zero (e.g. $\boldsymbol{J}_{-1}\equiv 0$ and $\boldsymbol{K}_{-1}\equiv 0$), and we have denoted (from Proposition 3.5), $\boldsymbol{R}_k=(\boldsymbol{R}_k^{(1),m},\dots,\boldsymbol{R}_k^{(d),m})$.

On the other hand, also from (4.2)-(4.3), we can compute the other term in $\boldsymbol{p}$ by using the orthonormality property in Proposition 3.6,

$$\begin{aligned}\frac{1}{m}\boldsymbol{W}(\kappa)\boldsymbol{U}(\kappa)=&\sum_{i,j\geq 0}\left(\boldsymbol{J}_i\boldsymbol{\gamma}_{ij}(\kappa)\boldsymbol{\alpha}_j(\kappa)+\boldsymbol{K}_i\overline{\gamma}_{ij}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)\right)\\&+\sum_{i,j\geq 0}\left(\boldsymbol{J}_i\hat{\boldsymbol{\gamma}}_{ij}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)+\boldsymbol{K}_i\hat{\mathring{\boldsymbol{\gamma}}}_{ij}(\kappa)\boldsymbol{\alpha}_j(\kappa)\right)+\mathfrak{E}_2(\kappa)\end{aligned}$$

where

$$\begin{aligned}\mathfrak{E}_2(\kappa) = \frac{1}{m}\Big(\mathfrak{W}(\kappa)\mathring{U}(\kappa) + \mathring{W}(\kappa)\mathfrak{U}(\kappa) + \mathfrak{W}(\kappa)\mathfrak{U}(\kappa)\Big) + \\ + \sum_{i,j,k\geq 0}\left[\boldsymbol{J}_i\boldsymbol{\gamma}_{ij}(\kappa) + \boldsymbol{K}_i\hat{\mathring{\boldsymbol{\gamma}}}_{ij}(\kappa)\right]\left[\boldsymbol{O}_{JJ}(j,k)\boldsymbol{\alpha}_k(\kappa) + \boldsymbol{O}_{JK}(j,k)\overline{\boldsymbol{\alpha}}_k(\kappa)\right] \\ + \sum_{i,j,k\geq 0}\left[\boldsymbol{J}_i\hat{\boldsymbol{\gamma}}_{ij}(\kappa) + \boldsymbol{K}_i\overline{\gamma}_{ij}(\kappa)\right]\left[(\boldsymbol{O}_{JK}(k,j))^\top\boldsymbol{\alpha}_k(\kappa) + O_{KK}(j,k)\overline{\boldsymbol{\alpha}}_k(\kappa)\right],\end{aligned}$$

and thus

$$\boldsymbol{p}(\kappa) = \mathring{\boldsymbol{p}}(\kappa) + \mathfrak{E}_p(\kappa) := \mathring{\boldsymbol{p}}(\kappa) + \mathfrak{E}_1(\kappa) + \mathfrak{E}_2(\kappa),$$

where

$$\begin{aligned}\mathring{\boldsymbol{p}}(\kappa) = \sum_{k\geq 0}\boldsymbol{J}_k\Big(\boldsymbol{\alpha}_{k+1}(\kappa) + \boldsymbol{\alpha}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\boldsymbol{\gamma}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa) + \hat{\boldsymbol{\gamma}}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)\right]\Big) \\ + \sum_{k\geq 0}\boldsymbol{K}_k\Big(\overline{\boldsymbol{\alpha}}_{k+1}(\kappa) + \overline{\boldsymbol{\alpha}}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\overline{\gamma}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa) + \hat{\mathring{\boldsymbol{\gamma}}}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa)\right]\Big).\end{aligned}$$

From here, using again the orthonormality property in Proposition 3.6, we can compute:

$$h_\kappa(x) = \frac{1}{m}(\boldsymbol{V}(\kappa))^\top\boldsymbol{p}(\kappa)x = \mathring{h}_\kappa(x) + \mathfrak{E}_h(\kappa)x$$

with

$$\begin{aligned}\mathring{h}_\kappa(x) := \sum_{k\geq 0}(\boldsymbol{\beta}_k)^\top\Big(\boldsymbol{\alpha}_{k+1}(\kappa) + \boldsymbol{\alpha}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\boldsymbol{\gamma}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa) + \hat{\boldsymbol{\gamma}}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)\right]\Big)x \\ + \sum_{k\geq 0}\overline{\beta}_k\Big(\overline{\boldsymbol{\alpha}}_{k+1}(\kappa) + \overline{\boldsymbol{\alpha}}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\overline{\gamma}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa) + \hat{\mathring{\boldsymbol{\gamma}}}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa)\right]\Big)x\end{aligned}$$

and

$$\begin{aligned}\mathfrak{E}_h(\kappa) = \frac{1}{m}(\mathring{\boldsymbol{V}}(\kappa))^\top\mathfrak{E}_p(\kappa) + \frac{1}{m}(\mathfrak{V}(\kappa))^\top\mathring{\boldsymbol{p}}(\kappa) + \frac{1}{m}(\mathfrak{V}(\kappa))^\top\mathfrak{E}_p(\kappa) \\ + \sum_{i,k\geq 0}\Big[(\boldsymbol{\beta}_i(\kappa))^\top\boldsymbol{O}_{JJ}(i,k) + \overline{\beta}_i(\kappa)(\boldsymbol{O}_{JK}(k,i))^\top\Big]\cdot \\ \cdot\Big(\boldsymbol{\alpha}_{k+1}(\kappa) + \boldsymbol{\alpha}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\boldsymbol{\gamma}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa) + \hat{\boldsymbol{\gamma}}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)\right]\Big) \\ + \sum_{i,k\geq 0}\Big[(\boldsymbol{\beta}_i(\kappa))^\top\boldsymbol{O}_{JK}(i,k) + \overline{\beta}_i(\kappa)O_{KK}(i,k)\Big]\cdot \\ \cdot\Big(\overline{\boldsymbol{\alpha}}_{k+1}(\kappa) + \overline{\boldsymbol{\alpha}}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\overline{\gamma}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa) + \hat{\mathring{\boldsymbol{\gamma}}}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa)\right]\Big).\end{aligned}$$

At this point it is important to notice that the expression for $\mathring{h}_\kappa(x)$ is independent of the basis, and thus, if m is large enough and κ is fixed, it is independent of m.

We can also write an expression for $\boldsymbol{V}(\kappa+1)$ using (2.5) directly, where it is easy to check that $\boldsymbol{V}(\kappa+1)$ can be written in the form (4.2)-(4.3) with coefficients satisfying (4.6) by induction.

Using a similar procedure (thanks to Propositions 3.6 and 3.5) we find the expression for

$$\boldsymbol{q}(\kappa) := \left[\frac{1}{\sqrt{m}}\boldsymbol{Z} + \frac{1}{m}\boldsymbol{W}(\kappa)\right]^\top \boldsymbol{V}(\kappa) = \mathring{\boldsymbol{q}}(\kappa) + \mathfrak{E}_q(\kappa) := \mathring{\boldsymbol{q}}(\kappa) + \mathfrak{E}_3(\kappa) + \mathfrak{E}_4(\kappa),$$

where

$$\begin{aligned}\mathring{\boldsymbol{q}}(\kappa) := &\sum_{k\geq 0}\boldsymbol{J}_k\left(\boldsymbol{\beta}_{k+1}(\kappa) + \boldsymbol{\beta}_{k-1}(\kappa) + \sum_{j\geq 0}\left[(\boldsymbol{\gamma}_{jk}(\kappa))^\top\boldsymbol{\beta}_j(\kappa) + (\hat{\mathring{\boldsymbol{\gamma}}}_{jk}(\kappa))^\top\overline{\beta}_j(\kappa)\right]\right)\\ &+\sum_{k\geq 0}\boldsymbol{K}_k\left(\overline{\beta}_{k+1}(\kappa) + \overline{\beta}_{k-1}(\kappa) + \sum_{j\geq 0}\left[\overline{\gamma}_{jk}(\kappa)\overline{\beta}_j(\kappa) + (\hat{\boldsymbol{\gamma}}_{jk}(\kappa))^\top\boldsymbol{\beta}_j(\kappa)\right]\right),\end{aligned}$$

and, as before, we have

$$\mathfrak{E}_3(\kappa) = \sum_{k\geq 0}\left(\boldsymbol{R}_k\boldsymbol{\beta}_k(\kappa) + \boldsymbol{S}_k\overline{\beta}_k(\kappa)\right) + \frac{1}{\sqrt{m}}\boldsymbol{Z}^\top\mathfrak{V}(\kappa),$$

and

$$\begin{aligned}\mathfrak{E}_4(\kappa) = &\frac{1}{m}\left((\mathfrak{W}(\kappa))^\top\mathring{\boldsymbol{V}}(\kappa) + (\mathring{\boldsymbol{W}}(\kappa))^\top\mathfrak{V}(\kappa) + (\mathfrak{W}(\kappa))^\top\mathfrak{V}(\kappa)\right) +\\ &+\sum_{i,j,k\geq 0}\left[\boldsymbol{J}_i(\boldsymbol{\gamma}_{ji}(\kappa))^\top + \boldsymbol{K}_i(\hat{\boldsymbol{\gamma}}_{ji}(\kappa))^\top\right]\left[\boldsymbol{O}_{JJ}(j,k)\beta_k(\kappa) + \boldsymbol{O}_{JK}(j,k)\overline{\beta}_k(\kappa)\right]\\ &+\sum_{i,j,k\geq 0}\left[\boldsymbol{J}_i(\hat{\mathring{\boldsymbol{\gamma}}}_{ji}(\kappa))^\top + \boldsymbol{K}_i\overline{\gamma}_{ji}(\kappa)\right]\left[(\boldsymbol{O}_{JK}(k,j))^\top\boldsymbol{\beta}_k(\kappa) + O_{KK}(j,k)\overline{\beta}_k(\kappa)\right].\end{aligned}$$

Step 3: The evolution. We can now use the expressions for $\boldsymbol{p}(\kappa)$, $\boldsymbol{q}(\kappa)$, and $\boldsymbol{V}(\kappa)x^\top(\boldsymbol{U}(\kappa))^\top$ to derive an evolution for the coefficients (4.6) from (2.5). In order to do that, let us observe that we can denote

$$\mathbb{R}^d \ni \boldsymbol{\xi}_\kappa = \int x\,\mathcal{L}'(h_\kappa(x), y)d\rho_\kappa(x,y) = \mathring{\boldsymbol{\xi}}_\kappa + \mathfrak{E}_\xi(\kappa),$$

with

$$\mathring{\boldsymbol{\xi}}_\kappa := \int x\,\mathcal{L}'(\mathring{h}_\kappa(x), y)d\rho_\kappa(x,y)$$

and

$$\mathfrak{E}_\xi(\kappa) := \int x\,\left(\mathcal{L}'(h_\kappa(x),y) - \mathcal{L}'(\mathring{h}_\kappa(x),y))\right)d\rho_\kappa(x,y),$$

so that, since $\mathcal{L}'$ is Lipschitz and (2.2) holds,

$$|\mathfrak{E}_\xi(\kappa)| \leq C\|\mathfrak{E}_h(\kappa)\|\int |x|^2\,d\rho_\kappa(x,y) \leq C\|\mathfrak{E}_h(\kappa)\|.$$

If we now denote

$$\delta\boldsymbol{\alpha}_k(\kappa) := \frac{1}{\tau}\left(\boldsymbol{\alpha}_k(\kappa+1) - \boldsymbol{\alpha}_k(\kappa)\right)$$

(analogously for $\overline{\boldsymbol{\alpha}}_k, \boldsymbol{\beta}_k, \overline{\beta}_k, \boldsymbol{\gamma}_{ij}, \overline{\gamma}_{ij}, \hat{\boldsymbol{\gamma}}_{ij}, \hat{\overline{\boldsymbol{\gamma}}}_{ij}$) we get, on the one hand,[9]

$$\begin{aligned}
\delta\boldsymbol{\alpha}_k(\kappa) &= -\Big(\boldsymbol{\beta}_{k+1}(\kappa) + \boldsymbol{\beta}_{k-1}(\kappa) + \sum_{j\geq 0}\big[(\boldsymbol{\gamma}_{jk}(\kappa))^\top\boldsymbol{\beta}_j(\kappa) + (\hat{\boldsymbol{\gamma}}_{jk}(\kappa))^\top\overline{\beta}_j(\kappa)\big]\Big)\mathring{\boldsymbol{\xi}}_\kappa^\top,\\
\delta\overline{\boldsymbol{\alpha}}_k(\kappa) &= -\Big(\overline{\beta}_{k+1}(\kappa) + \overline{\beta}_{k-1}(\kappa) + \sum_{j\geq 0}\big[\overline{\gamma}_{jk}(\kappa)\overline{\beta}_j(\kappa) + (\hat{\boldsymbol{\gamma}}_{jk}(\kappa))^\top\boldsymbol{\beta}_j(\kappa)\big]\Big)\mathring{\boldsymbol{\xi}}_\kappa^\top,\\
\delta\boldsymbol{\beta}_k(\kappa) &= -\Big(\boldsymbol{\alpha}_{k+1}(\kappa) + \boldsymbol{\alpha}_{k-1}(\kappa) + \sum_{j\geq 0}\big[\boldsymbol{\gamma}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa) + \hat{\gamma}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa)\big]\Big)\mathring{\boldsymbol{\xi}}_\kappa,\\
\delta\overline{\beta_k}(\kappa) &= -\Big(\overline{\boldsymbol{\alpha}}_{k+1}(\kappa) + \overline{\boldsymbol{\alpha}}_{k-1}(\kappa) + \sum_{j\geq 0}\big[\overline{\gamma}_{kj}(\kappa)\overline{\boldsymbol{\alpha}}_j(\kappa) + \hat{\overline{\boldsymbol{\gamma}}}_{kj}(\kappa)\boldsymbol{\alpha}_j(\kappa)\big]\Big)\mathring{\boldsymbol{\xi}}_\kappa,
\end{aligned} \tag{4.7}$$

and on the other hand, from (2.5) and (4.2)-(4.3) we immediately have

$$\begin{aligned}
\delta\boldsymbol{\gamma}_{ij}(\kappa) &= -\boldsymbol{\beta}_i(\kappa)\mathring{\boldsymbol{\xi}}_\kappa^\top\boldsymbol{\alpha}_j(\kappa)^\top, \quad & \delta\hat{\boldsymbol{\gamma}}_{ij}(\kappa) &= -\boldsymbol{\beta}_i(\kappa)(\mathring{\boldsymbol{\xi}}_\kappa)^\top(\overline{\boldsymbol{\alpha}}_j(\kappa))^\top,\\
\delta\overline{\gamma}_{ij}(\kappa) &= -\overline{\beta}_i(\kappa)\mathring{\boldsymbol{\xi}}_\kappa^\top(\overline{\boldsymbol{\alpha}}_j(\kappa))^\top, \quad & \delta\hat{\overline{\boldsymbol{\gamma}}}_{ij}(\kappa) &= -\overline{\beta}_i(\kappa)\mathring{\boldsymbol{\xi}}_\kappa^\top(\boldsymbol{\alpha}_j(\kappa))^\top,
\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}
\delta\mathfrak{U}(\kappa) &= -\mathfrak{E}_q(\kappa)\mathring{\boldsymbol{\xi}}_\kappa^\top - \mathring{\boldsymbol{q}}(\kappa)\mathfrak{E}_\xi^\top(\kappa) - \mathfrak{E}_q(\kappa)\mathfrak{E}_\xi^\top(\kappa)\\
\delta\mathfrak{W}(\kappa) &= -\boldsymbol{V}(\kappa)\boldsymbol{\xi}_\kappa^\top(\boldsymbol{U}(\kappa))^\top + \mathring{\boldsymbol{V}}(\kappa)\mathring{\boldsymbol{\xi}}_\kappa^\top(\mathring{\boldsymbol{U}}(\kappa))^\top\\
\delta\mathfrak{V}(\kappa) &= -\mathfrak{E}_p(\kappa)\mathring{\boldsymbol{\xi}}_\kappa - \mathring{\boldsymbol{p}}(\kappa)\mathfrak{E}_\xi(\kappa) - \mathfrak{E}_p(\kappa)\mathfrak{E}_\xi(\kappa).
\end{aligned} \tag{4.9}$$

It is now a simple check that (4.7)-(4.8) imply that, if the relations in (4.6) hold at time κ, they also hold at time $\kappa+1$.

Step 4: Initial conditions and boundedness of coefficients. We are considering the vectors $\boldsymbol{J}_k$ and $\boldsymbol{K}_k$ to be the ones constructed in subsection 3.2 where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the initializations $\boldsymbol{U}(0)$ and $\boldsymbol{V}(0)$. Thus, by construction,

$$\begin{array}{lll}
\boldsymbol{\alpha}_0(0) = \mathrm{Id}_d, & \boldsymbol{\alpha}_k(0) = \mathbf{0}_{d\times d} & \text{for} \quad k\geq 1,\\
\overline{\boldsymbol{\alpha}}_k(0) = \mathbf{0}_{1\times d}, & & \text{for} \quad k\geq 0,\\
\boldsymbol{\beta}_k(0) = \mathbf{0}_{d\times 1}, & & \text{for} \quad k\geq 0,\\
\overline{\beta}_0(0) = 1, & \overline{\beta}_k(0) = 0 & \text{for} \quad k\geq 1,
\end{array} \tag{4.10}$$

and all $\boldsymbol{\gamma}$, $\overline{\gamma}$, $\hat{\boldsymbol{\gamma}}$, and $\hat{\overline{\boldsymbol{\gamma}}}$ are initialized at 0. From the update (4.7), we immediately get that

$$\|\boldsymbol{\alpha}_k(\kappa)\| = \|\overline{\boldsymbol{\alpha}}_k(\kappa)\| = \|\boldsymbol{\beta}_k(\kappa)\| = \|\overline{\beta}_k(\kappa)\| = 0 \qquad \text{if} \quad k\geq\kappa+1, \tag{4.11}$$

and

$$\|\boldsymbol{\gamma}_{ij}(\kappa)\| = \|\overline{\gamma}_{ij}(\kappa)\| = \|\hat{\gamma}_{ij}(\kappa)\| = \|\hat{\overline{\boldsymbol{\gamma}}}_{ij}(\kappa)\| = 0 \qquad \text{if} \quad i\geq\kappa+1 \quad \text{or} \quad j\geq\kappa+1, \tag{4.12}$$

[9]Observe that the evolution of the system is "mostly" independent of m, and hence for m very large we have a trivial limit: the only problem is when all the elements in the corresponding arrays are non-zero, due to a boundary effect at $k=m$, but this is avoided for m large enough and after a fixed number of iterations thanks to the initialization (4.10).

and in particular, if m is large enough, the coefficients are all independent of m. This is because in the evolution (4.7)-(4.8) each element in a position k is only affected by elements in the surrounding positions (either for α, β, or γ).

Observe that, again by construction, the evolution of the coefficients (4.6) given by (4.7)-(4.8) is deterministic, and in particular all the coefficients are always bounded after finitely many time steps by a universal constant depending only on κ_* by (4.11)-(4.12) (and τ, but independent of m), and the same holds for $\mathring{\boldsymbol{\xi}}$ and $\mathring{h}$:

$$\begin{aligned}\|\boldsymbol{\alpha}_k(\kappa)\| + \|\overline{\boldsymbol{\alpha}}_k(\kappa)\| + \|\boldsymbol{\beta}_k(\kappa)\| + \|\overline{\beta}_k(\kappa)\| + \\ + \|\boldsymbol{\gamma}_{ij}(\kappa)\| + \|\overline{\boldsymbol{\gamma}}_{ij}(\kappa)\| + \|\hat{\gamma}_{ij}(\kappa)\| + \|\hat{\hat{\boldsymbol{\gamma}}}_{ij}(\kappa)\| + |\mathring{h}_\kappa| + \|\mathring{\boldsymbol{\xi}}_\kappa\| \leq C_{\kappa_*}\end{aligned} \tag{4.13}$$

for all $k, i, j \in \mathbb{N}$, and $1 \leq \kappa \leq \kappa_*$.

Hence, from (4.3) and by the proof of Theorem 3.2, we also have that

$$\mathbb{E}[|\mathring{U}_{i,\ell}(\kappa)|^\Upsilon + |\mathring{W}_{i,j}(\kappa)|^\Upsilon + |\mathring{V}_i(\kappa)|^\Upsilon + |\mathring{p}_{i,\ell}(\kappa)|^\Upsilon + |\mathring{q}_i(\kappa)|^\Upsilon] \leq C_{\kappa_*,\Upsilon} < +\infty, \tag{4.14}$$

for any $\Upsilon \geq 2$, $1 \leq i, j \leq m$, $1 \leq \ell \leq m$, and for some $C_{\kappa_*,\Upsilon}$ independent of m.

In particular, we have that

$$\begin{aligned}\mathbb{E}[\|\mathring{\boldsymbol{U}}(\kappa)\|^\Upsilon + \|\mathring{\boldsymbol{V}}(\kappa)\|^\Upsilon + \|\mathring{\boldsymbol{p}}(\kappa)\|^\Upsilon + \|\mathring{\boldsymbol{q}}(\kappa)\|^\Upsilon] &\leq C_{\kappa_*,\Upsilon} m^{\frac{\Upsilon}{2}}, \\ \mathbb{E}[\|\mathring{\boldsymbol{W}}(\kappa)\|^\Upsilon] &\leq C_{\kappa_*,\Upsilon} m^{\Upsilon}.\end{aligned} \tag{4.15}$$

Let us now bound the error terms. Let us assume that we have for some $\alpha \geq 0$ that will be small, and for any $\varrho \geq 2$,

$$\begin{aligned}\mathbb{E}[\|\mathfrak{U}(\kappa)\|^\varrho + \|\mathfrak{V}(\kappa)\|^\varrho] &\leq C_\varrho m^{\frac{\varrho}{2}-1+\alpha} \\ \mathbb{E}[\|\mathfrak{W}(\kappa)\|^\varrho] &\leq C_\varrho m^{\varrho-1+\alpha}.\end{aligned} \tag{4.16}$$

Then we will show that for any $\delta > 0$

$$\begin{aligned}\mathbb{E}[\|\mathfrak{U}(\kappa+1)\|^\varrho + \|\mathfrak{V}(\kappa+1)\|^\varrho] &\leq C'_{\varrho,\delta} m^{\frac{\varrho}{2}-1+\alpha+\delta} \\ \mathbb{E}[\|\mathfrak{W}(\kappa+1)\|^\varrho] &\leq C'_{\varrho,\delta} m^{\varrho-1+\alpha+\delta},\end{aligned} \tag{4.17}$$

for some new constant $C'_{\varrho,\delta}$ that might depend on everything, but it is independent of m.

In order to do that, we look at the different terms in the errors. We can always apply the same strategy to bound them, using our hypotheses (4.16) and that we know explicitly how the errors are being updated, (4.9). Let us for example show how to bound a representative case that includes all possible behaviors, to obtain a bound like (4.17) for the term $\mathfrak{E}_p(\kappa+1)$. Namely, we will show that

$$\mathbb{E}[\|\mathfrak{E}_p(\kappa+1)\|^\varrho] \leq C' m^{\frac{\varrho}{2}-1+\alpha+\delta} \tag{4.18}$$

for some δ arbitrarily small.

We know that $\mathfrak{E}_p(\kappa+1) = \mathfrak{E}_1(\kappa+1) + \mathfrak{E}_2(\kappa+1)$, let us control them separately.

Step 5: Bound on $\mathfrak{E}_1(\kappa+1)$. The term $\mathfrak{E}_1(\kappa+1)$ has two parts. The first part is

$$\sum_{k\geq 0} \left(\boldsymbol{R}_k \boldsymbol{\alpha}_k(\kappa) + \boldsymbol{S}_k \overline{\boldsymbol{\alpha}}_k(\kappa)\right).$$

In this case, we use the boundedness of coefficients (4.13) together with the fact that, from Proposition 3.5, we also have that for $k \leq \kappa_* + 1$ and any $\Upsilon \geq 2$,

$$\mathbb{E}[\|\boldsymbol{R}_k\|^\Upsilon + \|\boldsymbol{S}_k\|^\Upsilon] \leq C_{\kappa_*,\Upsilon} m^{\frac{\Upsilon}{2}-1},$$

(using the equivalence between Euclidean norms, $\|x\|_p \leq m^{\frac{1}{p}-\frac{1}{q}}\|x\|_q$ for any $x \in \mathbb{R}^m$ and $p < q$). This gives the desired result without losing any power.

For the second term in $\mathfrak{E}_1(\kappa+1)$,

$$\frac{1}{\sqrt{m}} \boldsymbol{Z}\mathfrak{U}(\kappa),$$

we can apply Lemma 4.2 to obtain, on the one hand,

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{m}} \boldsymbol{Z}\mathfrak{U}(\kappa)\right\|^2\right] \leq Cm^\alpha$$

and on the other hand, for any $r > 2$ and $\delta > 0$,

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{m}} \boldsymbol{Z}\mathfrak{U}(\kappa)\right\|^r\right] \leq Cm^{\frac{r}{2}-1+\frac{\delta}{r}+\frac{r}{r+\delta}\alpha} \leq Cm^{\frac{r}{2}-1+\delta+\alpha}$$

where C now might depend also on δ.

Step 6: Bound on $\mathfrak{E}_2(\kappa+1)$. The term $\mathfrak{E}_2(\kappa+1)$ also has two parts. Regarding all the terms involving the orthonormal errors coming from Proposition 3.6, we treat them as in the previous step but using Proposition 3.6 instead of Proposition 3.5. Let us then show how to bound the remaining term,

$$\frac{1}{m}\left(\mathfrak{W}(\kappa)\mathring{\boldsymbol{U}}(\kappa) + \mathring{\boldsymbol{W}}(\kappa)\mathfrak{U}(\kappa) + \mathfrak{W}(\kappa)\mathfrak{U}(\kappa)\right).$$

We do so separately, for each of the three elements. Let us start with the first term:, by means of Cauchy–Schwarz and Hölder:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{m}\mathfrak{W}(\kappa)\mathring{\boldsymbol{U}}(\kappa)\right\|^\varrho\right] &\leq \frac{1}{m^\varrho}\mathbb{E}\left[\|\mathfrak{W}(\kappa)\|^\varrho \left\|\mathring{\boldsymbol{U}}(\kappa)\right\|^\varrho\right] \\ &\leq \frac{1}{m^\varrho}\left(\mathbb{E}\left[\|\mathfrak{W}(\kappa)\|^{(1+\varepsilon)\varrho}\right]\right)^{\frac{1}{1+\varepsilon}}\left(\mathbb{E}\left[\left\|\mathring{\boldsymbol{U}}(\kappa)\right\|^{\eta\varrho}\right]\right)^{\frac{1}{\eta}},\end{aligned}$$

with $\frac{\varepsilon}{1+\varepsilon} = \frac{1}{\eta}$. By hypothesis (4.16) and using (4.15) we obtain

$$\mathbb{E}\left[\left\|\frac{1}{m}\mathfrak{W}(\kappa)\mathring{\boldsymbol{U}}(\kappa)\right\|^\varrho\right] \leq C\frac{1}{m^\varrho}m^{\varrho-1+\frac{\varepsilon+\alpha}{1+\varepsilon}}m^{\frac{\varrho}{2}} = Cm^{\frac{\varrho}{2}-1+\frac{\varepsilon+\alpha}{1+\varepsilon}}.$$

A similar computation works for the term $\frac{1}{m}\mathring{\boldsymbol{W}}(\kappa)\mathfrak{U}(\kappa)$. Finally,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{m}\mathfrak{W}(\kappa)\mathfrak{U}(\kappa)\right\|^\varrho\right] &\leq \frac{1}{m^\varrho}\mathbb{E}\left[\|\mathfrak{W}(\kappa)\|^\varrho\,\|\mathfrak{U}(\kappa)\|^\varrho\right] \\ &\leq \frac{1}{m^\varrho}\left(\mathbb{E}\left[\|\mathfrak{W}(\kappa)\|^{2\varrho}\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\|\mathfrak{U}(\kappa)\|^{2\varrho}\right]\right)^{\frac{1}{2}}.\end{aligned}$$

Using our hypotheses in (4.16) we have

$$\mathbb{E}\left[\left\|\frac{1}{m}\mathfrak{W}(\kappa)\mathfrak{U}(\kappa)\right\|^\varrho\right] \leq C\frac{1}{m^\varrho}m^{\varrho+\frac{-1+\alpha}{2}}m^{\frac{\varrho}{2}+\frac{-1+\alpha}{2}} \leq Cm^{\frac{\varrho}{2}-1+\alpha}.$$

Thus, assuming $\varepsilon < \delta$, we have shown that (4.18) holds.

We can do the same with all other terms in $\mathfrak{U}(\kappa+1)$ and $\mathfrak{V}(\kappa+1)$ to obtain the desired result, and a completely analogous argument also works on $\mathfrak{W}(\kappa+1)$.

Step 7: Conclusion. For $\kappa = 0$, there are no error terms, and in particular (4.16) holds with $\alpha = 0$ (recall $\alpha \geq 0$). We fix δ universally as $\delta = \frac{1}{2\kappa_*}$, in such a way that, from (4.16)-(4.17) with $\varrho = 2$,

$$\mathbb{E}[\|\mathfrak{U}(\kappa)\|^2 + \|\mathfrak{V}(\kappa)\|^2] \leq Cm^{\frac{1}{2}} \qquad \text{and} \qquad \mathbb{E}[\|\mathfrak{W}(\kappa)\|^2] \leq Cm^{\frac{3}{2}},$$

for all $\kappa \leq \kappa_*$ (notice that taking δ smaller, we can make the powers of m arbitrarily close to 1 and 2 respectively). In particular, by exchangeability of $\mathfrak{U}$, $\mathfrak{V}$, and $\mathfrak{W}$, we have that

$$\mathfrak{U}_{i,\ell}(\kappa), \mathfrak{V}_j(\kappa), \mathfrak{W}_{ij}(\kappa) \to 0 \quad \text{in } L^2 \quad \text{as } m \to \infty,$$

with

$$\mathbb{E}[|\mathfrak{U}_{i,\ell}(\kappa)|^2 + |\mathfrak{V}_j(\kappa)|^2 + |\mathfrak{W}_{ij}(\kappa)|^2] \leq Cm^{-\frac{1}{2}} \to 0 \quad \text{as } m \to \infty, \tag{4.19}$$

for all $i, j \in \mathbb{N}$, $1 \leq \ell \leq d$ fixed.

The same analysis also yields that, for every $\varepsilon > 0$ there exists some C_ε such that

$$\mathbb{E}[\|\mathfrak{E}_h(\kappa)\|^2] \leq C_\varepsilon m^{-1+\varepsilon} \to 0 \quad \text{as } m \to \infty,$$

so that $h_\kappa(x) \to \mathring{h}_\kappa(x)$ almost surely for every $x \in \mathbb{R}^d$, as $m \to \infty$. This gives the almost optimal quantitative convergence of the linear predictor, (2.14).

We finish by taking, on the one hand

$$\boldsymbol{A}(\kappa) = \begin{pmatrix} \boldsymbol{\alpha}_0(\kappa) \\ \overline{\boldsymbol{\alpha}}_1(\kappa) \\ \boldsymbol{\alpha}_2(\kappa) \\ \overline{\boldsymbol{\alpha}}_3(\kappa) \\ \vdots \end{pmatrix} \in \mathbb{R}^{\infty \times d}, \qquad \boldsymbol{B}(\kappa) = \begin{pmatrix} \overline{\beta}_0(\kappa) \\ \boldsymbol{\beta}_1(\kappa) \\ \overline{\beta}_2(\kappa) \\ \boldsymbol{\beta}_3(\kappa) \\ \vdots \end{pmatrix} \in \mathbb{R}^{\infty},$$

and

$$\boldsymbol{G}(\kappa) = \begin{pmatrix} \hat{\bar{\gamma}}_{00}(\kappa) & \overline{\gamma}_{01}(\kappa) & \hat{\bar{\gamma}}_{02}(\kappa) & \dots \\ \boldsymbol{\gamma}_{10}(\kappa) & \hat{\gamma}_{11}(\kappa) & \boldsymbol{\gamma}_{20}(\kappa) & \dots \\ \hat{\bar{\gamma}}_{20}(\kappa) & \overline{\gamma}_{21}(\kappa) & \hat{\bar{\gamma}}_{22}(\kappa) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \in \mathbb{R}^{\infty \times \infty},$$

which are well defined independently of m, if m is large enough for a fixed κ. On the other hand, recovering the superscripts m in the notation, we know rom Theorem 3.2 (more precisely, from Proposition 3.4), that the family of vectors $(\boldsymbol{J}_k^m, \boldsymbol{K}_k^m)$ converges, in distribution, to a family of independent, identically distributed (infinite) random vectors (3.6), that we denote $(\boldsymbol{J}_k^\infty, \boldsymbol{K}_k^\infty)$. Hence, we can take in (2.12)

$$(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \dots) = (\boldsymbol{J}_0^\infty, \boldsymbol{K}_1^\infty, \boldsymbol{J}_2^\infty, \dots),$$
$$(\tilde{\boldsymbol{\Gamma}}_1, \tilde{\boldsymbol{\Gamma}}_2, \dots) = (\boldsymbol{K}_0^\infty, \boldsymbol{J}_1^\infty, \boldsymbol{K}_2^\infty, \dots),$$

(notice that these equalities are not elementwise, but rather as matrices; that is, $\boldsymbol{\Gamma}_2 = \boldsymbol{J}_0^{(2),\infty}$ if $d \geq 2$). Thanks to Proposition 3.4 and (4.19), we are done. □

5. Properties of the infinite-width dynamics

In this section, we study the behavior of the time-continuous ($\tau \to 0$) version of the limit system (2.13)-(2.10), as $\tau \downarrow 0$, namely the gradient flow of $\mathcal{E}$ (2.11) with initialization (2.7).

5.1. **A gradient flow.** We start by showing that the time-continuous version of (2.13), (5.1) below, is a gradient flow of the energy with respect to the Euclidean norm of the parameters (in particular, in the limiting case $m \to \infty$, it is a gradient flow in ℓ^2), and that the variation of the squared ℓ_2-norm of each layer is the same; a property that follows from the 1-homogeneity of the output w.r.t. each layer, which is often used in the analysis of linear NNs [4, 16]. This property is often used in conjunction with a *balanced initialization* assumption [4, Eq. (7)], which does not hold here, in particular because the middle layer has infinite ℓ_2-norm at initialization.

Proposition 5.1. *Let $m \in \mathbb{N} \cup \{\infty\}$. Let $(\boldsymbol{A}(t), \boldsymbol{G}(t), \boldsymbol{B}(t))$ with $\boldsymbol{A}(t) : [0,\infty) \to \mathbb{R}^{m\times d}$, $\boldsymbol{G} : [0,\infty) \to \mathbb{R}^{m\times m}$, and $\boldsymbol{B}(t) : [0,\infty) \to \mathbb{R}^m$ be a solution to the following ODE system*

$$\left\{ \begin{array}{rcl} \dot{\boldsymbol{A}}(t) &=& -[\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top \boldsymbol{B}(t) \boldsymbol{\xi}_t^\top, \\ \dot{\boldsymbol{G}}(t) &=& -\boldsymbol{B}(t) \boldsymbol{\xi}_t^\top \boldsymbol{A}(t)^\top, \\ \dot{\boldsymbol{B}}(t) &=& -[\boldsymbol{\Lambda} + \boldsymbol{G}(t)] \boldsymbol{A}(t) \boldsymbol{\xi}_t, \end{array} \right. \tag{5.1}$$

with

$$\boldsymbol{\xi}_t = \int x \mathcal{L}'(h_t(x), y) d\rho(x,y) \in \mathbb{R}^d, \qquad h_t(x) = \boldsymbol{B}(t)^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)] \boldsymbol{A}(t) x. \tag{5.2}$$

and $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times m}$ is a fixed matrix, equal to:

$$\mathbb{R}^{m\times m} \ni \boldsymbol{\Lambda} = (\Lambda_{ij})_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } i + d = j \text{ or } j + 1 = i \\ 0 & \text{otherwise.} \end{array} \right.$$

Then, (5.1) is the gradient flow in the ℓ_2-norm of the energy functional

$$\mathcal{E}(\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{B}) := \int \mathcal{L}(\boldsymbol{B}^\top [\boldsymbol{\Lambda} + \boldsymbol{G}] \boldsymbol{A} x, y) d\rho(x,y).$$

In particular, we have

$$\frac{d}{dt} \int \mathcal{L}(h_t(x), y) d\rho(x,y) \le 0,$$

and

$$\frac{d}{dt} \|\boldsymbol{A}(t)\|^2 = \frac{d}{dt} \|\boldsymbol{B}(t)\|^2 = \frac{d}{dt} \|\boldsymbol{\Lambda} + \boldsymbol{G}(t)\|^2 = -2 \int h_t(x) \mathcal{L}'(h_t(x), y) d\rho(x,y). \tag{5.3}$$

Proof. Let us formally compute, using (5.1)

$$\frac{d}{dt} \|\boldsymbol{A}(t)\|^2 = \frac{d}{dt} \mathrm{tr} \left\{ \boldsymbol{A}(t)^\top \boldsymbol{A}(t) \right\} = 2 \mathrm{tr} \left\{ \dot{\boldsymbol{A}}(t)^\top \boldsymbol{A}(t) \right\} = -2 \boldsymbol{B}(t)^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)] \boldsymbol{A}(t) \boldsymbol{\xi}_t.$$

We can proceed similarly with $\boldsymbol{\Lambda} + \boldsymbol{G}(t)$ and $\boldsymbol{B}(t)$ to get (5.3).

The fact that (5.1) is the gradient flow in the 2-norm of $\mathcal{E}$ is a direct check. For future convenience, we explicitly compute the dissipation by first obtaining the

evolution of $h_t(x)$

$$-\frac{d}{dt}h_t(x) = \boldsymbol{\xi}_t^\top \boldsymbol{A}(t)^\top[\boldsymbol{\Lambda}+\boldsymbol{G}(t)]^\top[\boldsymbol{\Lambda}+\boldsymbol{G}(t)]\boldsymbol{A}(t)x + \boldsymbol{B}(t)^\top\boldsymbol{B}(t)\boldsymbol{\xi}_t^\top\boldsymbol{A}(t)^\top\boldsymbol{A}(t)x \\ + \boldsymbol{B}(t)^\top[\boldsymbol{\Lambda}+\boldsymbol{G}(t)][\boldsymbol{\Lambda}+\boldsymbol{G}(t)]^\top\boldsymbol{B}(t)\boldsymbol{\xi}_t^\top x, \tag{5.4}$$

so that denoting $\mathcal{E}(t) := \mathcal{E}(\boldsymbol{A}(t), \boldsymbol{G}(t), \boldsymbol{B}(t))$,

$$\frac{d}{dt}\mathcal{E}(t) = \int \frac{d}{dt}h_t(x)\mathcal{L}'(h_t(x), y)d\rho(x,y) \\ = -\|[\boldsymbol{\Lambda}+\boldsymbol{G}(t)]\boldsymbol{A}(t)\boldsymbol{\xi}_t\|^2 - \|\boldsymbol{B}(t)\|^2\|\boldsymbol{A}(t)\boldsymbol{\xi}_t\|^2 - \|\boldsymbol{\xi}_t\|^2\|[\boldsymbol{\Lambda}+\boldsymbol{G}(t)]^\top\boldsymbol{B}(t)\|^2.$$

All the above computations also work if $m = \infty$, in which case we consider the ℓ^2 norms of the parameters. □

Remark 5.2. When $m = \infty$, (5.3) should be paired with some initial conditions that ensure its finiteness, and since $\|\boldsymbol{\Lambda}\| = +\infty$, the third term should be interpreted as

$$\frac{d}{dt}\left(\|\boldsymbol{\Lambda}+\boldsymbol{G}(t)\|^2 - \|\boldsymbol{\Lambda}\|^2\right) = \frac{d}{dt}\mathrm{tr}\left(\boldsymbol{\Lambda}^\top\boldsymbol{G}(t) + \boldsymbol{G}^\top(t)\boldsymbol{G}(t)\right).$$

5.2. **Selection principle.** Recall that we initialize our system (5.1) with

$$\boldsymbol{A}(0) = \begin{pmatrix} \mathrm{Id}_d \\ \mathbf{0}_{d\times 1} \\ \mathbf{0}_{d\times 1} \\ \vdots \end{pmatrix} \in \mathbb{R}^{m\times d}, \quad \boldsymbol{B}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \in \mathbb{R}^m, \tag{5.5}$$

and

$$(\boldsymbol{G})_{ij}(0) = 0 \quad \text{for} \quad 1 \le i,j \le m. \tag{5.6}$$

If we denote

$$\boldsymbol{\lambda}_t := \boldsymbol{A}(t)^\top[\boldsymbol{\Lambda}+\boldsymbol{G}(t)]^\top\boldsymbol{B}(t) \in \mathbb{R}^d,$$

so that $h_t(x) = \boldsymbol{\lambda}_t^\top x$, we next show that $\boldsymbol{\lambda}_t^\top$ never leaves the span of our data. That is,

$$\boldsymbol{\lambda}_t^\top\boldsymbol{v} = 0 \qquad \text{for all} \quad \boldsymbol{v} \in \mathrm{span}\left(\mathrm{supp}((\pi_x)_\#\rho)\right)^\perp, \tag{5.7}$$

where $\pi_x : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is the projection operator $(x,y) \mapsto x$, and $(\pi_x)_\#\rho$ denotes the pushforward of ρ through π_x.

Proposition 5.3. *Under the assumptions of Proposition 5.1, let us further assume that $\mathcal{L}''$ is bounded and that $(\boldsymbol{A}(t), \boldsymbol{G}(t), \boldsymbol{B}(t))$ are initialized at* (5.5) *and* (5.6). *Then* (5.7) *holds for all $t \ge 0$.*

Proof. Since $\mathcal{L}'$ is Lipschitz we know that the evolution is globally defined in time. Moreover, since $\boldsymbol{\lambda}_0 = 0$ we only need to show

$$\frac{d}{dt}\boldsymbol{\lambda}_t^\top\boldsymbol{v} = 0 \quad \text{for all} \quad t > 0,$$

where $\boldsymbol{v} \in \mathrm{span}\left(\mathrm{supp}((\pi_x)_\#\rho)\right)^\perp$ will be fixed throughout the proof.

We can compute, using (5.1),

$$\frac{d}{dt}\boldsymbol{\lambda}_t^\top = -\boldsymbol{B}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}][\boldsymbol{\Lambda}+\boldsymbol{G}]^\top\boldsymbol{B}\boldsymbol{\xi}^\top - \boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\xi}^\top\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{\xi}^\top\boldsymbol{A}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]\boldsymbol{A},$$

where we have omitted the time dependence for the sake of readability, that will be made only explicit at time 0. Observe now that, since $\boldsymbol{v} \in \operatorname{span}(\operatorname{supp}((\pi_x)_\#\rho))^\perp$,

$$\boldsymbol{\xi}^\top\boldsymbol{v} = 0 \qquad \text{for all} \quad t \geq 0.$$

Hence, $\dot{\boldsymbol{A}}\boldsymbol{v} = 0$ for all $t \geq 0$, which implies that $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}_0\boldsymbol{v}$ (where $\boldsymbol{A}_0 = \boldsymbol{A}(0)$, given by (5.5)). In all, we have

$$\frac{d}{dt}\boldsymbol{\lambda}_t^\top\boldsymbol{v} = -\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\xi}^\top\boldsymbol{A}^\top\boldsymbol{A}_0\boldsymbol{v} - \boldsymbol{\xi}^\top\boldsymbol{A}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]\boldsymbol{A}_0\boldsymbol{v}. \tag{5.8}$$

Let us now define the following quantities:

$$\begin{array}{rclrcl} M_t &:=& \boldsymbol{B}^\top\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v} \in \mathbb{R}, & \boldsymbol{N}_t &:=& \boldsymbol{G}^\top\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v} \in \mathbb{R}^m, \\ \boldsymbol{O}_t &:=& \boldsymbol{\Pi}^\top\boldsymbol{A}^\top\boldsymbol{A}_0\boldsymbol{v} \in \mathbb{R}^d, & \boldsymbol{P}_t &:=& \boldsymbol{G}\boldsymbol{A}_0\boldsymbol{v} \in \mathbb{R}^m, \end{array}$$

where we have denoted by $\boldsymbol{\Pi} \in \mathbb{R}^{d\times d}$ the projection matrix to $\operatorname{span}(\operatorname{supp}((\pi_x)_\#\rho))$, so that

$$\boldsymbol{\Pi}\boldsymbol{w} = \boldsymbol{w} \quad \text{for all} \quad \boldsymbol{w} \in \operatorname{span}(\operatorname{supp}((\pi_x)_\#\rho)).$$

In particular, we always have that $\boldsymbol{\xi}^\top\boldsymbol{\Pi}^\top = \boldsymbol{\xi}^\top$. In the following, we will use that

$$\boldsymbol{A}_0\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{0}_{m-d} \end{pmatrix},$$

and hence, since the first $d \times d$ submatrix of $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$ is the identity (which is a simple check) we have

$$\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v} = \boldsymbol{A}_0\boldsymbol{v}. \tag{5.9}$$

A computation using (5.1) and (5.9) gives then the following system of ODEs:

$$\begin{cases} \dot{M}_t = -\boldsymbol{\xi}^\top\boldsymbol{A}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]^\top\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v} = -\boldsymbol{\xi}^\top\boldsymbol{O}_t - \boldsymbol{\xi}^\top\boldsymbol{A}^\top\boldsymbol{N}_t \\ \dot{\boldsymbol{N}}_t = -\boldsymbol{A}^\top\boldsymbol{\xi}\boldsymbol{B}^\top\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v} = -\boldsymbol{A}^\top\boldsymbol{\xi}M_t \\ \dot{\boldsymbol{O}}_t = -\boldsymbol{\Pi}^\top\boldsymbol{\xi}\boldsymbol{B}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]\boldsymbol{A}_0\boldsymbol{v} = -\boldsymbol{\Pi}^\top\boldsymbol{\xi}M_t - \boldsymbol{\Pi}^\top\boldsymbol{\xi}\boldsymbol{B}^\top\boldsymbol{P}_t \\ \dot{\boldsymbol{P}}_t = -\boldsymbol{B}\boldsymbol{\xi}^\top\boldsymbol{A}^\top\boldsymbol{A}_0\boldsymbol{v} = -\boldsymbol{B}\boldsymbol{\xi}^\top\boldsymbol{O}_t, \end{cases} \tag{5.10}$$

which is initialized at

$$M_0 = 0, \quad \boldsymbol{N}_0 = \boldsymbol{0}_m, \quad \boldsymbol{O}_0 = \boldsymbol{0}_d, \quad \boldsymbol{P}_0 = \boldsymbol{0}_m. \tag{5.11}$$

Here, we used that $\boldsymbol{G}(0) = 0$, that the first element of $\boldsymbol{\Lambda}\boldsymbol{A}_0\boldsymbol{v}$ is zero (and hence, $M_0 = 0$), that $\boldsymbol{\xi}^\top\boldsymbol{\Pi}^\top = \boldsymbol{\xi}^\top$, and that $\boldsymbol{A}_0^\top\boldsymbol{A}_0 = \mathrm{Id}_d$ so $\boldsymbol{O}_0 = \boldsymbol{\Pi}^\top\boldsymbol{v} = \boldsymbol{0}_d$. The system (5.10) is Lipschitz in its variables, coupled with locally bounded coefficients (thanks to (5.3)), and therefore it has a unique solution. Since the initial conditions (5.11) all vanish, the unique solution is $(M_t, \boldsymbol{N}_t, \boldsymbol{O}_t, \boldsymbol{P}_t) = (0, \boldsymbol{0}_m, \boldsymbol{0}_d, \boldsymbol{0}_m)$ for $t \geq 0$.

Finally, we can rewrite (5.8) in terms of $(M_t, \boldsymbol{N}_t, \boldsymbol{O}_t, \boldsymbol{P}_t)$ (recalling (5.9)) as

$$\frac{d}{dt}\boldsymbol{\lambda}_t^\top\boldsymbol{v} = -\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\xi}^\top\boldsymbol{O}_t - \boldsymbol{\xi}^\top\boldsymbol{O}_t - \boldsymbol{\xi}^\top\boldsymbol{A}^\top\boldsymbol{N}_t - \boldsymbol{\xi}^\top\boldsymbol{A}^\top[\boldsymbol{\Lambda}+\boldsymbol{G}]^\top\boldsymbol{P}_t = 0,$$

which is our desired result. □

Remark 5.4. We highlight that the selection principle in Proposition 5.3 is not a consequence of a general abstract result on gradient flows with this particular structure, but rather follows from the precise initialization that arises from the infinite width limit, as illustrated by the following example.

By denoting $\boldsymbol{e}_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top$, let us define

$$\mathcal{E}(\boldsymbol{A}, \boldsymbol{z}) := \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{z}, \boldsymbol{e}_1\rangle^2, \quad \text{with} \quad \mathbb{R}^{2\times 2} \ni \boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},\ \mathbb{R}^2 \ni \boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

which is the empirical risk of a two-layer linear NN with a single sample $(x_1, y_1) = (\boldsymbol{e}_1, 0)$ in the training set. Consider its gradient flow:

$$\dot{\boldsymbol{A}} = -\partial_{\boldsymbol{A}}\mathcal{E}(\boldsymbol{A}, \boldsymbol{z}) = -\langle \boldsymbol{A}\boldsymbol{z}, \boldsymbol{e}_1\rangle \boldsymbol{e}_1 \boldsymbol{z}^\top = -(A_{11}z_1 + A_{12}z_2)\begin{pmatrix} z_1 & z_2 \\ 0 & 0 \end{pmatrix},$$

$$\dot{\boldsymbol{z}} = -\partial_{\boldsymbol{z}}\mathcal{E} = -\langle \boldsymbol{A}\boldsymbol{z}, \boldsymbol{e}_1\rangle \boldsymbol{A}^\top \boldsymbol{e}_1 - (A_{11}z_1 + A_{12}z_2)\begin{pmatrix} A_{11} \\ A_{12} \end{pmatrix}.$$

Then, if we denote $\boldsymbol{\lambda} := \boldsymbol{A}\boldsymbol{z} = \begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix}^\top$, we can express the energy as

$$\mathcal{E}(\boldsymbol{A}, \boldsymbol{z}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{z}, \boldsymbol{e}_1\rangle^2 = \frac{1}{2}\lambda_1^2. \tag{5.12}$$

It is however not true that the evolution of $\boldsymbol{\lambda}$ must be such that it always moves along the span of $\boldsymbol{e}_1$. Indeed, using the previous gradient flow, we know that

$$\dot{\boldsymbol{\lambda}} = \dot{\boldsymbol{A}}\boldsymbol{z} + \boldsymbol{A}\dot{z} = -(A_{11}z_1 + A_{12}z_2)\left(\begin{pmatrix} z_1^2 + z_2^2 \\ 0 \end{pmatrix} + \begin{pmatrix} A_{11}^2 + A_{12}^2 \\ A_{21}A_{11} + A_{22}A_{12} \end{pmatrix}\right).$$

Hence, when $(A_{11}z_1 + A_{12}z_2)(A_{21}A_{11} + A_{22}A_{12}) \neq 0$, the second coordinate $\boldsymbol{\lambda}$ is moving. This can happen by choosing at time $t = 0$

$$\boldsymbol{A}(0) := \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \boldsymbol{z}(0) := \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{so that} \quad \boldsymbol{\lambda}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

and, since $\dot{\lambda}_2(0) \neq 0$, we have that $\lambda_2(t) \neq 0$ for some time $t > 0$, despite the fact that the energy in (5.12) depends only on λ_1.

5.3. **Quantitative convergence and implicit bias.** Whenever the loss function is uniformly convex (we take the quadratic case for convenience) then we expect exponential rate of convergence towards a minimizer.

In the following, given a measure ρ, we denote by $\boldsymbol{M}$ the covariance matrix,

$$\boldsymbol{M} := \int x x^\top d\rho(x, y) \in \mathbb{R}^{d\times d}. \tag{5.13}$$

Note that $\boldsymbol{M}$ is symmetric and positive semi-definite. In particular, if $\boldsymbol{M}$ is non-degenerate ($\det(\boldsymbol{M}) > 0$), then there is a unique minimizer $\boldsymbol{\lambda} \in \mathbb{R}^d$ of the quadratic energy

$$\mathcal{E} = \int (\boldsymbol{\lambda} \cdot x - y)^2 d\rho(x, y).$$

Otherwise, and as we have seen in Proposition 5.3, our system will converge to a minimizer in the span of $\operatorname{supp}((\pi_x)_\# \rho)$ (alternatively, in $\ker(\boldsymbol{M})^\perp$ or in the row space of $\boldsymbol{M}$), which is unique. We prove that it will do so at an exponential rate, depending on the lowest non-zero eigenvalue of $\boldsymbol{M}$.

Proposition 5.5. *Under the assumptions of Proposition 5.1, let us further assume that $\mathcal{L}$ is the quadratic loss function and that $\boldsymbol{M}$ has d' non-zero eigenvalues, with $1 \leq d' \leq d$, that we denote $0 < z_1 \leq z_2 \leq \cdots \leq z_{d'}$.*

Let $(\boldsymbol{A}(t), \boldsymbol{G}(t), \boldsymbol{B}(t))$ denote the evolution (5.1) initialized at (5.5) and (5.6). Then $\boldsymbol{\lambda}_t$ converges to the unique minimizer $\boldsymbol{\lambda} \in \mathbb{R}^d$ of the energy functional,

$$\mathcal{E}_t := \int (h_t(x) - y)^2 d\rho(x,y) = \int (\boldsymbol{\lambda}_t \cdot x - y)^2 d\rho(x,y)$$

such that $\boldsymbol{\lambda} \in \operatorname{span}(\operatorname{supp}((\pi_x)_\# \rho))$ (alternatively, $\boldsymbol{\lambda} \in \ker(\boldsymbol{M})^\perp$), and

$$\mathcal{E}_t - \mathcal{E}_\infty \leq (\mathcal{E}_0 - \mathcal{E}_\infty)\, e^{-\tilde{c}_\lambda t} \quad \text{for} \quad t \geq 0$$

for some constant $\tilde{c}_\lambda$ depending only on $\|\boldsymbol{\lambda}\|$, d, z_1, and $z_{d'}$ (and independent of m).

Proof. We divide the proof into four steps.

Step 1: The setting. We use the same notation as in Proposition 5.1 and Proposition 5.3. We recall that we had denoted

$$\boldsymbol{\lambda}_t := \boldsymbol{A}(t)^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top \boldsymbol{B}(t) \in \mathbb{R}^d.$$

(In particular, $\boldsymbol{\lambda}_0 = \mathbf{0}_{\mathrm{d}\times 1}$.) The condition on $\boldsymbol{M}$ can then be re-written as

$$0 < z_1 |\boldsymbol{w}|^2 \leq \boldsymbol{w} \cdot \boldsymbol{M}\boldsymbol{w} \leq z_{d'} |\boldsymbol{w}|^2 \qquad \text{for all} \quad \boldsymbol{w} \in \ker(\boldsymbol{M})^\perp. \tag{5.14}$$

The energy is given by

$$\mathcal{E}_t := \mathcal{E}(\boldsymbol{A}(t), \boldsymbol{G}(t), \boldsymbol{B}(t)) = \int (h_t(x) - y)^2 \, d\rho(x,y),$$

where we recall that $h_t(x) = \boldsymbol{\lambda}_t \cdot x$. In particular, we can explicitly compute the minimizer $\boldsymbol{\lambda}$ (with $\boldsymbol{\lambda} \in \ker(\boldsymbol{M})^\perp$) and the evolution of $\mathcal{E}_t$ in terms of $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda} := \int y\, \boldsymbol{M}^{-1} x \, \rho(x,y) \in \mathbb{R}^d, \qquad \mathcal{E}_t = (\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \cdot \boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) + \mathcal{E}_\infty, \tag{5.15}$$

where, by an abuse of notation, we denoted by $\boldsymbol{M}^{-1}x$ the inverse restricted to $\ker(\boldsymbol{M})^\perp$ of $x \in \operatorname{supp}(\pi_x)_\# \rho$, so that $\boldsymbol{\lambda} \in \ker(\boldsymbol{M})^\perp$ as well. From (5.14) and the fact that $\boldsymbol{\lambda}_t \in \ker(\boldsymbol{M})^\perp$ for all $t \geq 0$ (see Proposition 5.3), we have

$$z_1 \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\|^2 \leq \mathcal{E}_t - \mathcal{E}_\infty \leq z_{d'} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\|^2. \tag{5.16}$$

We also have (cf. (5.4))

$$\boldsymbol{\xi}_t = 2\boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \qquad \text{and} \qquad \dot{\boldsymbol{\lambda}}_t = -2\boldsymbol{R}_t \boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda})$$

where $\boldsymbol{R}_t$ is a symmetric matrix, $\boldsymbol{R}_t \geq 0$, defined by

$$\begin{aligned} \boldsymbol{R}_t = {} & \boldsymbol{A}(t)^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)] \boldsymbol{A}(t) + \boldsymbol{B}(t)^\top \boldsymbol{B}(t) \boldsymbol{A}(t)^\top \boldsymbol{A}(t) \\ & + \boldsymbol{B}(t)^\top [\boldsymbol{\Lambda} + \boldsymbol{G}(t)][\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top \boldsymbol{B}(t) \mathrm{Id}_{d\times d} \in \mathbb{R}^{d\times d}. \end{aligned} \tag{5.17}$$

Thus,

$$\dot{\mathcal{E}}_t = -4(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \cdot \boldsymbol{M}\boldsymbol{R}_t \boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}). \tag{5.18}$$

Observe also that (see the proof of Proposition 5.1)

$$\begin{aligned}\frac{d}{dt}\|\boldsymbol{A}(t)\|^2 = \frac{d}{dt}\|\boldsymbol{B}(t)\|^2 &= -4\boldsymbol{\lambda}_t \cdot \boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \\ &= -4(\mathcal{E}_t - \mathcal{E}_\infty) - 4\boldsymbol{\lambda} \cdot \boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \\ &\leq 4z_{d'}^{\frac{1}{2}}\|\boldsymbol{\lambda}\|\sqrt{\mathcal{E}_t - \mathcal{E}_\infty} \leq 4z_{d'}\|\boldsymbol{\lambda}\|^2,\end{aligned} \tag{5.19}$$

where we used that the energy is decreasing, Cauchy-Schwarz, and (5.14). Similarly, for any $\boldsymbol{e} \in \mathbb{S}^{d-1}$,

$$\begin{aligned}\frac{d}{dt}\|\boldsymbol{A}(t)\boldsymbol{e}\|^2 &= -4(\boldsymbol{\lambda}_t \cdot \boldsymbol{e})\, \boldsymbol{e}\boldsymbol{M}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}) \\ &\leq C\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\|^2 + \|\boldsymbol{\lambda}\|\,\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}\| \leq C\|\boldsymbol{\lambda}\|^2\end{aligned} \tag{5.20}$$

for some constant C depending only on z_1 and $z_{d'}$.

Step 2: Small times. We have $\boldsymbol{R}_t \geq \|\boldsymbol{B}(t)\|^2\boldsymbol{A}(t)^\top\boldsymbol{A}(t)$ and $\boldsymbol{R}_0 \geq \mathrm{Id}_{d\times d}$. In particular, thanks to (5.19)-(5.20),

$$\boldsymbol{R}_t \geq \frac{1}{2}\mathrm{Id}_{d\times d} \quad \text{for} \quad t \leq t_\circ, \tag{5.21}$$

where $t_\circ = c_\circ\|\boldsymbol{\lambda}\|^{-2}$ for some $c_\circ > 0$ depending only on z_1 and $z_{d'}$. Hence,

$$\dot{\mathcal{E}}_t \leq -c(\mathcal{E}_t - \mathcal{E}_\infty) \qquad \text{for} \qquad 0 \leq t < t_\circ,$$

for some c depending only on z_1 and $z_{d'}$, thanks to (5.15)-(5.16)-(5.18)-(5.21) (we use that if $\boldsymbol{M}$ and $\boldsymbol{R}_t$ are symmetric positive semi-definite matrices, then $\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{R}_t\boldsymbol{M}^{\frac{1}{2}}$ is positive semi-definite as well). In particular,

$$\mathcal{E}_t - \mathcal{E}_\infty \leq (\mathcal{E}_0 - \mathcal{E}_\infty)e^{-ct} \quad \text{for} \quad 0 \leq t < t_\circ. \tag{5.22}$$

Step 3: An ODE for all times. From the previous inequality and the dissipation of energy, we have

$$\|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\lambda}\|e^{-\frac{ct_\circ}{2}} \geq \|\boldsymbol{M}^{\frac{1}{2}}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda})\| \geq \|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\lambda}\| - \|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\lambda}_t\| \quad \text{for} \quad t \geq t_\circ,$$

so that

$$\|\boldsymbol{\lambda}_t\|^2 \geq C_\rho^{-1}\|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\lambda}_t\|^2 \geq C_\rho^{-1}\left(1 - e^{-\frac{ct_\circ}{2}}\right)^2\|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\lambda}\|^2 \geq C_\rho^{-2}\|\boldsymbol{\lambda}\|^2\left(1 - e^{-\frac{ct_\circ}{2}}\right)^2 =: c_\lambda$$

with $c_\lambda > 0$, for $t \geq t_\circ$. In particular, by Cauchy-Schwarz and up to a dimensional constant, from the definition of $\boldsymbol{\lambda}_t$,

$$c_\lambda \leq \|\boldsymbol{\lambda}_t\|^2 \leq C\|\boldsymbol{A}(t)\|^2\|[\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top\boldsymbol{B}(t)\|^2 \quad \text{for} \quad t \geq t_\circ.$$

From (5.17) we know that for some dimensional $c > 0$,

$$\boldsymbol{R}_t \geq \|[\boldsymbol{\Lambda} + \boldsymbol{G}(t)]^\top\boldsymbol{B}(t)\|^2\,\mathrm{Id}_{d\times d} \geq cc_\lambda\|\boldsymbol{A}(t)\|^{-2}\,\mathrm{Id}_{d\times d} \quad \text{for} \quad t \geq t_\circ.$$

On the other hand, from (5.19), and since $\|\boldsymbol{A}(0)\|^2 = d$,

$$\|\boldsymbol{A}(t)\|^2 \leq d + C\|\boldsymbol{\lambda}\|\int_0^t \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\,d\tau,$$

and hence

$$\boldsymbol{R}_t \geq \frac{cc_\lambda}{d + C\|\boldsymbol{\lambda}\|\int_0^t \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\,d\tau}\,\mathrm{Id}_{d\times d} \quad \text{for} \quad t \geq t_\circ.$$

Combined again with (5.14)-(5.16)-(5.18) we obtain the inequality

$$\dot{\mathcal{E}}_t \leq -\frac{cc_\lambda(\mathcal{E}_t - \mathcal{E}_\infty)}{1 + \|\boldsymbol{\lambda}\| \int_0^t \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\, d\tau} \quad \text{for} \quad t \geq t_\circ. \tag{5.23}$$

Step 4: Bootstrap argument. Observe that

$$\int_0^t \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\, d\tau \leq C\|\boldsymbol{\lambda}\| t, \tag{5.24}$$

since we have dissipation of the energy. Hence, from (5.23) we get

$$\dot{\mathcal{E}}_t \leq -\frac{cc_\lambda(\mathcal{E}_t - \mathcal{E}_\infty)}{1 + \|\boldsymbol{\lambda}\| t} \quad \text{for} \quad t \geq t_\circ,$$

which implies (also using that $c_\lambda \leq C\|\boldsymbol{\lambda}\|^2$ and $t_\circ = c\|\boldsymbol{\lambda}\|^{-2}$)

$$\mathcal{E}_t - \mathcal{E}_\infty \leq (\mathcal{E}_{t_\circ} - \mathcal{E}_\infty)\left(\frac{1 + \|\boldsymbol{\lambda}\| t_\circ}{1 + \|\boldsymbol{\lambda}\| t}\right)^{\frac{cc_\lambda}{\|\boldsymbol{\lambda}\|}} \leq C\|\boldsymbol{\lambda}\|^2 (1 + \|\boldsymbol{\lambda}\| t)^{-\frac{cc_\lambda}{\|\boldsymbol{\lambda}\|}}, \quad \text{for} \quad t \geq t_\circ.$$

Plugging it back into (5.23), we now have that instead of (5.24) (also using (5.22)),

$$\int_0^t \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\, d\tau \leq C\frac{\|\boldsymbol{\lambda}\|}{1 + \|\boldsymbol{\lambda}\|^2} + C\|\boldsymbol{\lambda}\|(1 + \|\boldsymbol{\lambda}\| t)^{1-\varepsilon_\lambda}$$

where we have denoted $\varepsilon_\lambda := \frac{cc_\lambda}{\|\boldsymbol{\lambda}\|} < \frac{1}{2}$ (if c is sufficiently small). Again from (5.23),

$$\frac{\dot{\mathcal{E}}_t}{\mathcal{E}_t - \mathcal{E}_\infty} \leq -\frac{cc_\lambda}{1 + \|\boldsymbol{\lambda}\|^2 (1 + \|\boldsymbol{\lambda}\|^2)^{-2} + \|\boldsymbol{\lambda}\|^2 (1 + \|\boldsymbol{\lambda}\| t)^{1-\varepsilon_\lambda}} \quad \text{for} \quad t \geq t_\circ.$$

In particular, there exists some $\tilde{c}_\lambda$ depending on $\|\boldsymbol{\lambda}\|$, z_1, and $z_{d'}$, such that

$$\mathcal{E}_t - \mathcal{E}_\infty \leq (\mathcal{E}_0 - \mathcal{E}_\infty)\, e^{-\tilde{c}_\lambda t^{\varepsilon_\lambda}} \quad \text{for} \quad t \geq 0.$$

Iterating again the procedure, now $\int_0^\infty \sqrt{\mathcal{E}_\tau - \mathcal{E}_\infty}\, d\tau < +\infty$, and hence

$$\mathcal{E}_t - \mathcal{E}_\infty \leq (\mathcal{E}_0 - \mathcal{E}_\infty)\, e^{-\tilde{c}_\lambda t} \quad \text{for} \quad t \geq 0$$

for some (possibly different) $\tilde{c}_\lambda$ depending only on $\|\boldsymbol{\lambda}\|$, d, z_1, and $z_{d'}$ □

Finally, we have:

Proof of Theorem 2.3. If follows from Proposition 5.5. □

6. Multi-layer case

Let us now consider the multi-layer case, that is, the evolution of a neural network with $L+1$ hidden layers (being the previous case, $L = 1$). For the sake of readability, we do it in the case $d = 1$, but the same holds for $d > 1$. The aim of this section is to introduce and justify all the objects, notably the limit evolution equation and the basis in which such evolution is expressed, for the analogous of Theorem 2.2 to hold with $L + 1$ hidden layers. We remark that the following arguments are formal, and that their rigorous justifications can be obtained by the same methods developed in the core of the paper.

Using the notation in subsection 2.2, and dropping the superscript m, we now have $\boldsymbol{U} \in \mathbb{R}^m$, $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{m\times m}$ for $1 \le \ell \le L$, and $\boldsymbol{V} \in \mathbb{R}^m$, initialized as

$$U_j(0) \sim \mathcal{N}(0,1), \qquad W_{ij}^{(\ell)}(0) = 0, \qquad V_i(0) \sim \mathcal{N}(0,1).$$

We also fix $L \in \mathbb{N}$ independent random matrices of size $m \times m$ with independent entries $\mathcal{N}(0,1)$, $(\boldsymbol{Z}^{(\ell)})_{1\le\ell\le L}$. The neural network is (recall $x \in \mathbb{R}$):

$$y = h(x, \boldsymbol{U}, (\boldsymbol{W}^{(\ell)})_{1\le\ell\le L}, \boldsymbol{V}) = \left\langle \frac{1}{m}\boldsymbol{V}, \prod_{\ell=1}^{L}\left(\frac{1}{\sqrt{m}}\boldsymbol{Z}^{(\ell)} + \frac{1}{m}\boldsymbol{W}^{(\ell)}\right)\boldsymbol{U}x \right\rangle.$$

And the evolution $(\boldsymbol{U}(\kappa), (\boldsymbol{W}^{(\ell)}(\kappa))_{1\le\ell\le L}, \boldsymbol{V}(\kappa))_{\kappa\in\mathbb{N}}$ is a GD (with layer-wise learning rates) on the objective function

$$F(\boldsymbol{U}, (\boldsymbol{W}^{(\ell)})_{1\le\ell\le L}, \boldsymbol{V}) := \int_{\mathbb{R}^d\times\mathbb{R}} \mathcal{L}\left(h(x, \boldsymbol{U}, (\boldsymbol{W}^{(\ell)})_{1\le\ell\le L}, \boldsymbol{V}), y\right) d\rho(x,y),$$

given by

$$\begin{cases} \boldsymbol{U}(\kappa+1) = \boldsymbol{U}(\kappa) - \tau \displaystyle\prod_{\ell=1}^{L}\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^{(\ell)} + \frac{1}{m}\boldsymbol{W}^{(\ell)}(\kappa)\right]^\top \boldsymbol{V}(\kappa)(\boldsymbol{\xi}_{\kappa,\tau})^\top, \\ \boldsymbol{W}^{(\ell)}(\kappa+1) = \boldsymbol{W}^{(\ell)}(\kappa) - \tau \displaystyle\prod_{i=\ell+1}^{L}\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^{(i)} + \frac{1}{m}\boldsymbol{W}^{(i)}(\kappa)\right]^\top \boldsymbol{V}(\kappa)(\boldsymbol{\xi}_{\kappa,\tau})^\top \\ \qquad\qquad (\boldsymbol{U}(\kappa))^\top \displaystyle\prod_{i=1}^{\ell-1}\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^{(i)} + \frac{1}{m}\boldsymbol{W}^{(i)}(\kappa)\right]^\top, \qquad 1 \le \ell \le L, \\ \boldsymbol{V}(\kappa+1) = \boldsymbol{V}(\kappa) - \tau \displaystyle\prod_{\ell=L}^{1}\left[\frac{1}{\sqrt{m}}\boldsymbol{Z}^{(\ell)} + \frac{1}{m}\boldsymbol{W}^{(\ell)}(\kappa)\right]\boldsymbol{U}(\kappa)\boldsymbol{\xi}_{\kappa,\tau}, \end{cases} \tag{6.1}$$

with $\boldsymbol{\xi}_{\kappa,\tau} = \int x\, \mathcal{L}'(h_{\kappa,\tau}(x), y) d\rho_\kappa(x,y) \in \mathbb{R}$, where we have also denoted $h_{\kappa,\tau}(x) = h(x, \boldsymbol{U}(\kappa), (\boldsymbol{W}^{(\ell)}(\kappa))_{1\le\ell\le L}, \boldsymbol{V}(\kappa))$, and we always assume uniformly finite second moments, (2.2).

In analogy with the three-layer case, we expect the dynamics to be expressed, up to errors which vanish as m gets large, in a suitable Gaussian basis with certain orthogonality properties, and with an explicit behavior with respect to multiplication by $\boldsymbol{Z}^{(\ell)}$. The construction of such a basis (and more precisely, of one basis for each layer ℓ) is a nontrivial generalization of Theorem 3.2 and it is defined in subection 6.1 below. We describe now how to obtain the limit dynamics, assuming the existence of such a basis, whose properties are detailed in (6.2) and (6.4) below.

We assume therefore the existence of $L+1$ appropriate orthonormal bases, that we denote

$$\boldsymbol{\Psi}^0, \boldsymbol{\Psi}^1, \dots, \boldsymbol{\Psi}^L, \quad \text{with} \quad \boldsymbol{\Psi}^\ell = (\boldsymbol{\Psi}_1^\ell, \boldsymbol{\Psi}_2^\ell, \boldsymbol{\Psi}_3^\ell, \dots) \quad \text{for any} \qquad 0 \le \ell \le L,$$

such that $\boldsymbol{\Psi}^\ell \in \mathbb{R}^{m\times\infty}$ is a matrix formed of independent m-dimensional Gaussian vectors (as columns), $\boldsymbol{\Psi}_i^\ell \in \mathbb{R}^m$ for all $i \in \mathbb{N}$, with entries $\mathcal{N}(0,1)$, and that are going

to act as the approximate bases for $m < \infty$, satisfying

$$\frac{1}{m}(\boldsymbol{\Psi}^\ell)^\top \boldsymbol{\Psi}^\ell = \mathrm{Id}_\infty, \quad \frac{1}{m}(\boldsymbol{\Psi}^\ell)^\top \boldsymbol{\Psi}^{\ell'} = \mathbf{0}_{\infty\times\infty}, \qquad 0 \le \ell \ne \ell' \le L \tag{6.2}$$

up to errors that vanish as $m \to \infty$ (cf. Theorem 3.2). Namely, we assume that we can write, up to errors that are of order $O(m^{-\frac{1}{2}+\delta})$ for any $\delta > 0$,

$$\begin{cases} \boldsymbol{U}(\kappa) = \boldsymbol{\Psi}^0 \boldsymbol{A}(\kappa), \\ \boldsymbol{W}^\ell(\kappa) = \boldsymbol{\Psi}^\ell \boldsymbol{G}_\ell(\kappa)(\boldsymbol{\Psi}^{\ell-1})^\top, \qquad 1 \le \ell \le L, \\ \boldsymbol{V}(\kappa) = \boldsymbol{\Psi}^L \boldsymbol{B}(\kappa), \end{cases} \tag{6.3}$$

for some coefficients $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^\infty$, $\boldsymbol{G}_\ell \in \mathbb{R}^{\infty\times\infty}$ for $1 \le \ell \le L$, initialized as (2.7)-(2.8) for $d = 1$ and all $1 \le \ell \le L$. Finally, we also assume the following recurrence relationship between bases under multiplication by $\boldsymbol{Z}^{(\ell)}$ (cf. subsection 3.4),

$$\begin{aligned} \frac{1}{\sqrt{m}} \boldsymbol{Z}^{(\ell)} \boldsymbol{\Psi}^{\ell-1} &= \boldsymbol{\Psi}^\ell \boldsymbol{\Lambda}_\ell, \\ \frac{1}{\sqrt{m}} (\boldsymbol{Z}^{(\ell)})^\top \boldsymbol{\Psi}^\ell &= \boldsymbol{\Psi}^{\ell-1} \boldsymbol{\Lambda}_\ell^\top, \qquad 1 \le \ell \le L, \end{aligned} \tag{6.4}$$

for some fixed matrices $\boldsymbol{\Lambda}_\ell \in \mathbb{R}^{\infty\times\infty}$ (cf. equation (2.9)). We can then write an evolution for the coefficients $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{G}_\ell$, using (6.1)-(6.2)-(6.4) and the representation (6.3):

$$\begin{cases} \boldsymbol{A}(\kappa+1) = \boldsymbol{A}(\kappa) - \tau \displaystyle\prod_{\ell=1}^{L} (\boldsymbol{\Lambda}_\ell^\top + \boldsymbol{G}_\ell^\top(\kappa)) \boldsymbol{B}(\kappa), \\ \boldsymbol{G}_\ell(\kappa+1) = \boldsymbol{G}_\ell(\kappa) - \tau \displaystyle\prod_{i=\ell+1}^{L} (\boldsymbol{\Lambda}_i^\top + \boldsymbol{G}_i^\top(\kappa)) \boldsymbol{B}(\kappa) \boldsymbol{\xi}_{\kappa,\tau}^\top \boldsymbol{A}^\top(\kappa) \displaystyle\prod_{i=1}^{\ell-1} (\boldsymbol{\Lambda}_i^\top + \boldsymbol{G}_i^\top(\kappa)), \\ \boldsymbol{B}(\kappa+1) = \boldsymbol{B}(\kappa) - \tau \displaystyle\prod_{\ell=L}^{1} (\boldsymbol{\Lambda}_\ell + \boldsymbol{G}_\ell(\kappa)) \boldsymbol{A}(\kappa) \boldsymbol{\xi}_{\kappa,\tau}, \end{cases} \tag{6.5}$$

for $1 \le \ell \le L$, with

$$\chi_{\kappa,\tau}(x) = \chi(x, \boldsymbol{A}(\kappa), (\boldsymbol{G}_\ell(\kappa))_{1\le\ell\le L}, \boldsymbol{B}(\kappa)),$$
$$\boldsymbol{\xi}_{\kappa,\tau} = \int x \mathcal{L}'(\chi_{\kappa,\tau}(x), y) d\rho_\kappa(x,y) \in \mathbb{R}.$$

When $\rho_\kappa = \rho$ for all $\kappa \in \mathbb{N}$, this recursion is exactly the GD on the (deterministic) objective function $\mathcal{E}$ defined by

$$\mathcal{E}(\boldsymbol{A}, (\boldsymbol{G}_\ell)_{1\le\ell\le L}, \boldsymbol{B}) = \int \mathcal{L}\left(\boldsymbol{B}^\top \prod_{\ell=L}^{1} (\boldsymbol{\Lambda}_\ell + \boldsymbol{G}_\ell) \boldsymbol{A} x), y \right) d\rho(x,y), \tag{6.6}$$

and the linear predictor of the neural network is given by

$$\boldsymbol{A}^\top \prod_{\ell=1}^{L} (\boldsymbol{\Lambda}_\ell^\top + \boldsymbol{G}_\ell^\top) \boldsymbol{B},$$

up to errors that disappear as $m \to \infty$. Thus, the description of the linear neural network in the general multi-layered case, (6.5), is reduced to finding bases such that (6.2) and (6.4) hold, up to errors (which is precisely what we did in Section 3 above).

6.1. **The choice of the bases.** Given $L \in \mathbb{N}$ and $0 \leq \ell \leq L$, let us define the following set of finite sequences:

$$\mathcal{S}^L(\ell) := \{(s_0, s_1, s_2, \dots, s_M) : s_0 \in \{0, L\}, s_M = \ell, s_i \in \{0, \dots, L\}, |s_i - s_{i-1}| = 1\},$$

that is, $\mathcal{S}^L(\ell)$ is the set of finite sequences of numbers belonging to $\{0, \dots, L\}$, starting at 0 or L, finishing at ℓ, and such that each element of the sequence is obtained by adding or subtracting 1 to the previous element (in particular, if $s_0 = 0$, $s_1 = 1$ necessarily). This set is going to be, for each $0 \leq \ell \leq L$, our index set for the basis $\mathbf{\Psi}^\ell$. For example, when $L = 1$, the sequences in $\mathcal{S}^1(0)$ (and analogously in $\mathcal{S}^1(1)$) are just of the form 0101...0 or 1010...0, and can be identified with their length. This is the reason why the index set in the case $L = 1$ is just given by the natural numbers, which was the case in Section 3.

We therefore consider $\mathbf{\Psi}^\ell$ to have as columns the elements $\mathbf{\Psi}^\ell_s$ for $s \in \mathcal{S}^L(\ell)$, and we denote it,

$$\mathbf{\Psi}^\ell = (\mathbf{\Psi}^\ell_s)_{s \in \mathcal{S}^L(\ell)}, \qquad 0 \leq \ell \leq L,$$

where we still need to define what $\mathbf{\Psi}^\ell_s$ is for a given $s \in \mathcal{S}^L(\ell)$. To do so, for notational convenience, given the matrices $\boldsymbol{Z}^{(\ell)}$ for $1 \leq \ell \leq L$, we denote

$$\boldsymbol{Z}^{\ell-1,\ell} := (\boldsymbol{Z}^{(\ell)})^\top \qquad \text{and} \qquad \boldsymbol{Z}^{\ell,\ell-1} := \boldsymbol{Z}^{(\ell)}.$$

Moreover, we let $\mathbf{\Psi}^0_0$ and $\mathbf{\Psi}^L_L$ be two fixed independent Gaussian vectors of size m (that is, those associated to the sequences $\{0\}$ and $\{L\}$).

Then, given $s \in \mathcal{S}^L(\ell)$ of length $M+1$, $s = (s_0, \dots, s_M)$, we define

$$\mathbf{\Psi}^\ell_s := m^{-M/2} \sum_{(i_0,\dots,i_M) \in \mathcal{I}(s,m)} \left(\prod_{j=1}^{M} \boldsymbol{Z}^{s_j, s_{j-1}}_{i_j, i_{j-1}} \right) (\mathbf{\Psi}^{s_0}_{s_0})_{i_0}, \tag{6.7}$$

where $\mathcal{I}(s,m)$ is the set of indices $(i_0, \dots, i_M)$ with $i_j \in \{0, \dots, m\}$ such that $(i_j, s_j) \neq (i_k, s_k)$ for all $1 \leq j \neq k \leq M$. In other words, the main novelty of the current definition with respect to the corresponding definition (3.4) for $L = 1$ lies in the fact that the basis is parametrized by an element $s \in \mathcal{S}^L(\ell)$, which identifies a fixed sequence of consecutive layers. Once the sequence is fixed, the sum in (6.7) runs over all possible loopless choices of one element between $1, ..., m$ in each of the layers signposted by s.

Formally, we obtain orthonormal bases in the sense (6.2) (as in Proposition 3.6), and the relationships in (6.4) are of the form

$$\frac{1}{\sqrt{m}} \boldsymbol{Z}^{(\ell)} \mathbf{\Psi}^{\ell-1}_s = \begin{cases} \mathbf{\Psi}^\ell_{(s,\ell)} & \text{if} \quad s = (s', \ell-2, \ell-1), \\ \mathbf{\Psi}^\ell_{(s',\ell)} + \mathbf{\Psi}^\ell_{(s,\ell)} & \text{if} \quad s = (s', \ell, \ell-1), \end{cases} \tag{6.8}$$

and

$$\frac{1}{\sqrt{m}} (\boldsymbol{Z}^{(\ell)})^\top \mathbf{\Psi}^\ell_s = \begin{cases} \mathbf{\Psi}^{\ell-1}_{(s,\ell-1)} & \text{if} \quad s = (s', \ell+1, \ell), \\ \mathbf{\Psi}^{\ell-1}_{(s',\ell-1)} + \mathbf{\Psi}^{\ell-1}_{(s,\ell-1)} & \text{if} \quad s = (s', \ell-1, \ell), \end{cases} \tag{6.9}$$

for $1 \leq \ell \leq L$.

6.2. **The case $L = 2$.** In the case $L = 2$ (that is, a four layers neural network, or a neural network with three hidden layers) we have a more explicit expression. In this case, any element $s \in \mathcal{S}^2(\ell)$ is of the form

$$(s_0, 1, s_2, 1, s_4, 1, s_6, 1, s_8, 1, \dots), \quad \dots \quad s_{2i} \in \{0, 2\},$$

and therefore, we can identify any element s in $\mathcal{S}^2(0)$, $\mathcal{S}^2(1)$, or $\mathcal{S}^2(2)$, with a natural number $N(s)$, seeing it as a binary representation. Thus, we associate

$$\mathcal{S}^2(0) \ni s \mapsto N_0(s) := 2^\sigma + \sum_{i=1}^{\sigma} 2^{i-2} s_{2(\sigma-i)}$$

$$\mathcal{S}^2(1) \ni s \mapsto N_1(s) := 2^{\sigma+1} + \sum_{i=0}^{\sigma} 2^{i-1} s_{2(\sigma-i)}$$

$$\mathcal{S}^2(2) \ni s \mapsto N_2(s) := 2^\sigma + \sum_{i=1}^{\sigma} 2^{i-2} s_{2(\sigma-i)}$$

where we have denoted $\sigma = \lfloor M/2 \rfloor$ for $s = (s_0, \dots, s_M)$. With this indexing, we can obtain more explicit relations (6.8)-(6.9), since we now have that $\mathbf{\Psi}^0$, $\mathbf{\Psi}^1$, and $\mathbf{\Psi}^2$ can be indexed by the natural numbers. That is, as an abuse of notation we denote

$$\mathbf{\Psi}^i_j = \mathbf{\Psi}^i_s \qquad \text{if} \quad N_i(s) = j, \quad \text{for} \quad i = 0, 1, 2,$$

which is well-defined for any $j \geq 2$.

The relations (6.8)-(6.9) correspond to

$$\frac{1}{\sqrt{m}} \boldsymbol{Z}^{(1)} \mathbf{\Psi}^0_j = \mathbf{\Psi}^1_j + \mathbf{\Psi}^1_{2j}, \tag{6.10}$$

$$\frac{1}{\sqrt{m}} (\boldsymbol{Z}^{(2)})^\top \mathbf{\Psi}^2_j = \mathbf{\Psi}^1_j + \mathbf{\Psi}^1_{2j+1}, \tag{6.11}$$

and

$$\frac{1}{\sqrt{m}} (\boldsymbol{Z}^{(1)})^\top \mathbf{\Psi}^1_j = \begin{cases} \mathbf{\Psi}^0_j & \text{if } j \text{ is odd,} \\ \mathbf{\Psi}^0_j + \mathbf{\Psi}^0_{j/2} & \text{if } j \text{ is even,} \end{cases} \tag{6.12}$$

$$\frac{1}{\sqrt{m}} \boldsymbol{Z}^{(2)} \mathbf{\Psi}^1_j = \begin{cases} \mathbf{\Psi}^2_j & \text{if } j \text{ is even,} \\ \mathbf{\Psi}^2_j + \mathbf{\Psi}^2_{(j-1)/2} & \text{if } j \text{ is odd.} \end{cases} \tag{6.13}$$

Thanks to (6.10)-(6.11)-(6.12)(6.13), the matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ in (6.5) can be determined, which are the only missing unknowns to be able to obtain an evolution of the system (6.5):

$$(\Lambda_1)_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } 2i = j, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad (\Lambda_2)_{ij} = \begin{cases} 1 & \text{if } i = j \text{ or } 2j+1 = i, \\ 0 & \text{otherwise,} \end{cases}$$

that is,

$$\mathbf{\Lambda}_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & \\ & & & & \vdots & & & & & \ddots \end{pmatrix},$$

and

$$\mathbf{\Lambda}_2^\top = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & \\ & & & & \vdots & & & & & \ddots \end{pmatrix}.$$

References

[1] David J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[3] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2018.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Francis Bach and Lénaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. *arXiv preprint arXiv:2110.08084*, 2021.

[6] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

[7] Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

[8] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[9] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[10] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

[11] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

[12] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

[13] Nadav Cohen, Govind Menon, and Zsolt Veraszto. Deep linear networks for matrix completionâ€"an infinite depth limit. *SIAM Journal on Applied Dynamical Systems*, 22(4):3208–3232, 2023.

[14] Amit Daniely. SGD learns the conjugate kernel class of the network. *Advances in Neural Information Processing Systems*, 30, 2017.

[15] Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.

[16] Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31, 2018.

[17] Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[18] Weinen E, Chao Ma, Lei Wu, and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *CSIAM Trans. Appl. Math.*, 1:561–615, 2020.

[19] Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In *International Conference on Machine Learning*, pages 2836–2847. PMLR, 2020.

[20] Xavier Fernández-Real and Alessio Figalli. The continuous formulation of shallow neural networks as wasserstein-type gradient flows. In Arthur Avila, Michael Rassias, and Sinai Yakov, editors, *Analysis at Large*. Springer, 2022.

[21] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[22] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Eugene Golikov. Towards a general theory of infinite-width limits of neural classifiers. In *International Conference on Machine Learning*, pages 3617–3626. PMLR, 2020.

[24] Eugene Golikov and Greg Yang. Non-gaussian tensor programs. *Advances in Neural Information Processing Systems*, 35, 2022.

[25] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[26] Karl Hajjar, Lénaïc Chizat, and Christophe Giraud. Training integrable parameterizations of deep neural networks in the infinite-width limit. *arXiv preprint arXiv:2110.15596*, 2021.

[27] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[29] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

[30] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[31] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

[32] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

[33] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[34] Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

[35] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[36] Sewoong Oh, Soumik Pal, Raghav Somani, and Raghav Tripathi. Gradient flows on graphons: existence, convergence, continuity equations. *arXiv preprint arXiv:2111.09459*, 2021.
[37] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.
[38] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in Neural Information Processing Systems*, 33:21174–21187, 2020.
[39] Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems*, 2018.
[40] Andrew M Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.
[41] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.
[42] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
[43] Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.
[44] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.
[45] Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

EPFL SB MATH, Institute of Mathematics, Station 8, CH-1015 Lausanne, Switzerland
Email address: lenaic.chizat@epfl.ch

EPFL SB MATH, Institute of Mathematics, Station 8, CH-1015 Lausanne, Switzerland
Email address: maria.colombo@epfl.ch

EPFL SB MATH, Institute of Mathematics, Station 8, CH-1015 Lausanne, Switzerland
Email address: xavier.fernandez-real@epfl.ch

ETH Zurich, Department of Mathematics, Rämistrasse 101, 8092 Zürich, Switzerland
Email address: alessio.figalli@math.ethz.ch