On the Impact of Overparameterization on the Training of a Shallow Neural Network in High Dimensions

Simon Martin, Francis Bach, Giulio Biroli

On the Impact of Overparameterization on the Training of a Shallow Neural Network in High Dimensions

Simon Martin[1, 2] Francis Bach[1] Giulio Biroli[2]

[1]INRIA - Ecole Normale Supérieure, PSL Research University

[2]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

Abstract

We study the training dynamics of a shallow neural network with quadratic activation functions and quadratic cost in a teacher-student setup. In line with previous works on the same neural architecture, the optimization is performed following the gradient flow on the population risk, where the average over data points is replaced by the expectation over their distribution, assumed to be Gaussian.

We first derive convergence properties for the gradient flow and quantify the overparameterization that is necessary to achieve a strong signal recovery. Then, assuming that the teachers and the students at initialization form independent orthonormal families, we derive a high-dimensional limit for the flow and show that the minimal overparameterization is sufficient for strong recovery. We verify by numerical experiments that these results hold for more general initializations.

1. Introduction

While neural networks have revolutionized various domains such as image recognition (Krizhevsky et al., 2017), image generation (Goodfellow et al., 2014), and natural language processing (Gozalo-Brizuela and Garrido-Merchan, 2023), a strong gap remains between their practical achievements and the theoretical understanding of their behaviours. In addition to the formal guarantees that such an understanding can provide, new theoretical insights may lead to an improvement of optimization algorithms as well as the development of more robust and reliable models.

One main obstacle to a general theoretical study of neural networks is the fact that they are characterized by highly non-convex loss functions. As a consequence, trajectories of gradient-based algorithms and their convergence properties are often hard to understand. Indeed, even for shallow architectures, it is known that loss landscapes of neural networks possess spurious local minima in which parameters can be trapped during optimization (Baity-Jesi et al., 2018; Choromanska et al., 2015; Christof and Kowalczyk, 2023; Safran and Shamir, 2018). However, some recent works show that in some limits, either those local minima tend to disappear from the landscape, or gradient-based algorithms avoid them despite their presence. This is for instance the case in highly overparameterized networks (Chizat and Bach, 2018; Du et al., 2019; Mei et al., 2019) or when optimizing with a very large amount of data (Du et al., 2018; Li and Yuan, 2017; Tian, 2017) or in high-dimensional inference (Mannelli et al., 2019). Note that those results are often obtained in purely theoretical limits, where the number of neurons or data points go to infinity. A key open question to determine is how many neurons are needed to achieve global optima; this paper provides an answer for an idealized problem.

1.1. Contributions

We investigate the effect of overparameterization on the training of a shallow neural network in a teacher-student setup. Our model is a one-hidden layer neural network with quadratic activation functions. The optimization is performed on the population loss (i.e., in the infinite data limit) under the assumption of Gaussian data points. While we assume without loss of generality that the number of neurons of the teacher network (denoted m^*) is less than the dimension d, we do not constraint the number of neurons of the student (denoted m). In this setup:

- In Section 3 we derive a general solution of the gradient flow depending on an unknown scalar function which is solution of an implicit equation.
- In Section 4 we show that the gradient flow always converges to the global minimizer of the loss. In the case where the student network has more neurons m than the teacher, this global optimum corresponds

to a perfect recovery of the teacher network. In the overparameterized case, i.e., $m \geq m^*$, we derive tight convergence rates for the gradient flow.

- In Section 5, assuming that the teacher vectors and students at initialization form orthonormal families, drawn from an appropriate distribution, we derive a high-dimensional limit for the flow. In this limit, we show that strong recovery is achieved as soon as there are more students than teachers. A key feature of our analysis is to go beyond the case $m = m^* = 1$ by letting these two quantities to grow with dimension d.

1.2. Related Works

Shallow neural networks with quadratic activations. One hidden-layer neural networks with quadratic activation functions have already been studied in the literature. Du and Lee (2018) as well as Soltanolkotabi et al. (2018) focus on the empirical loss and derive landscape properties in the overparameterized case. More precisely, Du and Lee (2018) obtain that for $m \geq d$, the landscape does not admit spurious local minimizers in the case where the output weights are fixed (which corresponds to our setup), and Soltanolkotabi et al. (2018) derive the same property for $m \geq 2d$, if the output weights can also be learned. This result is also obtained by Venturi et al. (2019) for the case of the population loss.

Moreover, Gamarnik et al. (2019) and Mannelli et al. (2020b) worked in the exact same setup as ours. The main differences are that Gamarnik et al. (2019) focus on the case where students and teachers have more neurons than the dimension ($m, m^* \geq d$), and Mannelli et al. (2020b) only assume that $m \geq d$. In this overparameterized setting, Mannelli et al. (2020b) show that the gradient flow on the population loss converges towards optimum and derives a convergence rate. They also show rigorously that the gradient flow on the empirical loss leads to a minimizer with optimal prediction for a number of sample larger than $2d$ for $m^* = 1$ (and heuristically $(m^* + 1)d$ for $m^* > 1$). Finally, Gamarnik et al. (2019), only considering sub-Gaussian observations, proves that the minimal value of both the empirical and population loss over rank deficient matrices is bounded away from zero with high probability. In addition, they obtain a generalization bound for the weights optimized on the empirical loss.

Phase retrieval. Another topic related to ours is the phase retrieval problem, which corresponds to the simpler case where $m = m^* = 1$. This problem has been extensively studied in the literature (see Fienup, 1982, for a review). As a result of interest, we mention Mannelli et al. (2020a), who exhibit a phase transition for the number of observations per dimension in order to achieve strong recovery. This criterion is obtained in the mean-field limit, where the dimension goes jointly to infinity with the number of observations, with a constant ratio.

High-dimensional limits. More generally, several learning properties of shallow neural networks have been obtained through high-dimensional limits. For instance, in the case of a one hidden layer neural network, Berthier et al. (2023) obtain a set of low dimensional equations for the gradient flow dynamics in the mean-field limit, and Arnaboldi et al. (2023) derive similar equations for the SGD algorithm with different scaling between the dimension, the number of neurons and the stepsize. Other works rely on the use of statistical physics methods to derive high-dimensional equations for the learning dynamics of neural networks (Gabrié et al., 2023; Gamarnik et al., 2022; Mannelli et al., 2020a; Mignacco et al., 2021).

2. Setting

2.1. Notations

For a matrix $A \in \mathbb{R}^{d\times m}$, we denote $A^T \in \mathbb{R}^{m\times d}$ its transpose and $\|A\|_F = [\mathrm{Tr}(AA^T)]^{1/2}$ its Frobenius norm. We denote $\mathcal{S}_d(\mathbb{R})$, $\mathcal{S}_d^+(\mathbb{R})$, $\mathcal{S}_d^{++}(\mathbb{R})$ the spaces of $d \times d$ symmetric, positive semi-definite and positive definite matrices. If $\big(E, \langle\cdot,\cdot\rangle\big)$ is a Euclidean space and $\Phi : E \to \mathbb{R}$ is twice continuously differentiable, we denote $\nabla L(x) \in E$ its gradient and $\mathrm{d}^2 L_x$ its Hessian at x, which is a linear self-adjoint map on E. We say that $x \in E$ is a critical point of L if $\nabla L(x) = 0$, and a local minimizer if in addition, $\langle \mathrm{d}^2 L_x(h), h\rangle \geq 0$ for all $h \in E$.

2.2. Model

We consider a teacher-student setup where the input $x \in \mathbb{R}^d$ is randomly generated and fed to a teacher network u^* whose output is learned by a student u with the same structure, with:

$$u^*(x) = \frac{1}{m^*}\sum_{j=1}^{m^*}(w_j^* \cdot x)^2 = \mathrm{Tr}\big(xx^T W^* W^{*T}\big),$$
$$u(x) = \frac{1}{m}\sum_{j=1}^{m}(w_j \cdot x)^2 = \mathrm{Tr}\big(xx^T WW^T\big), \tag{1}$$

where m and m^* are the number of neurons of the student and the teacher networks, $(w_j)_{1\leq j\leq m} \in \big(\mathbb{R}^d\big)^m$

and $(w_j^*)_{1\leq j\leq m^*} \in (\mathbb{R}^d)^{m^*}$ are the corresponding weights, and W, W^* the associated renormalized matrices:

$$W^* = \frac{1}{\sqrt{m^*}}\big(w_1^* \,\big|\, \dots \,\big|\, w_{m^*}^*\big) \in \mathbb{R}^{d\times m^*},$$
$$W = \frac{1}{\sqrt{m}}\big(w_1 \,\big|\, \dots \,\big|\, w_m\big) \in \mathbb{R}^{d\times m}. \tag{2}$$

This common architecture corresponds to a one hidden layer neural network where the output weights are fixed, with quadratic activation functions.

The error made by the student network is evaluated using the quadratic cost function. In line with previous work on this specific model (see Gamarnik et al., 2019; Mannelli et al., 2020b), our study exclusively focuses on the population risk, which is obtained by taking the expectation over the data points distribution, assumed to be Gaussian with zero mean and identity covariance matrix:

$$\begin{aligned}\mathcal{L}(W) &= \frac{1}{4}\mathbb{E}_x\Big[\mathrm{Tr}\Big(xx^T(W^*W^{*T} - WW^T)\Big)^2\Big] \\ &\equiv \frac{1}{2}\big\|\Delta_W\big\|_F^2 + \frac{1}{4}\big[\mathrm{Tr}(\Delta_W)\big]^2,\end{aligned} \tag{3}$$

with $\Delta_W = WW^T - W^*W^{*T} \in \mathbb{R}^{d\times d}$. As shown by Gamarnik et al. (2019, Theorem 3.1), equation (3) is verified whenever x has i.i.d. coordinates drawn from a distribution matching the standard Gaussian distribution (with mean 0 and variance 1) up to the fourth moment.[1]

Although the loss $\mathcal{L}$ is non convex, it can be written as $\mathcal{L}(W) = F(WW^T)$ where F is convex. Some results derived in Section 4 will show that $\mathcal{L}$ enjoys some properties shared by convex functions. For instance, we show that all its local minimizers are in fact global. Such functions depending only on WW^T have already been studied with various approaches, see Edelman et al. (1998); Journée et al. (2010); Massart and Absil (2020).

Optimization on W is performed using the gradient flow, which has the form:

$$\dot{W} = -\nabla\mathcal{L}(W) = -2\Delta_W W - \mathrm{Tr}(\Delta_W)W. \tag{4}$$

Note that, due to the form of $\mathcal{L}$, we do not expect to exactly retrieve through optimization the students vectors $w_1^*, \dots, w_{m^*}^*$, but rather to obtain a student matrix W such that $WW^T = W^*W^{*T}$. As shown in equation (1), such a matrix will lead to an optimal predictor. If W^* is of rank m^* and since $W \in \mathbb{R}^{d\times m}$, this is only possible for $m \geq m^*$. Only in this case, one can hope for a convergence of the flow towards strong recovery.

In the following, we will suppose that the initialization of the flow $W^0 = W(t=0) \in \mathbb{R}^{d\times m}$ as well as the teacher matrix $W^* \in \mathbb{R}^{d\times m^*}$ are random. When necessary, we will make assumptions about their distributions. In Section 4, we will assume these matrices to be drawn from a distribution which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$ and $\mathbb{R}^{d\times m^*}$. In Section 5, we will require the initial students and the teachers to be orthonormal families. In general, we will always assume the independence of the matrices $W^0 = W(t=0)$ and W^*.

We restrict our study to the case $m^* \leq d$. Indeed, whenever $m^* > d$, the matrix $Z^* = W^*W^{*T} \in \mathcal{S}_d^+(\mathbb{R})$ can be determined using only d vectors, thus the optimization problem will be equivalent to the case $m^* = d$. We will assume Z^* to be of rank m^* (that is the teacher vectors form a linearly independent family of $\mathbb{R}^d$), and denote $\mu_1 \geq \dots \geq \mu_{m^*} > 0$ its non-zero eigenvalues, and $v_1^*, \dots, v_{m^*}^* \in \mathbb{R}^d$ the associated orthonormal eigenvectors, so that:

$$Z^* = \sum_{k=1}^{m^*} \mu_k v_k^*(v_k^*)^T.$$

3. Self-Consistent Solution of the Flow

In this section, we present a first result for the gradient flow dynamics defined in equation (4). This flow is non-linear and associated with a non-convex loss, thus we cannot rely on general results to understand its dynamics. To the exception of Mannelli et al. (2020b) who gave an interpretation of the solution of the flow using a stochastic differential equation, and derived convergence properties in the over-parameterized case $m \geq d$, the dynamical properties of the flow in the general case have not been studied before.

However, a similar problem, the Oja's flow, was treated by Yan et al. (1994). It corresponds to the gradient flow associated with the loss $\mathcal{L}^{\mathrm{oja}}(W) = \frac{1}{2}\|\Delta_W\|_F^2$ (where the second term in equation (3) is removed), leading to the differential equation $\dot{W} = -2\Delta_W W$. They show that this flow admits a closed-form solution and convergence properties can be deduced from their formula.

In the following proposition, we obtain a solution of the

[1] Gamarnik et al. (2019) show that in the general case, an extra term appears in the loss, proportional to $\kappa_4\|\mathrm{diag}(\Delta_W)\|_2^2$, where κ_4 is the fourth cumulant of the distribution from which the coefficients of x are drawn (which is zero in the Gaussian case), and $\mathrm{diag}(\Delta_W)$ is the vector of the diagonal elements of Δ_W. This leads to a different flow which is not covered by our results.

flow very similar to the one of the Oja's flow. As done by Yan et al. (1994), this solution is expressed in terms of the matrix $Z(t) = W(t)W(t)^T$. However, in our case the solution also depends on a function ψ which is solution of an implicit equation.

Proposition 3.1. *Let $W(t)$ be solution of the flow in equation (4) with initial condition $W^0 \in \mathbb{R}^{d\times m}$. Let $Z(t) = W(t)W(t)^T$. Then, for all $t \geq 0$:*

$$Z(t) = M^*(t)\left(I_m + 4\int_0^t M^*(s)^T M^*(s)\mathrm{d}s\right)^{-1} M^*(t)^T, \tag{5}$$

with:

$$M^*(t) = e^{-\psi(t)} e^{t\mathrm{Tr}(Z^*)} e^{2Z^* t} W^0 \in \mathbb{R}^{d\times m},$$

and $\psi(t) = \int_0^t \mathrm{Tr}\big(W(s)W(s)^T\big)\,\mathrm{d}s$, solution of the equation:

$$\psi(t) = \frac{1}{4}\mathrm{Tr}\log\left(I_m + 4\int_0^t M^*(s)^T M^*(s)\mathrm{d}s\right), \tag{6}$$

where the log *is understood as the spectral map on $\mathcal{S}_m^{++}(\mathbb{R})$.*

Proof. To derive equation (5), note that if $W(t) \in \mathbb{R}^{d\times m}$ is solution of equation (4), then $Z(t) = W(t)W(t)^T \in \mathcal{S}_d^+(\mathbb{R})$ is solution of:

$$\dot{Z} = 2\mathrm{Tr}(Z^* - Z)Z + 2Z^*Z + 2ZZ^* - 4Z^2.$$

The result is obtained by deriving equation (5) along with the definition of ψ. For equation (6), we have:

$$\dot{\psi}(t) = \mathrm{Tr}[Z(t)] = \frac{1}{4}\mathrm{Tr}\big[\dot{U}(t)U(t)^{-1}\big],$$

with $U(t) = I_m + 4\int_0^t M^*(s)M^*(s)^T\mathrm{d}s$. Since the map $X \mapsto \mathrm{Tr}\log X$ has gradient X^{-1} on $\mathcal{S}_m^{++}(\mathbb{R})$, this leads to the result by integrating the previous equation. □

This proposition gives an implicit formula for the solution of the flow $W(t)$, but in terms of the matrix $Z(t) = W(t)W(t)^T$. Obtaining an understanding of $W(t)$ in itself is not needed: as mentioned in subsection 2.2, the signal recovery is achieved whenever $WW^T = W^*W^{*T}$.

The first consequence of formula (5) is that if W^0 is of full rank, i.e., $\mathrm{rank}(W^0) = \min(m, d)$, then $W(t)$ stays of full rank along the flow. Thus, due to the lower semicontinuity of the rank, $W_\infty = \lim_{t\to\infty} W(t)$ will be of rank $\leq \min(m, d)$, provided that the limit exists. As a consequence, we cannot hope for a full signal recovery (i.e., $Z(t) \xrightarrow[t\to\infty]{} Z^*$) whenever $m < m^*$.

Equation (6) defines an implicit equation on ψ. This scalar function is the only unknown of the solution obtained in equation (5), therefore, gathering information about its behaviour will be useful to understand the properties of the flow. However, equation (6) cannot be solved in closed form in general cases, but it can be analyzed in the high-dimensional limit and solved numerically. In Section 4, we first obtain general results on the convergence properties at long times by directly studying the local minimizers of the loss function. We then work out the high-dimensional limit in Section 5.

4. Convergence of the Gradient Flow

We now determine the convergence properties of the gradient flow defined in equation (4). In Proposition 4.1, we derive the limit of the solution $W(t)$ of the flow as $t \to \infty$. We also mention the rate associated with this convergence, a result we detail in Appendix A.1.

The first result makes use of the stable manifold theorem (see Smale, 2011), which states that, under the following assumption, the gradient flow almost surely converges towards a local minimizer of the loss function.

Assumption 4.1. *The initialization of the flow W^0 is drawn with a distribution which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$.*

Proposition 4.1. *Let $W(t)$ be solution of the flow (4) with initial condition W^0. Then, under Assumption 4.1, with probability one, $W_\infty = \lim_{t\to\infty} W(t)$ exists and verifies:*

- *if $m \geq m^*$: $W_\infty W_\infty^T = Z^*$,*
- *if $m < m^*$ and the non-zero eigenvalues of Z^* are simple (i.e., $\mu_1 > \cdots > \mu_{m^*} > 0$):*

$$W_\infty W_\infty^T = \sum_{k=1}^{m} (\mu_k + \tau) v_k^* (v_k^*)^T,$$

where τ is defined as:

$$\tau = \frac{1}{m+2}\sum_{k=m+1}^{m^*} \mu_k. \tag{7}$$

This result ensures a strong recovery of the signal by the gradient flow when $m \geq m^*$, which is the minimal overparameterization one needs. Indeed, as mentioned earlier, perfect recovery is impossible when $m < m^*$,

due to rank constraints. Instead we recover only the largest eigenvectors.

We prove this proposition in Appendix C.1 using that, under Assumption 4.1, one only needs to determine the local minimizers associated with the loss, that is solving the equations:

$$\nabla\mathcal{L}(W) = 0, \qquad \mathrm{Tr}\big(\mathrm{d}^2\mathcal{L}_W(K)K^T\big) \geq 0,$$

for all $K \in \mathbb{R}^{d\times m}$, which is much easier than studying the full flow using the solution in Proposition 3.1. The proof is done in two steps: we first characterize the critical points of the loss, i.e., the matrices $W \in \mathbb{R}^{d\times m}$ verifying $\nabla\mathcal{L}(W) = 0$. Then, we select those satisfying the second optimality condition.

In the case $m < m^*$, the assumption on the simplicity of the eigenvalues of Z^* is not mandatory, but it allows to obtain a single local minimizer up to the representation $W \longmapsto WW^T$. For instance, if all of the non-zero eigenvalues of Z^* are equal to some $\mu > 0$, which happens for instance when the teachers are orthogonal, then for each subset $I \subset [\![1, m^*]\!]$ of size m, the matrix:

$$\sum_{k\in I}(\mu+\tau)v_k^* v_k^{*T},$$

corresponds to a local minimizer of the loss. Overall, each choice of subset I leads to the same value of the loss, which allows to conclude that every local minimizer of $\mathcal{L}$ is global. For $m \geq m^*$, they are optimal and achieve zero error on the loss.

Another natural question regarding the long time behaviour of the gradient flow is the understanding of the convergence rate. Proposition A.1 in Appendix A.1 derives them in the overparameterized case $m \geq m^*$, improving the results obtained by Mannelli et al. (2020b). Some numerical experiments, also featured in Appendix A.1 prove our bounds to be tight.

5. High-Dimensional Limit

We now investigate the high dimensional limit of the flow defined in equation (4). Originally used in statistical physics and mean-field theories where the dimension accounts for the number of degrees of freedom (going to infinity in the thermodynamic limit), high dimensional limits have been widely applied to learning and optimization problems (see Gabrié et al., 2023; Gamarnik et al., 2022). They allow to obtain equations where the main objects (called order parameters in physics) are finite-dimensional and satisfy self-consistent equations. In some cases, one also finds that the relevant random quantities concentrate, thus leading to a deterministic description in the $d \to \infty$ limit, only depending on the distribution of the randomness associated with the initialization or the signal. We show below that this is indeed what happens for the flow by taking the limit $m, m^*, d \to \infty$ with an appropriate scaling.

In this section, the main goals are to (i) identify the limiting deterministic dynamical equations describing the $d \to \infty$ limit, (ii) determine the timescale t_d at which convergence occurs, and (iii) characterize the behavior of a scalar quantity accounting for the signal retrieval, which is the overlap between the students and teachers:

$$\chi(Z, Z^*) = \frac{\big|\mathrm{Tr}(ZZ^*)\big|}{\|Z\|_F\|Z^*\|_F} \in [0,1], \tag{8}$$

with $Z = WW^T$. This quantity depends on time and dimension, and the first challenge is to derive the relevant scaling between those two parameters.

5.1. Sampling Assumptions

In all the following, we make the assumption that the families $(w_1^0, \dots, w_m^0)$ and $(w_1^*, \dots, w_{m^*}^*)$ are orthonormal and drawn independently. We also assume that they are uniformly drawn, which can be achieved by considering the uniform measure on the Stiefel manifold (see Götze and Sambale, 2023). From now on, the dependency in the dimension is made explicit when relevant. Thus, defining $U_d^0 = \sqrt{m_d}\,W_d^0$ and $U_d^* = \sqrt{m_d^*}\,W_d^*$ (the initialization W_d^0 and the teacher matrix W_d^* are linked to their respective vectors in equation (2)), we assume that:

$$U_d^{0T}U_d^0 = I_{m_d}, \qquad U_d^{*T}U_d^* = I_{m_d^*}.$$

The orthonormality assumption will help simplify the computations, as the relevant quantities that will be studied in the high dimensional limit will only depend on the matrix $Y_d = U_d^{0T}U_d^*U_d^{*T}U_d^0 \in \mathbb{R}^{m_d\times m_d}$. Indeed, we will make use of the implicit solution for the flow obtained in equation (5), which depends on the matrix $W_d^{0T}\exp\big(4W_d^*W_d^{*T}t\big)W_d^0$. Under our orthonormality assumption, we have the more convenient formula:

$$U_d^{0T}\exp\big(\lambda U_d^*U_d^{*T}\big)U_d^0 = I_{m_d} + \big(e^\lambda - 1\big)Y_d, \tag{9}$$

which is specific to our choice of distribution. Moreover, the limit distribution for the eigenvalues of the random matrix Y_d is well understood (see Aubrun, 2021; Hiai and Petz, 2005; Kunisky, 2023).

For obvious dimensional reasons, we necessarily have $m_d, m_d^* \leq d$. As $d \to \infty$, we will assume that m_d and

m_d^* diverge but keep a fixed ratio with d:

$$\alpha = \lim_{d\to\infty} \frac{m_d}{d} \in]0,1], \qquad \alpha^* = \lim_{d\to\infty} \frac{m_d^*}{d} \in]0,1].$$

Thus, we let the number of teachers and students to vary with the dimension, leading to predictors parameterized by a continuum of neurons.

In this setting, the matrix Y_d has a convergent empirical spectral distribution, namely there exists a probability measure μ (depending on α, α^*) supported on $[0,1]$, such that for any continuous $f : \mathbb{R} \to \mathbb{R}$:

$$\frac{1}{m_d}\mathrm{Tr}\big(f(Y_d)\big) \xrightarrow[d\to\infty]{} \int f(x)\mathrm{d}\mu(x).$$

This convergence result is presented by Kunisky (2023, Theorem 1.7), and the distribution μ is known as a limiting distribution in the $\beta-$Jacobi ensemble (see Collins, 2005; Jiang, 2013):

$$\begin{aligned}\mathrm{d}\mu(x) &= \frac{\sqrt{(r_+ - x)(x - r_-)}}{2\pi\alpha x(1-x)}\mathbb{1}_{[r_-,r_+]}(x)\mathrm{d}x \\ &+ \Big(1 - \frac{\alpha^*}{\alpha}\Big)^+\delta_0(x) + \frac{1}{\alpha}\big(\alpha+\alpha^*-1\big)^+\delta_1(x),\end{aligned} \tag{10}$$

where $u^+ = \max(u, 0)$ and:

$$r_\pm = \Big(\sqrt{\alpha(1-\alpha^*)} \pm \sqrt{\alpha^*(1-\alpha)}\Big)^2.$$

In Figure 1, we compare the probability density function of μ (when the coefficients associated with the Dirac measures are zero) and its empirical counterpart when drawing the eigenvalues of Y_d in finite dimension. This

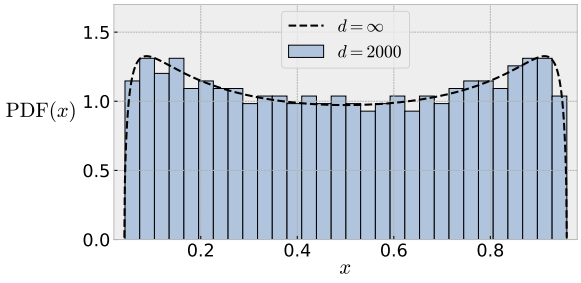

Figure 1: Probability density function of μ defined in equation (10) (dashed line) and histogram of the eigenvalues of Y_d for $d = 2000$ with 30 bins, with $\alpha = 0.3$, $\alpha^* = 0.5$.

convergence result is a key ingredient to characterize the high-dimensional limit. In fact, as shown in Section 3, the solution of the flow derived in equation (5) mainly relies on the unknown function ψ_d, which is solution of the implicit equation (6). Our first result identifies the high-dimensional limit of ψ_d defined in Proposition 3.1.

5.2. High-Dimensional Limit of the Flow

In order to investigate signal recovery in the high-dimensional limit, we first determine the limit equations solved by ψ_d. The convergence will be shown to happen at timescale $t \sim d$, we therefore allow time to vary with dimension by setting $t_d = \gamma d$, where γ is fixed and accounts for a renormalized time variable. We also set:

$$\begin{aligned}\phi_d(\gamma) &= \int_0^{\gamma d} \mathrm{Tr}\big(W_d(s)W_d(s)^T - W_d^* W_d^{*T}\big)\mathrm{d}s \\ &= \psi_d(\gamma d) - \gamma d,\end{aligned} \tag{11}$$

since under our assumptions, we have $\mathrm{Tr}(Z_d^*) = 1$. Note that we cannot hope for a high-dimensional limit for ψ_d in itself: in the case of a signal recovery at finite dimension, one can expect $\psi_d(\gamma d)$ to be of the order d, whereas $\phi_d(\gamma)$ should remain of order 1. That is indeed what we show in the following proposition.

Proposition 5.1. *Let $W_d(t)$ be solution of the gradient flow in equation* (4) *with initial condition W_d^0. Then, with probability one, as $d \to \infty$, ϕ_d uniformly converges on any compact of $\mathbb{R}_+$ towards a function ϕ which is solution of the equation:*

$$\begin{aligned}4\gamma = \alpha \log F_\phi(\gamma) + \big(\alpha+\alpha^*-1\big)^+ \log\left(\frac{G_\phi(\gamma)}{F_\phi(\gamma)}\right) \\ + \Theta\left(\frac{G_\phi(\gamma)}{F_\phi(\gamma)} - 1\right),\end{aligned} \tag{12}$$

with:

$$F_\phi(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}\mathrm{d}s,$$

$$G_\phi(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}\mathrm{d}s,$$

$$\Theta(u) = \frac{1}{2\pi}\int_{r_-}^{r_+} \frac{\sqrt{(r_+ - x)(x - r_-)}}{x(1-x)}\log(1+ux)\mathrm{d}x.$$

This result is proven in Appendix C.3. The idea is to start from equation (6) which allows to obtain an implicit equation of ϕ_d. The key observation is that this equation only depends on the matrix Y_d (this is a consequence of equation (9)).

The equation above, in which the dimension only appears through the parameters α and α^*, allows to analyze the high-dimensional limit of the flow.

Although it cannot be solved analytically, it is to be compared with the implicit equation (6) on ψ_d: this new equation is purely deterministic and does not depend on the initialization of the flow. We can obtain an

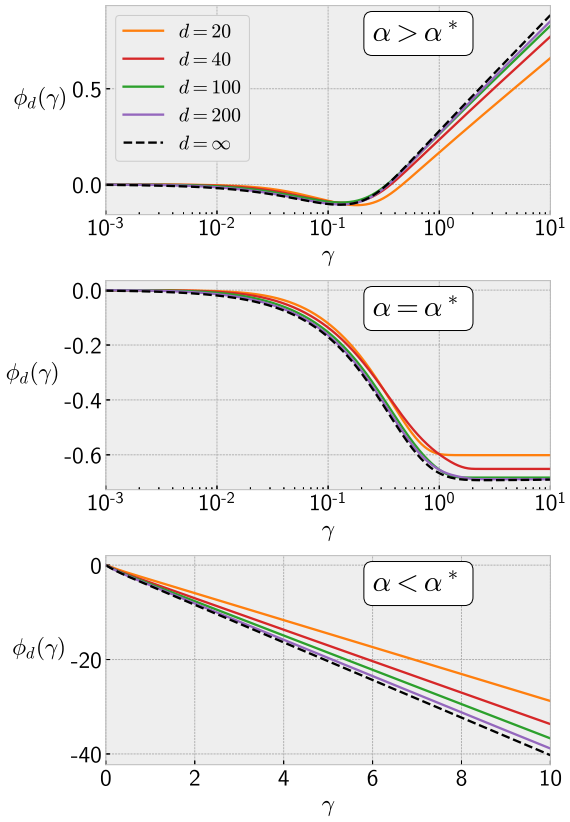

Figure 2: Simulations of $\phi_d(\gamma)$ (defined in equation (11)) with $\gamma = t/d$ for increasing values of d. Finite d: simulation using gradient descent with stepsize $\eta = 10^{-2}$, orthonormal teachers and students at initialization. $d = \infty$: simulations using discretized version of equation (12). Top: $m_d = d/2$, $m_d^* = d/4$, middle: $m_d = d/2$, $m_d^* = d/2$, bottom: $m_d = d/4$, $m_d^* = d/2$. Results averaged over 5 simulations.

approximate numerical solution of ϕ by differentiating this equation with respect to γ, and obtaining an equation of the form:

$$\phi(\gamma) = \Omega\big(F_\phi(\gamma), G_\phi(\gamma), \gamma\big).$$

From the expression of F_ϕ and G_ϕ, this can be numerically handled as a standard differential equation. We have compared this numerical solution to the one obtained by a direct solution of the flow equations in finite dimension. Figure 2 displays the evolution of the function ϕ_d for increasing values of d, as well as the numerical solution of equation (12). It shows a quite fast convergence with the dimension (determining the convergence rate as $d \to \infty$ is a challenge that we leave for future studies). Figure 2 also showcases very different behaviours for ϕ_d depending on the values of m_d and m_d^*: logarithmic for $m_d > m_d^*$, converging for $m_d = m_d^*$ and linear for $m_d < m_d^*$. Those behaviours of ϕ_d in the different regimes can be also understood from finite dimension estimation. In Appendix A.1, we determine an asymptotic development of the function $\psi(t)$ as $t \to \infty$, which allows to recover the behaviours displayed in Figure 2.

5.3. High-Dimensional Signal Recovery

We now investigate the question of the signal retrieval in the high-dimensional limit. Proposition 4.1 in Section 4 shows that strong recovery as $t \to \infty$ is guaranteed whenever $m \geq m^*$. The question is now whether this criterion still holds as $d \to \infty$. To this end, we introduce the overlap, a scalar quantity describing the alignment between the teachers and students:

$$\chi(Z_d(t), Z_d^*) = \frac{\big|\mathrm{Tr}\big(Z_d(t) Z_d^*\big)\big|}{\big\|Z_d(t)\big\|_F \big\|Z_d^*\big\|_F},$$

with $Z_d(t) = W_d(t) W_d(t)^T$. A strong signal recovery is obtained whenever $\chi(Z_d, Z_d^*) = 1$, corresponding to a perfect alignment. Moreover, we say that the flow achieves a weak recovery if $\chi(Z_d, Z_d^*)$ is larger than a purely random overlap χ_d^0, i.e., the overlap between two independent projection matrices. In the vector case, where it is commonly used (replacing the trace by the standard inner product on $\mathbb{R}^d$ and $\|\cdot\|_F$ by the Euclidean norm), this purely random overlap goes to zero as $d \to \infty$. However, in our case, we show that this quantity stays positive in the high dimensional limit. More precisely, χ_d^0 can be expressed as the overlap between the teachers and the students at initialization, which leads to the limit $\chi_d^0 \to \sqrt{\alpha \alpha^*}$, see Appendix A.3.

In the same spirit as Proposition 5.1, we determine the convergence of the overlap between the students and teachers as $d \to \infty$ and at timescale $t_d = \gamma d$.

Proposition 5.2. *Let $W_d(t)$ be solution of the gradient flow in equation (4) with initial condition W_d^0. Then, as $d \to \infty$, almost surely, the function $\chi_d(\gamma) = \chi\big(Z_d(\gamma d), Z_d^*\big)$ uniformly converges on every compact of $\mathbb{R}_+$. Let $\chi(\gamma) = \lim_{d \to \infty} \chi_d(\gamma)$. We have:*

$$\chi(\gamma) \xrightarrow[\gamma \to \infty]{} \min\left(\sqrt{\frac{\alpha}{\alpha^*}}, 1\right).$$

We prove this proposition in Appendix C.4. The proof

is a consequence of the convergence of the empirical measure associated with the eigenvalues of Y_d as well as the result of Proposition 5.1. This last result gives

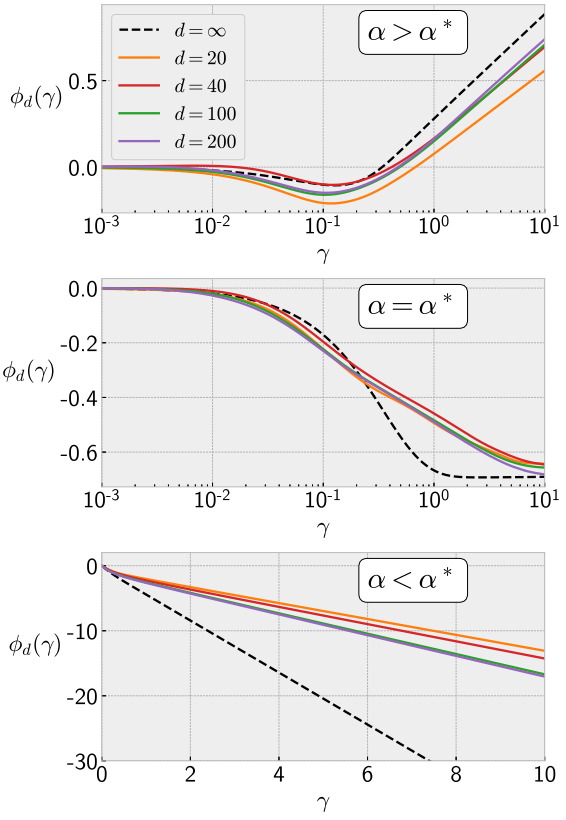

Figure 3: Simulations of $\phi_d(\gamma)$ (defined in equation (11)) with $\gamma = t/d$ for increasing values of d. Finite d: simulation using gradient descent with stepsize $\eta = 10^{-2}$, and Gaussian $\mathcal{N}(0, I_d/d)$ teachers and students at initialization. $d = \infty$: simulations using discretized version of equation (12) Top: $m_d = d/2$, $m_d^* = d/4$, middle: $m_d = d/2$, $m_d^* = d/2$, bottom: $m_d = d/4$, $m_d^* = d/2$. Results averaged over 5 simulations.

the behaviour of the overlap in the regime $t \gg d$. On this timescale, we obtain a strong recovery of the signal when $\alpha \geq \alpha^*$, and a weak recovery when $\alpha < \alpha^*$. This corresponds to the same threshold as the one obtained in Proposition 4.1. More precisely, when the teachers are orthonormal, one can compute the overlap between Z_d^* and the limit $Z_d^\infty = \lim_{t\to\infty} Z_d(t)$ derived in Proposition 4.1 in finite dimension, and obtain the same result as $d \to \infty$, i.e., that the limits $d, t \to \infty$ commute:

$$\lim_{d\to\infty} \lim_{t\to\infty} \chi\big(Z_d(t), Z_d^*\big) = \lim_{\gamma\to\infty} \lim_{d\to\infty} \chi\big(Z_d(\gamma d), Z_d^*\big).$$

Moreover, still in the orthonormal case, one can show that the quantity obtained in the previous proposition realizes the maximum overlap possible for a given number of students. We detail this result in Appendix A.3.

The previous results are valid for teachers and initial students which are drawn orthonormally. However, we believe it still holds for a larger class of distributions. Indeed, the determination of the convergence rates at finite dimension in Appendix A.1 allows to obtain the asymptotic behaviour of the flow as $t \to \infty$, without any distributional assumption. Interestingly, those behaviours coincide with the ones obtained throughout the section. To support the conjectured broader generality of our results, we show in Figure 3 a simulation of the function ϕ_d (directly computed from the flow), where both teacher and students are independently drawn from a Gaussian distribution $\mathcal{N}(0, I_d/d)$. This is compared to the approximate numerical solution of equation (12), where they are drawn orthonormally. We find again a quite fast convergence for $d \to \infty$, and a similar qualitative behaviour. However, the infinite dimensional limit of ϕ_d appears to be different from the orthonormal case. Generalizing the self-consistent analysis of Section 5 to the Gaussian case is left for future works.

6. Conclusion and Further Work

In this paper we presented new theoretical results on the optimization of one-hidden layer neural networks with quadratic activation functions. Focusing on the population loss gradient flow, we derived convergence properties and showed that a slight overparameterization is enough to achieve signal recovery. Then, we derived a high-dimensional limit for the flow and showed that our criterion still holds whenever the initialized students and teachers are orthonormal families.

Further work. The assumptions we made along this paper leaves several challenges of interest:

- As mentioned previously, we believe that the orthonormality assumption we made throughout Section 5 is not mandatory and that the results of this section can be generalized to a larger class of distributions, as suggested by Figure 3.
- Our study essentially focused on the optimization of the population loss. The next step is to understand the gradient flow associated with the empirical loss on a finite dataset. This extension has already been studied in several papers (see Du and Lee, 2018; Gamarnik et al., 2019; Mannelli et al., 2020b), but very few results were obtained regarding the dynamics of the flow. One promising strategy relies on the use of statistical physics methods, and more precisely

the dynamical mean-field theory, which allows to obtain a low-dimensional set of equations describing the dynamics in the limit where the dimension and the number of samples jointly go to infinity (see Ben Arous et al., 2006; Mignacco et al., 2021).

- More generally, it is of a high interest to obtain similar results for general activation functions, although several methods we used throughout this work are specific to the quadratic activation.

Acknowledgements

The authors acknowledge support from the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA0001 (PRAIRIE 3IA Institute). SM also thanks Louis-Pierre Chaintron for fruitful mathematical discussions.

References

L. Arnaboldi, L. Stephan, F. Krzakala, and B. Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. Technical Report 2302.05882, arXiv, 2023.

G. Aubrun. Principal angles between random subspaces and polynomials in two free projections. *Confluentes Mathematici*, 13(2):3–10, 2021.

M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. Ben Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, pages 314–323, 2018.

G. Ben Arous, A. Dembo, and A. Guionnet. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probab. Theory Relat. Fields*, 136(4): 619–660, Dec. 2006.

R. Berthier, A. Montanari, and K. Zhou. Learning time-scales in two-layers neural networks. Technical Report 2303.00055, arXiv, 2023.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204, 2015.

C. Christof and J. Kowalczyk. On the omnipresence of spurious local minima in certain neural network training problems. *Constructive Approximation*, June 2023.

B. Collins. Product of random projections, Jacobi ensembles and universality problems arising from free probability. *Probability Theory and Related Fields*, 133(3):315–344, Nov. 2005.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

S. S. Du and J. D. Lee. On the power of over-parametrization in neural networks with quadratic activation. In *Proceedings of the International Conference on Machine Learning*, pages 1329–1338, 2018.

S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the International Conference on Machine Learning*, pages 1339–1348, 2018.

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

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. & Appl.*, 20(2):303–353, Jan. 1998.

J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21(15):2758, Aug. 1982.

M. Gabrié, S. Ganguli, C. Lucibello, and R. Zecchina. Neural networks: from the perceptron to deep nets. Technical Report 2304.06636, arXiv, 2023.

D. Gamarnik, E. C. Kızıldağ, and I. Zadik. Stationary points of shallow neural networks with quadratic activation function. Technical Report 1912.01599, arXiv, 2019.

D. Gamarnik, C. Moore, and L. Zdeborová. Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114015, 2022.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

F. Götze and H. Sambale. Higher order concentration on stiefel and grassmann manifolds. *Electronic Journal of Probability*, 28:1–30, 2023.

R. Gozalo-Brizuela and E. C. Garrido-Merchan. Chatgpt is not all you need. a state of the art review of large generative AI models. Technical Report 2301.04655, arXiv, 2023.

F. Hiai and D. Petz. Large deviations for functions of two random projection matrices. Technical Report math/0504435, arXiv, 2005.

T. Jiang. Limit theorems for Beta-Jacobi ensembles. *Bernoulli*, 19(3):1028–1046, 2013.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM J. Optim.*, 20(5): 2327–2351, Jan. 2010.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017.

D. Kunisky. Generic MANOVA limit theorems for products of projections. Technical Report 2301.09543, arXiv, 2023.

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in Neural Information Processing Systems*, 30, 2017.

S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, and L. Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. *Advances in Neural Information Processing Systems*, 32, 2019.

S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová. Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval. In *Advances in Neural Information Processing Systems*, 2020a.

S. S. Mannelli, E. Vanden-Eijnden, and L. Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. In *Advances in Neural Information Processing Systems*, 2020b.

E. Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM J. Matrix Anal. Appl.*, 41(1):171–198, Jan. 2020.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the Conference on Learning Theory*, pages 2388–2464, 2019.

F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. *J. Stat. Mech.*, 2021(12):124008, Dec. 2021.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

I. Safran and O. Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441, 2018.

S. Smale. Stable manifolds for differential equations and diffeomorphisms. In *Topologia differenziale*, pages 93–126. Springer Berlin Heidelberg, 2011.

M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Y. Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *Proceedings of the International Conference on Machine Learning*, pages 3404–3413, 2017.

L. Venturi, A. S. Bandeira, and J. Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

W.-Y. Yan, U. Helmke, and J. B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Trans. Neural Netw.*, 5(5):674–683, Sept. 1994.

A. Additional Results

In this section we provide some additional results and insights. As mentioned in Section 4, we derive convergence rates in the overparameterized case $m \geq \min(m^*, d)$ in Appendix A.1. In Appendix A.2, we give a more detailed description of the results obtained for the limit distributions of product of projections that we introduce in Section 5.1. Finally, we discuss in Appendix A.3 the notion of overlap introduced in equation (8) and prove some results mentioned throughout Section 5.

A.1. Convergence Rates

Following the convergence result derived in Proposition 4.1, the natural question is to understand how fast this convergence is. In the following, we investigate the convergence rates associated with the gradient flow in the case $m \geq \min(m^*, d)$. The method we use in the proof can also be applied to the underparameterized setting $m < m^*$, but we choose to focus on the relevant case $m \geq \min(m^*, d)$, where the gradient flow converges towards the teacher matrix.

In the following proposition, the convergence rates are derived in terms of the loss, i.e., we obtain a bound on $\mathcal{L}(W(t))$ where $W(t)$ is solution of the gradient flow in equation (4). Due to the expression of the loss, this leads to a bound on the distance $\|W(t)W(t)^T - Z^*\|$.

Proposition A.1. *Let $W(t)$ be solution of the flow (4) with initial condition $W(t=0) = W^0$. Suppose that $m \geq \min(m^*, d)$. Denote μ the smallest non-zero eigenvalues of Z^*. Then, under Assumption 4.1, with probability one:*

- *If $m, m^* \geq d$:*
$$\mathcal{L}(W(t)) \underset{t\to\infty}{=} O\big(e^{-8\mu t}\big).$$
- *If $m = m^* < d$:*
$$\mathcal{L}(W(t)) \underset{t\to\infty}{=} O\big(e^{-4\mu t}\big).$$
- *If $m, d > m^*$:*
$$\mathcal{L}(W(t)) \underset{t\to\infty}{=} O\left(\frac{1}{t^2}\right).$$

This proposition is proven in Appendix C.2. The proof uses the fact that if $W(t)$ is solution of the flow in equation (4), then with probability one, $W(t)W(t)^T \to Z^*$ as $t \to \infty$. As explained in Section 3, the main challenge to understand the long time dynamics of the flow is to obtain the behaviour of the function $\psi(t)$ introduced in Proposition 3.1.

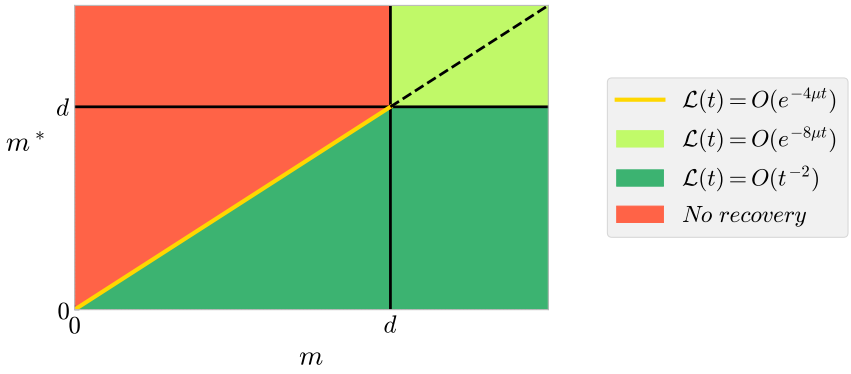

Figure 4: Diagram for the convergence rates of $\mathcal{L}(t) = \mathcal{L}(W(t))$ depending on the value of m, m^*. Valid under the initialization Assumption 4.1.

Figure 4 displays the diagram of convergence rates in the overparamterized case. While the regions $m, m^* \geq d$ and $m = m^* \leq d$ exhibit exponentially fast convergence, the third one ($m, d > m^*$) reveals a slower convergence. This discrepancy can be understood in terms of rank: in the two first case, we have that $\mathrm{rank}(W(t)W(t)^T) = \mathrm{rank}(Z^*)$ all along the flow. On the contrary, when $m, d > m^*$, then $W(t)W(t)^T$ still converges towards Z^*, which is now of lower rank. In this case the proof shows that the convergence is slowed down by the positive eigenvalues of $W(t)W(t)^T$ that go to zero as $t \to \infty$.

Whenever $m^* < d$, a high overparameterization is not ideal in terms of convergence rate. In this case, this is the smallest overparameterization (exactly $m = m^*$) which gives the most efficient convergence. Obviously, the number of neurons of the teacher m^* is not known in advance, and it is not clear how it can be inferred from the observations of the output of the predictors $u^*(x) = \mathrm{Tr}\big(W^* W^{*T} x x^T\big)$ (see equation (1)).

Finally, we believe that the bounds derived in Proposition A.1 are tight. As depicted in Figure 5, in the case $m^* \leq m \leq d$ (left panel), the function $t^2 \mathcal{L}(W(t))$ stays bounded with time. Likewise, for the case $m = m^* \leq d$ (right panel), the loss $\mathcal{L}(W(t))$ follows the line drawn by $e^{-4\mu t}$, where μ is the smallest non-zero eigenvalue of Z^* (in log-scale on the y-axis).

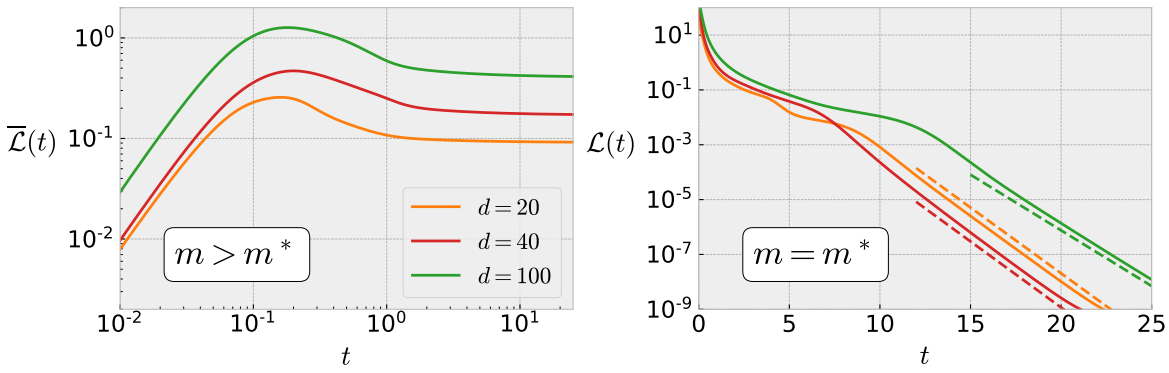

Figure 5: Evolution of the loss as a function of time. Left: $m = d/2$ and $m^* = d/4$, renormalized loss $\overline{\mathcal{L}}(t) = t^2 \mathcal{L}(W(t))$. Right: $m = m^* = d/2$, $\mathcal{L}(t) = \mathcal{L}(W(t))$ and dotted lines correspond to $\mathrm{cst} \times e^{-4\mu t}$ where μ is the smallest non-zero eigenvalue of Z^*. The solution $W(t)$ was simulated using gradient descent with stepsize $\eta = 10^{-2}$. Results averaged over 5 simulations.

A.2. Limit Distribution for Product of Projections

We now give more detailed results on the limit distribution for products of projections. In Section 5, we present a convergence result for the product of matrices $Y_d = U_d^{0T} U_d^* U_d^{*T} U_d^0 \in \mathbb{R}^{m_d \times m_d}$, where $U_d^{0T} U_d^0 = I_{m_d}$ and $U_d^{*T} U_d^* = I_{m_d^*}$. We also assume that U_d^0 and U_d^* are uniformly drawn under this constraint. This can be achieved by two different but equivalent ways. For $k \leq d$, the manifold $\mathcal{M}_{d,k} = \big\{U \in \mathbb{R}^{d \times k},\, U^T U = I_k\big\}$ can be endowed with a uniform measure, and U_d^0, U_d^* can be respectively drawn from this measure on $\mathcal{M}_{d,m_d}$ and $\mathcal{M}_{d,m_d^*}$. Another equivalent way of drawing uniformly on $\mathcal{M}_{d,k}$ is by drawing $V \in \mathbb{R}^{d \times k}$ with i.i.d. Gaussian coefficients $\mathcal{N}(0,1)$ and to let $U = V(V^T V)^{-1/2}$, so that, conditionally on the event $\{V^T V \text{ is invertible}\}$ (which has probability one), U is uniform on $\mathcal{M}_{d,k}$. Thus, in all the following, we make the assumption:

Assumption A.1. *The initial condition of the flow $W_d^0 \in \mathbb{R}^{d \times m_d}$ and the teacher matrix $W_d^* \in \mathbb{R}^{d \times m_d^*}$ verify:*

$$W_d^0 = \frac{1}{\sqrt{m_d}} U_d^0, \qquad W_d^* = \frac{1}{\sqrt{m_d^*}} U_d^*,$$

where the matrices U_d^0 and U_d^* are uniformly drawn on $\mathcal{M}_{d,m_d}$ and $\mathcal{M}_{d,m_d^*}$ respectively, with:

$$\frac{m_d}{d} \xrightarrow[d \to \infty]{} \alpha, \qquad \frac{m_d^*}{d} \xrightarrow[d \to \infty]{} \alpha^*.$$

This is the assumption under which the results of Section 5 are valid. As mentioned earlier, with probability one, the empirical spectral distribution of Y_d weakly converges towards some probability measure as $d \to \infty$. Kunisky (2023), Aubrun (2021), and Hiai and Petz (2005) have shown the convergence of the empirical spectral distribution of the matrix $X_d = U_d^0 Y_d U_d^{0T} \in \mathcal{S}_d^+(\mathbb{R})$. Informally, they obtain that for all $f : [0,1] \to \mathbb{R}$ continuous:

$$\frac{1}{d}\mathrm{Tr}\big(f(X_d)\big) \xrightarrow[d\to\infty]{} \int f(x)\mathrm{d}\nu(x),$$

where:

$$\mathrm{d}\nu(x) = \big(1 - \min(\alpha, \alpha^*)\big)\delta_0(x) + \max(\alpha + \alpha^* - 1, 0)\delta_1(x) + \frac{\sqrt{(r_+ - x)(x - r_-)}}{2\pi x(1-x)}\mathbb{1}_{[r_-, r_+]}(x)\mathrm{d}x.$$

More precisely, Kunisky (2023) obtained a convergence in moments and in probability under general assumption on the matrices distribution, and Hiai and Petz (2005) derived a large deviation result whenever the matrices U_d^*, U_d^0 are drawn uniformly on the Stiefel manifold.

To recover the measure defined in equation (10), note that X_d and Y_d have the same non-zero eigenvalues (this is true only when $U_d^{0T} U_d^0 = I_{m_d}$). Thus, we have for $f : [0,1] \to \mathbb{R}$ continuous:

$$\begin{aligned}\frac{1}{m_d}\mathrm{Tr}\big(f(Y_d)\big) &= \frac{d}{m_d}\frac{1}{d}\mathrm{Tr}\big(f(X_d)\big) - \frac{d - m_d}{m_d}f(0) \\ &\xrightarrow[d\to\infty]{} \frac{1}{\alpha}\int f(x)\mathrm{d}\nu(x) - \frac{1-\alpha}{\alpha}f(0).\end{aligned}$$

Thus, the empirical spectral distribution of Y_d converges as $d \to \infty$ towards a probability measure μ which satisfies $\alpha\mu = \nu - (1-\alpha)\delta_0$, which indeed corresponds to equation (10).

We now state a lemma which will be used in the proof of Propositions 5.1 and 5.2. It generalizes the previous convergence of the empirical spectral distribution of Y_d for a compact family of functions.

Lemma A.1. *Suppose that Assumption A.1 is verified and denote $Y_d = U_d^{0T} U_d^* U_d^{*T} U_d^0 \in \mathbb{R}^{m_d \times m_d}$. Let $\mathcal{F} \subset \mathcal{C}([0,1], \mathbb{R})$ be a compact set for the norm $\|\cdot\|_\infty$. Then, with probability one:*

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{m_d}\mathrm{Tr}\big(f(Y_d)\big) - \int f(x)\mathrm{d}\mu(x) \right| \xrightarrow[d\to\infty]{} 0.$$

Proof. We let $\eta > 0$. Since $\mathcal{F}$ is compact, we can find $f_1, \ldots, f_n \in \mathcal{F}$ such that any element of $\mathcal{F}$ is at distance $\leq \eta/4$ of one of the f_i's. Thus:

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{m_d}\mathrm{Tr}\big(f(Y_d)\big) - \int f(x)\mathrm{d}\mu(x) \right| \leq \frac{\eta}{2} + \sup_{1\leq i\leq n} \left| \frac{1}{m_d}\mathrm{Tr}\big(f_i(Y_d)\big) - \int f_i(x)\mathrm{d}\mu(x) \right|,$$

and:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{m_d}\mathrm{Tr}\big(f(Y_d)\big) - \int f(x)\mathrm{d}\mu(x) \right| \geq \eta \right) \leq \mathbb{P}\left(\sup_{1\leq i\leq n} \left| \frac{1}{m_d}\mathrm{Tr}\big(f_i(Y_d)\big) - \int f_i(x)\mathrm{d}\mu(x) \right| \geq \frac{\eta}{2} \right). \tag{13}$$

We now use the large deviation result obtained by Hiai and Petz (2005). We denote μ_d the empirical spectral distribution of Y_d and for $\delta > 0$, we define:

$$E_\eta = \left\{ \nu \in \mathcal{M} \;\middle|\; \sup_{1\leq i\leq n} \left| \int f_i(x)\mathrm{d}\nu(x) - \int f_i(x)\mathrm{d}\mu(x) \right| \geq \frac{\eta}{2} \right\},$$

where μ is defined in equation (10) and $\mathcal{M}$ denotes the space of probability measures over $[0,1]$. First, since the supremum is over a finite number of functions, E_η is closed for the weak topology on $\mathcal{M}$.

Now Hiai and Petz (2005) state that:

$$\limsup_{d\to\infty} \frac{1}{d^2} \log \mathbb{P}(\mu_d \in E_\eta) \leq - \inf_{\nu \in E_\eta} I(\nu) \equiv -\beta_\eta,$$

where I is positive, lower semi-continuous (with respect to the weak topology on $\mathcal{M}$), such that $I(\mu) = 0$ and $I > 0$ elsewhere. Remains to show that $\beta_\eta > 0$. By contradiction, suppose that there is $(\nu_p)_{p\in\mathbb{N}}$ a sequence of E_η such that $I(\nu_p) \xrightarrow[p\to\infty]{} 0$. Since $\mathcal{M}$ is compact, one can extract a converging subsequence: $\nu_{\phi(p)} \xrightarrow[p\to\infty]{} \nu$. Thus, by lower semi-continuity of I:

$$I(\nu) \leq \liminf_{p\to\infty} I(\nu_{\phi(p)}) = 0,$$

and $\nu = \mu$. Since E_η is closed for the weak topology, this implies that $\mu \in E_\eta$ which leads to a contradiction. Thus for d large enough:

$$\mathbb{P}\left(\sup_{1\leq i\leq n} \left| \frac{1}{m_d} \mathrm{Tr}\big(f_i(Y_d)\big) - \int f_i(x) \mathrm{d}\mu(x) \right| \geq \frac{\eta}{2} \right) \leq \exp\left(-\frac{\beta_\eta d^2}{2} \right),$$

and using equation (13), we get that, for all $\eta > 0$:

$$\sum_{d\geq 0} \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{m_d} \mathrm{Tr}\big(f(Y_d)\big) - \int f(x) \mathrm{d}\mu(x) \right| \geq \eta \right) < \infty,$$

and the conclusion using Borel-Cantelli lemma. □

A.3. Overlap

We study the overlap function in Section 5.3. In a general Euclidean space E, the overlap between two vectors $x, y \in E$ is given by:

$$\chi(x, y) = \frac{\big|\langle x, y\rangle\big|}{\|x\|\|y\|},$$

where $\|\cdot\|$ is the norm associated with the inner product on E. The overlap is a powerful scalar quantity measuring the alignment between two vectors. Indeed, Cauchy-Schwartz inequality ensures that $\chi(x, y) \leq 1$, and $\chi(x, y) = 1$ if and only if x and y are aligned.

It is well known that if $E = \mathbb{R}^d$ and x, y are independent and uniformly drawn on the sphere $\mathbb{S}^{d-1}$, then:

$$\chi(x, y) \underset{d\to\infty}{=} O\left(\frac{1}{\sqrt{d}}\right).$$

In this case, what we called the purely random overlap goes to zero in the high-dimensional limit. Throughout Section 5, we studied the overlap between two matrices in $\mathcal{S}_d^+(\mathbb{R})$:

$$\chi^{\mathrm{PSD}}(X, Y) = \frac{\mathrm{Tr}(XY)}{\big\|X\big\|_F \big\|Y\big\|_F}.$$

For $X, Y \in \mathcal{S}_d^+(\mathbb{R})$, we always have that $\mathrm{Tr}(XY) \geq 0$. Note that we recover the $\mathbb{R}^d$ overlap with $\chi^{\mathrm{PSD}}(xx^T, yy^T) = \chi(x, y)^2$ for $x, y \in \mathbb{R}^d$. The first challenge is to determine the purely random overlap mentioned in the beginning of Section 5.3. In our special case, the goal is to study the overlap between the teacher and student along the flow: therefore, our purely random overlap will be defined as the one between the teacher and student at initialization. Those are independent random projection matrices of rank m_d^* and m_d respectively. We now show that this overlap admits a limit as $d \to \infty$, whenever m_d, m_d^* grow linearly with the dimension.

Lemma A.2. *The overlap between the teacher matrix and the student at initialization admits the limit, with*

probability one:

$$\chi^{\mathrm{PSD}}(Z_d^0, Z_d^*) \xrightarrow[d\to\infty]{} \sqrt{\alpha\alpha^*}.$$

Proof. Due to the definition of Z_d^* and Z_d^0, we have:

$$\chi^{\mathrm{PSD}}(Z_d^0, Z_d^*) = \frac{\mathrm{Tr}\big(U_d^0 U_d^{*T} U_d^{*T} U_d^0\big)}{\big\|U_d^0 U_d^{0T}\big\|_F \big\|U_d^* U_d^{*T}\big\|_F}.$$

For the denominator, we obtain:

$$\big\|U_d^0 U_d^{0T}\big\|_F = \sqrt{m_d}, \qquad\qquad \big\|U_d^* U_d^{*T}\big\|_F = \sqrt{m_d^*}.$$

Due to the rotational invariance of the distribution of $U_d^* U_d^{*T}$, one can suppose without loss of generality that $U_d^* U_d^{*T}$ is the orthogonal projection onto the m_d^* first vectors of the standard basis in $\mathbb{R}^d$, denoted Π_d. Thus, one can rewrite:

$$\chi^{\mathrm{PSD}}(Z_d^0, Z_d^*) = \frac{1}{\sqrt{m_d m_d^*}} \sum_{k=1}^{m_d} \big\|\Pi_d(v_k)\big\|^2,$$

with $(v_1, \dots, v_{m_d})$ orthonormal such that $U_d^0 U_d^{0T} = \sum_{k=1}^{m_d} v_k v_k^T$. Moreover, each v_j is uniform on the sphere. Thus, for $\epsilon > 0$, applying an union bound:

$$\mathbb{P}\left(\left|\chi^{\mathrm{PSD}}(Z_d^0, Z_d^*) - \frac{\sqrt{m_d m_d^*}}{d}\right| \geq \epsilon\right) \leq m_d\, \mathbb{P}\left(\left|\big\|\Pi_d(v_k)\big\|^2 - \frac{m_d^*}{d}\right| \geq \epsilon\sqrt{\frac{m_d^*}{m_d}}\right).$$

Following Dasgupta and Gupta (2003), we have, if v is uniform on the sphere and for $\eta \in [0,1]$:

$$\mathbb{P}\left(\left|\big\|\Pi_d(v)\big\|^2 - \frac{m_d^*}{d}\right| \geq \frac{m_d^*}{d}\eta\right) \leq 2\exp\left(-\frac{m_d^*\eta^2}{4}\right),$$

which leads to, with probability one:

$$\left|\chi^{\mathrm{PSD}}(Z_d^0, Z_d^*) - \frac{\sqrt{m_d m_d^*}}{d}\right| \xrightarrow[d\to\infty]{} 0,$$

and the result. □

Note that the behaviour of this overlap is very different from the $\mathbb{R}^d$ case. The overlap between two independent vectors in $\mathbb{R}^d$ vanishes as $d \to \infty$. Here, due to the fact we consider a number of vectors growing with the dimension, we instead obtain a positive limit.

Finally, we prove a claim done at the end of Section 5.3: the overlap reached by the gradient flow at timescale $t \gg d$ (see Proposition 5.2) is the best we can get whenever the teachers are orthonormal (for a given number of students).

Proposition A.2. *Suppose that $Z^* \in \mathcal{S}_d^+(\mathbb{R})$ is an orthogonal projection matrix of rank $m^* \leq d$. Then, for $m \leq d$:*

$$\sup_{\substack{Z\in\mathcal{S}_d^+(\mathbb{R}) \\ \mathrm{rank}(Z)\leq m}} \frac{\mathrm{Tr}(ZZ^*)}{\|Z\|_F \|Z^*\|_F} = \min\left(\sqrt{\frac{m}{m^*}}, 1\right).$$

Proof. If $m \geq m^*$, taking $Z = Z^*$ leads to the result since the overlap is always smaller than one. Suppose that $m < m^*$ and take $Z \in \mathcal{S}_d^+(\mathbb{R})$ of rank $\leq m$. We write:

$$Z = \sum_{k=1}^{m} \lambda_k v_k v_k^T \qquad\qquad Z^* = \sum_{k=1}^{m^*} v_k^*(v_k^*)^T,$$

so that:

$$\mathrm{Tr}(ZZ^*) = \sum_{j=1}^{m} \lambda_j \sum_{k=1}^{m^*} \left(v_j^T v_k^*\right)^2 .$$

The sum over k corresponds to the norm of the projection of v_j onto $\mathrm{Vect}(v_1^*, \dots, v_{m^*}^*)$, thus it is smaller than the norm of v_j, equal to one. Thus, since $\left\|Z^*\right\|_F = \sqrt{m^*}$

$$\frac{\left|\mathrm{Tr}(ZZ^*)\right|}{\left\|Z\right\|_F \left\|Z^*\right\|_F} \leq \frac{1}{\sqrt{m^*}} \frac{\sum_{j=1}^{m} \lambda_j}{\left(\sum_{j=1}^{m} \lambda_j^2\right)^{1/2}},$$

and the upper bound using Cauchy-Schwarz inequality. The case of equality is reached if and only if Z is of the form $Z = \lambda \sum_{k=1}^{m} v_k v_k^T$, with $(v_1, \dots, v_m)$ an orthonormal family of $\mathrm{Vect}(v_1^*, \dots, v_{m^*}^*)$. □

B. Useful Lemmas

In this section we mention several lemmas that will be used throughout the proofs of the results in Appendix C.

Lemma B.1. *Let $A \in \mathcal{S}_d^+(\mathbb{R})$. Then:*

$$\frac{1}{d}\left[\mathrm{Tr}(A)\right]^2 \leq \mathrm{Tr}(A^2) \leq \left[\mathrm{Tr}(A)\right]^2.$$

Moreover, if $B \in \mathcal{S}_d(\mathbb{R})$:

$$\lambda_{min}(B)\mathrm{Tr}(A) \leq \mathrm{Tr}(AB) \leq \lambda_{max}(B)\mathrm{Tr}(A).$$

Lemma B.2 (Courant-Fischer min-max formula)**.** *Let $A \in \mathcal{S}_d(\mathbb{R})$ and denote $\lambda_1 \geq \cdots \geq \lambda_d$ its eigenvalues. Then:*

$$\lambda_j = \max_{\substack{V \subset \mathbb{R}^d \\ \dim V = j}} \min_{\substack{x \in V \\ \|x\|=1}} x^T A x \;=\; \min_{\substack{V \subset \mathbb{R}^d \\ \dim V = d-j+1}} \max_{\substack{x \in V \\ \|x\|=1}} x^T A x.$$

Lemma B.3. *Let $\lambda > 0$ and $v : \mathbb{R}^+ \to \mathbb{R}$ continuously differentiable such that $\dot{v}(t) \xrightarrow[t\to\infty]{} 0$. Then :*

$$\int_0^t e^{\lambda s} e^{v(s)} \mathrm{d}s \underset{t\to\infty}{\sim} \frac{1}{\lambda} e^{\lambda t} e^{v(t)}.$$

Moreover, for all $\mu > 0$:

$$e^{-\mu t} \int_0^t e^{v(s)} \mathrm{d}s \xrightarrow[t\to\infty]{} 0.$$

Proof. The condition on v implies that the first integral diverges as $t \to \infty$. Integrating by parts:

$$\int_0^t e^{\lambda s} e^{v(s)} \mathrm{d}s = \frac{1}{\lambda}\left[e^{\lambda s} e^{v(s)}\right]_0^t - \frac{1}{\lambda}\int_0^t \dot{v}(s) e^{\lambda s} e^{v(s)} \mathrm{d}s.$$

Due to the assumption on v, the second integral is negligible with respect to the first, which gives the first result. For the second claim, let $\epsilon \in]0, \mu[$ and t_0 such that $v(t) \leq \epsilon t$ for $t \geq t_0$. Then, splitting the integral:

$$e^{-\mu t} \int_0^t e^{v(s)} \mathrm{d}s \leq e^{-\mu t} \int_0^{t_0} e^{v(s)} \mathrm{d}s + \frac{1}{\epsilon} e^{-(\mu-\epsilon)(t-t_0)} \xrightarrow[t\to\infty]{} 0.$$

□

C. Proofs of the Main Results

C.1. Proof of Proposition 4.1

We now prove the convergence result stated in Proposition 4.1. As mentioned earlier, we make use of the stable manifold theorem (see Smale, 2011): under the assumption that the initialization of the flow W^0 is drawn from a distribution which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$, and since the loss function $\mathcal{L}$ defined in equation (3) is analytic in the coefficients of W, then with probability one (with respect to the initialization) the solution of the gradient flow (4) converges towards a local minimizer of $\mathcal{L}$, i.e., a point $W \in \mathbb{R}^{d\times m}$ satisfying :

$$\nabla\mathcal{L}(W) = 0 \qquad \text{and} \qquad \mathrm{Tr}\big(\mathrm{d}^2 L_W(K)K^T\big) \geq 0,$$

for all $K \in \mathbb{R}^{d\times m}$. Proposition 4.1 is proven as follows:

- In Lemma C.1, we determine the critical points of the loss, i.e., the $W \in \mathbb{R}^{d\times m}$ such that $\nabla\mathcal{L}(W) = 0$.
- In Lemma C.2, we make use of the structure of the loss and show that a rank deficient local minimizer necessarily leads to an optimal predictor (i.e., $WW^T = Z^*$).
- Finally, we determine the local minimizer of the loss.

Lemma C.1. *Let $W \in \mathbb{R}^{d\times m}$ be a critical point of $\mathcal{L}$. Then there exists $I \subset [\![1, d]\!]$ of size $\leq \min(m, d)$ such that:*

$$WW^T = \sum_{k\in I}(\mu_k + \tau)v_k^*(v_k^*)^T,$$

with:

$$\tau = \frac{1}{2 + |I|}\sum_{\substack{1\leq k\leq m^*\\ k\notin I}} \mu_k.$$

Proof. Let $W \in \mathbb{R}^{d\times m}$ such that $\nabla\mathcal{L}(W) = 0$, that is:

$$2\big(Z^* - WW^T\big)W + \mathrm{Tr}\big(Z^* - WW^T\big)W = 0.$$

Computing $\nabla\mathcal{L}(W)W^T - W\nabla\mathcal{L}(W)^T = 0$, we obtain that Z^* and WW^T commute, therefore we can write a singular value decomposition of the form:

$$W = \sum_{k=1}^{d}\sqrt{\lambda_k}v_k^*w_k^T,$$

where $\lambda_1, \ldots, \lambda_d \geq 0$ and if $m \leq d$, at most m of them are non-zero. Setting $\tau = \frac{1}{2}\mathrm{Tr}\big(Z^* - WW^T\big)$, the constraint $\nabla\mathcal{L}(W) = 0$ writes:

$$\sum_{k=1}^{d}\sqrt{\lambda_k}\big(\lambda_k - \mu_k - \tau\big)v_k^*w_k^T = 0.$$

Thus, if $\lambda_k > 0$, necessarily $\lambda_k = \mu_k + \tau$. Defining $I = \{k \in [\![1, d]\!], \lambda_k > 0\}$ (which is of size $\leq \min(m, d)$) leads to the first claim. For the expression of τ, simply remark that:

$$\tau = \frac{1}{2}\mathrm{Tr}\big(Z^* - WW^T\big) = \frac{1}{2}\left(\mathrm{Tr}(Z^*) - \sum_{k\in I}(\mu_k + \tau)\right) = \frac{1}{2}\sum_{\substack{1\leq k\leq m^*\\ k\notin I}}\mu_k - \frac{1}{2}|I|\tau,$$

which gives the desired result. □

We now state a more general lemma which bears on the functions of the form $L(W) = F(WW^T)$ where F is convex. This naturally includes our setup since for the loss defined in equation (3), the function F is quadratic with a positive-semidefinite Hessian, hence convex.

Lemma C.2. *Let $L : \mathbb{R}^{d\times m} \to \mathbb{R}$ such that $L(W) = F(WW^T)$ for all $W \in \mathbb{R}^{d\times m}$ with $F : \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}$ convex and twice continuously differentiable. Suppose that $W \in \mathbb{R}^{d\times m}$ is a local minimizer of L. Then, if $\mathrm{rank}(W) < m$, WW^T is a global minimizer of F over the space $\mathcal{S}_d^+(\mathbb{R})$.*

Proof. The gradient and Hessian of L can be deduced from the ones of F:

$$\nabla L(W) = 2\nabla F(WW^T)W$$
$$\mathrm{Tr}\big(\mathrm{d}^2 L_W(K)K^T\big) = 2\mathrm{Tr}\big(\nabla F(WW^T)KK^T\big) + \mathrm{Tr}\left(\mathrm{d}^2 F_{WW^T}(WK^T + KW^T)\big(WK^T + KW^T\big)\right).$$

If $W \in \mathbb{R}^{d\times m}$ is a local minimizer of L, then $\nabla F(WW^T)W = 0$ and $\mathrm{Tr}\big(\mathrm{d}^2 L_W(K)K^T\big) \geq 0$ for all $K \in \mathbb{R}^{d\times m}$. If $\mathrm{rank}(W) < m$, then there is a non-zero $v \in \mathbb{R}^m$ such that $Wv = 0$. Let $u \in \mathbb{R}^d$ and evaluate the second equation at $K = uv^T$, so that WK^T and KW^T are zero. Thus, we get that $\mathrm{Tr}\big(\nabla F(WW^T)uu^T\big) \geq 0$ so that $\nabla F(WW^T) \in \mathcal{S}_d^+(\mathbb{R})$ (it is symmetric since F is a function on $\mathcal{S}_d(\mathbb{R})$). Since F is convex, we have, for any $S \in \mathcal{S}_d^+(\mathbb{R})$:

$$\begin{aligned} F(S) &\geq F(WW^T) + \mathrm{Tr}\big(\nabla F(WW^T)(S - WW^T)\big) \\ &= F(WW^T) + \underbrace{\mathrm{Tr}\big(\nabla F(WW^T)S\big)}_{\geq 0}, \end{aligned}$$

where we used that $\nabla F(WW^T)W = 0$. Thus WW^T is a global minimizer of F over $\mathcal{S}_d^+(\mathbb{R})$. □

We can apply this result to our case, with $F(Z) = \frac{1}{2}\big\|Z - Z^*\big\|_F^2 + \frac{1}{4}\mathrm{Tr}\big(Z - Z^*\big)^2$. Note that the only global minimizer of F on $\mathcal{S}_d^+(\mathbb{R})$ is $Z = Z^*$.

We are now ready to prove Proposition 4.1. Let $W \in \mathbb{R}^{d\times m}$ be a local minimizer of the loss $\mathcal{L}$. Already if $m > d$, then $\mathrm{rank}(W) < m$ and necessarily $WW^T = Z^*$ by the previous result, so we can assume that $m \leq d$. Likewise; if $m \geq m^*$ and $\mathrm{rank}(W) < m$, we also get that $WW^T = Z^*$. Finally, for $m < m^*$, such an equality is not possible since $\mathrm{rank}(Z^*) = m^*$ so we necessarily have $\mathrm{rank}(W) = m$.

We now write the local minimizer condition $\mathrm{Tr}\big(\mathrm{d}^2\mathcal{L}_W(K)K^T\big) \geq 0$:

$$\mathrm{Tr}\big((WW^T - Z^*)KK^T\big) - \tau\mathrm{Tr}(KK^T) + \mathrm{Tr}(KW^TWK^T) + \mathrm{Tr}(KW^TKW^T) + \mathrm{Tr}(WK^T)^2 \geq 0,$$

where τ is defined in Lemma C.1. Reusing the notations of this lemma, we take $j \in I$ (i.e $\lambda_j = \mu_j + \tau > 0$) and $k \in [\![1, d]\!]$. Evaluating the previous equation with $K = v_k^* w_j^T$, we get:

$$\lambda_k + \lambda_j + 2\lambda_j\delta_{j,k} \geq \mu_k + \tau.$$

If $j \neq k$, then replacing $\lambda_j = \mu_j + \tau$, we have that $\lambda_k \geq \mu_k - \mu_j$. By the assumption on the eigenvalues of Z^* ($\mu_1 > \cdots > \mu_{m^*} > 0$), taking $k < j$ (if possible), we obtain that $\lambda_k > 0$ thus $k \in I$ by definition of I. Therefore, whenever $j \in I$, we have that $[\![1, j]\!] \subset I$. Since $|I| = \mathrm{rank}(W) = m$ by assumption, necessarily $I = [\![1, m]\!]$. For $m \leq m^*$, this implies the desired by the expression of τ (obviously $\tau = 0$ for $m = m^*$). For $m > m^*$, we also obtain that $\tau = 0$ and we cannot have $I = [\![1, m]\!]$ since Z^* have only m^* non-zero eigenvalues. In this case, any local minimizer of the loss must satisfy $WW^T = Z^*$.

C.2. Proof of Proposition A.1

Following the previous result, we now determine the convergence rates associated with the convergence of the flow in the case $m \geq m^*$. As explained in Section 3, the understanding of the function ψ defined in Proposition 3.1 is essential to the determination of the dynamics. Under the assumption that $W(t)W(t)^T \xrightarrow[t\to\infty]{} Z^*$, which happens with probability one when the flow is correctly initialized, we have:

$$\psi(t) = \int_0^t \mathrm{Tr}\big(W(s)W(s)^T\big)\mathrm{d}s \underset{t\to\infty}{\sim} \mathrm{Tr}(Z^*)t. \tag{14}$$

The first challenge will be to gather more information about ψ, using equation (6):

$$\psi(t) = \frac{1}{4}\mathrm{Tr}\log\left(I_m + 4W^{0T}\int_0^t e^{-2\psi(s)}e^{2s\mathrm{Tr}(Z^*)}e^{4Z^*s}\mathrm{d}s\, W^0\right). \tag{15}$$

We will denote $E^*(t) = \int_0^t e^{-2\psi(s)}e^{2s\mathrm{Tr}(Z^*)}e^{4Z^*s}\mathrm{d}s \in \mathbb{R}^{d\times d}$. The proof is done as follows:

- In Lemma C.3, we give an asymptotic behaviour of the eigenvalues of the matrix $W^{0T}E^*(t)W^0 \in \mathbb{R}^{m\times m}$ as $t \to \infty$.
- In Lemma C.4, we obtain an asymptotic development of ψ up to the precision $o(1)$.
- We finally prove Proposition A.1 by using the differential equation on $Z(t) = W(t)W(t)^T$:

$$\dot{Z} = Z\mathrm{Tr}(Z^* - Z)Z + 2Z^*Z + 2ZZ^* - 4Z^2.$$

Our results are verified under Assumption 4.1. As mentioned before, for $m \geq m^*$, we have $W(t)W(t)^T \longrightarrow Z^*$ as $t \to \infty$ with probability one. Moreover, almost surely, any subfamily of $(W^{0T}v_1^*, \ldots, W^{0T}v_d^*)$ with size $\leq \min(m,d)$ is linearly independent in $\mathbb{R}^m$ (we say that this family is in general position, and this occurs almost surely as soon as assumption 4.1 is verified and W^0 is drawn independently from Z^*). We restrict ourselves to this event from now on, to avoid sets of zero probability.

In order to determine the behaviour of ψ, it is important to understand how the eigenvalues of $W^{0T}E^*(t)W^0$ evolve as $t \to \infty$. This is done in the following lemma: the core idea is that the eigenvalues of E^* (directly related to those of Z^*) are well separated at long timescales, even if we have little information on ψ. Indeed, the fact that $\psi(t) \underset{t\to\infty}{\sim} \mathrm{Tr}(Z^*)t$ is enough to prove the result.

In Lemma C.3 and Lemma C.4, we will assume that $m^* \leq d$. Indeed, for the case $m^* > d$, the proof of convergence rates will be straightforward.

Lemma C.3. *Denote $r = \min(m,d)$ and $\lambda_1(t), \ldots, \lambda_r(t) > 0$ the non-zero eigenvalues of $W^{0T}E^*(t)W^0$. Then, for $j \in [\![1, r]\!]$:*

$$\lambda_j(t) \underset{t\to\infty}{=} \Theta\left(\int_0^t e^{-2\psi(s)}e^{2s\mathrm{Tr}(Z^*)}e^{4\mu_j s}\mathrm{d}s\right).$$

The notation Θ means that $\lambda_j(t)$ is bounded from below and above by constants times the integral. This lemma is proven in Appendix D.1. The goal is now to use this result along with equation (15) to obtain a more precise idea of how $\psi(t)$ evolve as $t \to \infty$. This is done in the following lemma.

Lemma C.4. *For $m \geq m^*$:*

$$\psi(t) \underset{t\to\infty}{=} \mathrm{Tr}(Z^*)t + \frac{1}{2}\frac{\min(m,d) - m^*}{\min(m,d)+2}\log(t) + O(1).$$

Proof. Set $\xi(t) = \psi(t) - \mathrm{Tr}(Z^*)t$ which is $o(t)$ by equation (14). From the previous lemma, let C_j, D_j such that:

$$C_j \leq \lambda_j(t)\left(\int_0^t e^{-2\psi(s)}e^{2s\mathrm{Tr}(Z^*)}e^{4\mu_j s}\mathrm{d}s\right)^{-1} \leq D_j.$$

In the following, we only use the upper bound, the lower bound will be done similarly. By equation (15) and using the notations of the previous lemma, as well as $r = \min(m,d)$:

$$\begin{aligned}\psi(t) &= \frac{1}{4}\sum_{j=1}^{r}\log\left(1 + 4\lambda_j(t)\right)\\ &\leq \frac{1}{4}\sum_{j=1}^{m^*}\log\left(1 + 4D_j\int_0^t e^{-2\xi(s)}e^{4\mu_j s}\mathrm{d}s\right) + \frac{1}{4}\sum_{j=m^*+1}^{r}\log\left(1 + 4D_j\int_0^t e^{-2\xi(s)}\mathrm{d}s\right).\end{aligned}$$

From Lemma B.3, since $\dot{\xi}(t)$ goes to zero as $t \to \infty$, we can obtain the behaviour of the first term:

$$\frac{1}{4}\sum_{j=1}^{m^*} \log\left(1 + 4D_j \int_0^t e^{-2\xi(s)} e^{4\mu_j s} \mathrm{d}s\right) \underset{t\to\infty}{=} \mathrm{Tr}(Z^*)t - \frac{m^*}{2}\xi(t) + \frac{1}{4}\sum_{j=1}^{m^*} \log\left(\frac{D_j}{\mu_j}\right) + o(1).$$

For the second we need to understand how $\int_0^t e^{-2\xi(s)}\mathrm{d}s$ behave. As $t \to \infty$, this integral either converges or goes to infinity. Suppose by contradiction that it converges towards some finite value. Using the two previous equations, along with the lower bound obtained from the previous lemma, this implies that $\xi(t)$ stays bounded, which leads to a contradiction since $\int_0^t e^{-2\xi(s)}\mathrm{d}s$ should diverge. Thus, we obtain:

$$\left(1 + \frac{m^*}{2}\right)\xi(t) \underset{t\to\infty}{\leq} \frac{r-m^*}{4}\log\left(\int_0^t e^{-2\xi(s)}\mathrm{d}s\right) + \frac{1}{4}\sum_{j=1}^{m^*}\log\left(\frac{D_j}{\mu_j}\right) + \frac{1}{4}\sum_{j=m^*+1}^{r}\log(4D_j) + o(1).$$

Thus, using the lower bound of the previous lemma, we obtain $h_1(t), h_2(t)$ two bounded functions such that:

$$h_1(t) \leq \xi(t) - \frac{r-m^*}{2(m^*+2)}\log\left(\int_0^t e^{-2\xi(s)}\mathrm{d}s\right) \leq h_2(t).$$

Setting $a(t) = \int_0^t e^{-2\xi(s)\mathrm{d}s}$, we obtain:

$$e^{-2h_2(t)} \leq \dot{a}(t)a(t)^\beta \leq e^{-2h_1(t)},$$

with $\beta = \dfrac{r-m^*}{m^*+2}$. Integrating, and using that $\xi(t) = -\dfrac{1}{2}\log\dot{a}(t)$:

$$\zeta + h_1(t) + \frac{\beta}{2(\beta+1)}\log\left(\int_0^t e^{-2h_2(s)}\mathrm{d}s\right) \leq \xi(t) \leq \zeta + h_2(t) + \frac{\beta}{2(\beta+1)}\log\left(\int_0^t e^{-2h_1(s)}\mathrm{d}s\right),$$

where ζ is a constant depending only on β. We now use that h_1, h_2 are $O(1)$. There exists constants K_1, K_2 such that:

$$K_1 \leq \xi(t) - \frac{\beta}{2(\beta+1)}\log(t) \leq K_2,$$

which gives the result by the definition of β and $r = \min(m, d)$. □

Now that the behaviour of ψ is known, we are ready to determine the convergence rates in the case $m \geq m^*$. We showed that we can assume Z^* to be diagonal without loss of generality. Let $Z^* = Q^T D^* Q$ with $Q \in O_d(\mathbb{R})$. If $W(t)$ is solution of the flow of equation (4) (associated with Z^*), then $V(t) = QW(t)$ is solution of the same equation but with teacher D^*. Moreover, since $VV^T - D^* = Q(WW^T - Z^*)Q^T$, this implies that the loss of V (with teacher D^*) is equal to the loss of W (with teacher Z^*). Thus, as long as the convergence properties we derive are invariant under conjugation for Z^* (for instance if they only depend on its eigenvalues, which is the case in Proposition A.1), we may assume that:

$$Z^* = \begin{pmatrix} Z_0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{16}$$

where $Z_0 = \mathrm{diag}(\mu_1, \dots, \mu_{m^*}) \in \mathbb{R}^{m^*\times m^*}$. We split the proof into two parts: first, we study the case $m, m^* \geq d$. In the second part, we jointly cover the two other cases.

C.2.1. Highly overparameterized case

We start with the first case, i.e $m, m^* \geq d$. If $W(t)$ if solution of the flow then, $\dot{W}(t) = -M(t)W(t)$ with $M(t) = 2\big(W(t)W(t)^T - Z^*\big) + \mathrm{Tr}\big(W(t)W(t)^T - Z^*\big)I_d$. Thus:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(W(t)) = -\big\|\dot{W}(t)\big\|^2 = -\mathrm{Tr}\big(M(t)^2 W(t)W(t)^T\big) \leq -\lambda_{min}\big(W(t)W(t)^T\big)\mathrm{Tr}\big(M(t)^2\big).$$

We now use the inequality $\mathrm{Tr}\big(M(t)^2\big) \geq 8\,\mathcal{L}(W(t))$, and the fact that $\lambda_{min}\big(W(s)W(s)^T\big) \xrightarrow[t\to\infty]{} \mu$. Therefore:

$$\mathcal{L}(W(t)) \leq \mathcal{L}(W^0)\exp\left(-8\int_0^t \lambda_{min}\big(W(s)W(s)^T\big)\mathrm{d}s\right) = \mathcal{L}(W^0)\exp\big(-8\mu t + o(t)\big).$$

We now show that the $o(t)$ is in fact a $O(1)$:

$$\left|\int_0^t \lambda_{min}\big(W(s)W(s)^T\big)\mathrm{d}s - \mu t\right| \leq \int_0^t \big\|W(s)W(s)^T - Z^*\big\|\mathrm{d}s \leq \sqrt{2}\int_0^t \sqrt{\mathcal{L}(W(s))}\mathrm{d}s,$$

which remains bounded as $t \to \infty$. This proves the first claim of Proposition A.1.

C.2.2. General case

In the following, we suppose that $m^* < d$. To derive the convergence rates in the case $m \geq m^*$, the first step uses the implicit solution obtained in Proposition 3.1:

$$Z(t) = e^{-2\xi(t)}e^{2Z^*t}W^0\underbrace{\left(I_m + 4W^{0T}\int_0^t e^{-2\xi(s)}e^{4Z^*s}\mathrm{d}s\, W^0\right)^{-1}}_{\equiv N(t)^{-1}}W^{0T}e^{2Z^*t},$$

with $\xi(t) = \psi(t) - \mathrm{Tr}(Z^*)t$ whose behaviour is known as $t \to \infty$. We now decompose $W^0 = \begin{pmatrix} U_0 \\ V_0 \end{pmatrix}$ where $U_0 \in \mathbb{R}^{m^*\times m}$ and $V_0 \in \mathbb{R}^{(d-m^*)\times m}$. Note that under our assumption on the initialization, we have with probability one $\mathrm{rank}(W_0) = \min(m,d)$ and $\mathrm{rank}(U_0) = m^*$ (since $m \geq m^*$). In order to avoid sets of probability one, we restrict ourselves to this event. Then, from the previous equation:

$$Z(t) = e^{-2\xi(t)}\begin{pmatrix} e^{2Z_0t}U_0N(t)^{-1}U_0^Te^{2Z_0t} & e^{2Z_0t}U_0N(t)^{-1}V_0^T \\ V_0N(t)^{-1}U_0^Te^{2Z_0t} & V_0N(t)^{-1}V_0^T \end{pmatrix} \equiv \begin{pmatrix} A(t) & B(t) \\ B(t)^T & C(t) \end{pmatrix}, \tag{17}$$

with $A(t) \in \mathcal{S}^+_{m^*}(\mathbb{R}), B(t) \in \mathbb{R}^{m^*\times(d-m^*)}$ and $C(t) \in \mathcal{S}^+_{d-m^*}(\mathbb{R})$. To start, we are going to bound the bottom right term. From Lemma B.1:

$$\mathrm{Tr}\big(C(t)\big) = e^{-2\xi(t)}\mathrm{Tr}\big(V_0^TN(t)^{-1}V_0\big) \leq e^{-2\xi(t)}\mathrm{Tr}(V_0V_0^T)\lambda_{min}(N(t))^{-1}.$$

Since $N(t) = I_m + 4W^{0T}E^*(t)W^0$, the behaviour of the smallest eigenvalue of $N(t)$ can be deduced from Lemma C.3 along with the behaviour of ξ. We get that:

$$\mathrm{Tr}\big(C(t)\big) \underset{t\to\infty}{=} \begin{cases} O\big(e^{-4\mu t}\big) & \text{for } m = m^* \\ O\big(t^{-1}\big) & \text{for } m > m^* \end{cases}, \tag{18}$$

where μ is the smallest non-zero eigenvalue of Z^*. Now, in order to obtain the behaviour of $A(t)$ and $B(t)$, we use the differential equation on $Z(t)$:

$$\dot{Z}(t) = -2\dot{\xi}(t)Z(t) + 2Z(t)(Z^* - Z(t)) + 2(Z^* - Z(t))Z(t).$$

Due to the form of Z^* in equation (16), this induces the evolution equations for $A(t)$ and $B(t)$:

$$\begin{aligned} \dot{A} &= -2\dot{\xi}A + 2A(Z_0 - A) + 2(Z_0 - A)A - 4BB^T \\ \dot{B} &= -2\dot{\xi}B + 2(Z_0 - A)B - 2AB - 4BC. \end{aligned} \tag{19}$$

We begin by controlling $\big\|B(t)\big\|_F$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big\|B\big\|_F^2 &= -2\dot{\xi}\big\|B\big\|_F^2 + 4\mathrm{Tr}\big((Z_0 - A)BB^T\big) - 4\mathrm{Tr}\big(ABB^T\big) - 8\mathrm{Tr}\big(BCB^T\big) \\ &\leq \big(-2\dot{\xi} + 4\lambda_{max}(Z_0 - A) - 4\lambda_{min}(A)\big)\big\|B\big\|_F^2. \end{aligned}$$

Therefore:

$$\left\|B(t)\right\|_F^2 \leq \left\|B_0\right\|_F^2 e^{-4\mu t} \exp\left(-\int_0^t \alpha(s)\mathrm{d}s\right), \tag{20}$$

with:

$$\alpha(t) = 2\dot{\xi}(t) - 4\lambda_{max}(Z_0 - A(t)) + 4\lambda_{min}(A(t)) - 4\mu \xrightarrow[t\to\infty]{} 0, \tag{21}$$

since $A(t) \xrightarrow[t\to\infty]{} Z_0$ and $\dot{\xi}(t) = \mathrm{Tr}(Z(t) - Z^*) \xrightarrow[t\to\infty]{} 0$ (note that $\lambda_{min}(Z_0) = \mu$). We now take care of $A(t)$ by introducing:

$$\phi(t) = \left\|A(t) - Z_0\right\|_F^2 + \frac{1}{2}\left[\mathrm{Tr}(A(t) - Z_0)\right]^2.$$

Differentiating and using the equation (19) on $A(t)$:

$$\begin{aligned}\dot{\phi} &= -2\mathrm{Tr}(AM^2) - 2\mathrm{Tr}(C)\mathrm{Tr}(AM) - 4\mathrm{Tr}(BB^TM) \\ &\leq -2\lambda_{min}(A)\mathrm{Tr}(M^2) + 2\mathrm{Tr}(C)\mathrm{Tr}(A^2)^{1/2}\mathrm{Tr}(M^2)^{1/2} + 4\mathrm{Tr}(BB^TBB^T)^{1/2}\mathrm{Tr}(M^2)^{1/2},\end{aligned}$$

where we defined $M = 2(A - Z_0) + \mathrm{Tr}(A - Z_0)I_{m^*}$ and used Cauchy-Schwartz inequality as well as Lemma B.1. Using the inequality $4\phi(t) \leq \mathrm{Tr}(M^2) \leq 2(4 + m^*)\phi(t)$, we obtain that:

$$\dot{\phi}(t) \leq -8\lambda_{min}(A(t))\phi(t) + 2\underbrace{\sqrt{2(4+m^*)}\big(\mathrm{Tr}(C(t))\mathrm{Tr}(A(t)) + 2\mathrm{Tr}(B(t)B(t)^T)\big)}_{\equiv \delta(t)}\sqrt{\phi(t)}.$$

We set $\Lambda(t) = \int_0^t \lambda_{min}(A(s))\mathrm{d}s \underset{t\to\infty}{\sim} \mu t$, and obtain:

$$\phi(t) \leq e^{-8\Lambda(t)}\left(\sqrt{\phi(0)} + \int_0^t e^{4\Lambda(s)}\delta(s)\mathrm{d}s\right)^2.$$

Using the definition of δ, we have two terms to bound. From equation (20):

$$\int_0^t e^{4\Lambda(s)}\mathrm{Tr}\big(B(s)B(s)^T\big)\mathrm{d}s \leq \left\|B_0\right\|_F^2 \int_0^t e^{4\Lambda(s)}e^{-4\mu s}\exp\left(-\int_0^s \alpha(u)\mathrm{d}u\right)\mathrm{d}s.$$

Since $\alpha(t) \xrightarrow[t\to\infty]{} 0$ and due to the behaviour of $\Lambda(t)$, this term is $o(e^{\beta t})$ for all $\beta > 0$ thanks to Lemma B.3. For the other term, we use equation (18) and split the cases. For $m = m^*$:

$$\int_0^t e^{4\Lambda(s)}\mathrm{Tr}(C(s))\mathrm{Tr}(A(s))\mathrm{d}s \underset{t\to\infty}{\lesssim} \int_0^t e^{4\Lambda(s)}e^{-4\mu s}\mathrm{Tr}(A(s))\mathrm{d}s,$$

with again is $o(e^{\beta t})$ for all $\beta > 0$. Thus, in this case, we have the bound:

$$\phi(t) \lesssim e^{-8\mu t}e^{v(t)},$$

where $v(t) = o(t)$ as $t \to \infty$. In the case $m > m^*$, again with equation (18) and Lemma B.3:

$$\int_0^t e^{4\Lambda(s)}\mathrm{Tr}(C(s))\mathrm{Tr}(A(s))\mathrm{d}s \underset{t\to\infty}{\lesssim} \frac{\mathrm{Tr}(Z_0)}{4\mu}\frac{e^{4\Lambda(t)}}{t},$$

which predominates over the first term. Therefore, for $m > m^*$, we get that $\phi(t) = O\big(t^{-2}\big)$. Finally, putting everything together:

$$\begin{aligned}\left\|Z(t) - Z^*\right\|_F^2 &\leq \left\|A(t) - Z_0\right\|_F^2 + 2\left\|B(t)\right\|_F^2 + \mathrm{Tr}\big(C(t)^2\big) \\ &\leq \phi(t) + 2\left\|B(t)\right\|_F^2 + \left[\mathrm{Tr}(C(t))\right]^2.\end{aligned}$$

For $m > m^*$, the second term is negligible and we get that $\|Z(t) - Z^*\|_F^2 = O(t^{-2})$. For $m = m^*$, the first and third term are of the order $e^{-8\mu t}$ (up to some corrections), and the leading term is given by $\|B(t)\|_F^2$. In this case:

$$\|Z(t) - Z^*\|_F^2 \underset{t\to\infty}{=} O\left(e^{-4\mu t}\exp\left(-\int_0^t \alpha(s)\mathrm{d}s\right)\right),$$

where α is defined in equation (21). We already have that $\int_0^t \alpha(s)\mathrm{d}s = o(t)$ as $t \to \infty$ so that we know the main behaviour of $\|Z(t) - Z^*\|_F$. We now show that $\int_0^t \alpha(s)\mathrm{d}s$ is bounded. From the expression of ξ:

$$\begin{aligned}\left|\int_0^t \alpha(s)\mathrm{d}s\right| &\leq 2\int_0^t \left|\mathrm{Tr}(Z(s) - Z^*)\right|\mathrm{d}s + 4\int_0^t \left|\lambda_{max}(Z_0 - A(s))\right|\mathrm{d}s + 4\int_0^t \left|\lambda_{min}(A(t)) - \mu\right|\mathrm{d}s \\ &\leq 2\sqrt{d}\int_0^t \|Z(s) - Z^*\|_F \mathrm{d}s + 8\int_0^t \|A(s) - Z_0\|_F \mathrm{d}s.\end{aligned}$$

Due to the bounds obtained in the case $m = m^*$, the integrals converge which leads to $\|Z(t) - Z^*\|_F^2 = O\big(e^{-4\mu t}\big)$. The result of the proposition follows from the bound:

$$\mathcal{L}(W(t)) = \frac{1}{2}\|Z(t) - Z^*\|_F^2 + \frac{1}{4}\big[\mathrm{Tr}(Z(t) - Z^*)\big]^2 \leq \left(\frac{1}{2} + \frac{d}{4}\right)\|Z(t) - Z^*\|_F^2.$$

C.3. Proof of Proposition 5.1

The main idea behind the proof is that ψ_d is solution of the implicit equation (6):

$$\psi_d(t) = \frac{1}{4}\mathrm{Tr}\log\left(I_{m_d} + 4W_d^{0T}\int_0^t e^{-2\psi_d(s)}e^{2s\mathrm{Tr}(Z_d^*)}e^{4Z_d^* s}\mathrm{d}s\; W_d^0\right). \tag{22}$$

In order to understand the high-dimensional limit of ψ_d one should be able to derive the limiting distribution of the matrix inside the log, which is not easily solved since this matrix depends on ψ_d. However, this becomes easier under Assumption A.1 which states that the matrices $U_d^0 U_d^{0T}$ and $U_d^* U_d^{*T}$ are uniformly drawn orthonormal projections (see Appendix A.2 for a more precise definition). Indeed, equation (9) shows that:

$$W_d^{0T}e^{4Z_d^* s}W_d^0 = \frac{1}{m_d}\left(I_{m_d} + \big(e^{4s/m_d^*} - 1\big)U_d^{0T}U_d^* U_d^{*T}U_d^0\right). \tag{23}$$

Thus, the integrals involving ψ_d can be detached from the matrices and the analysis will be simpler. The proof is split in different steps. We start by deriving properties of ϕ_d using the implicit equation (22). Through those steps, the goal is to obtain sufficient information on ϕ_d so that we can extract a converging subsequence (using Arzelà-Ascoli theorem). Finally, the goal will be to identify an equation which is verified in the limit (Lemma C.7) and show that its solution is unique (Lemma C.9).

In the following, we set $T > 0$ and denote $\mathcal{C}_T$ the space of continuous functions on $[0, T]$ equipped with the norm:

$$\|\chi\|_T = \sup_{\gamma\in[0,T]} |\chi(t)|.$$

C.3.1. Useful bounds

As mentioned earlier, the two following lemmas allow to gather sufficient information on ϕ_d in order to extract converging subsequences in $\mathcal{C}_T$ for $T > 0$. The first naive bound we obtain uses the fact that the loss function is always decreasing along a gradient flow.

Lemma C.5. *There exists $\kappa > 0$, such that:*

$$\sup_{\gamma\geq 0}\left|\dot{\phi}_d(\gamma)\right| \leq \kappa\sqrt{d}.$$

Proof. This property is a consequence of the gradient flow structure. Indeed, following equation (11):

$$\phi_d(\gamma) = \int_0^\gamma \mathrm{Tr}\big(W_d(s)W_d(s)^T - W_d^* W_d^{*T}\big)\mathrm{d}s,$$

where $W_d(t)$ is solution of the gradient flow (4) with initial condition W_d^0. Now, deriving the loss with respect to time:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(W_d(t)) = -\big\|\nabla\mathcal{L}(W_d(t))\big\|^2 \leq 0.$$

Thus, $\mathcal{L}(W_d(t))$ in non-increasing in time, and from the expression of the loss in equation (3):

$$\big[\mathrm{Tr}\big(W_d(t)W_d(t)^T - W_d^* W_d^{*T}\big)\big]^2 \leq 4\mathcal{L}(W_d(t)) \leq 4\mathcal{L}(W_d^0).$$

We now use the orthonormality assumption:

$$\begin{aligned}
\mathcal{L}(W_d^0) &= \frac{1}{2}\left\|\frac{1}{m_d}U_d^0 U_d^{0T} - \frac{1}{m_d^*}U_d^* U_d^{*T}\right\|_F^2 + \underbrace{\frac{1}{4}\left(\frac{1}{m_d}\mathrm{Tr}\big(U_d^0 U_d^{0T}\big) - \frac{1}{m_d^*}\mathrm{Tr}\big(U_d^* U_d^{*T}\big)\right)^2}_{=0} \\
&= \frac{1}{2}\left(\frac{1}{m_d^2}\mathrm{Tr}\big(U_d^0 U_d^{0T} U_d^0 U_d^{0T}\big) - \frac{2}{m_d m_d^*}\mathrm{Tr}\big(U_d^0 U_d^{0T} U_d^* U_d^{*T}\big) + \frac{1}{m_d^{*2}}\mathrm{Tr}\big(U_d^* U_d^{*T} U_d^* U_d^{*T}\big)\right) \\
&\leq \frac{1}{2}\left(\frac{1}{m_d} + \frac{1}{m_d^*}\right).
\end{aligned}$$

Since $m_d/d \xrightarrow[d\to\infty]{} \alpha$ and $m_d^*/d \xrightarrow[d\to\infty]{} \alpha^*$, we obtain that:

$$\big|\mathrm{Tr}\big(W_d(t)W_d(t)^T - W_d^* W_d^*\big)\big| \leq \frac{\kappa}{\sqrt{d}},$$

for some κ independent from t and d. Since $\dot\phi_d(\gamma) = d\,\mathrm{Tr}\big(W_d(\gamma d)W_d(\gamma d)^T - W_d^* W_d^*\big)$, we obtain the desired result. □

Lemma C.6. *For all $T > 0$:*

$$\sup_{d\in\mathbb{N}}\sup_{\gamma\in[0,T]}\big|\phi_d(\gamma)\big| < \infty.$$

Moreover, the family $(\phi_d)_{d\in\mathbb{N}}$ is equicontinuous on $[0,T]$, that is:

$$\sup_{d\in\mathbb{N}}\sup_{\substack{\gamma,\gamma'\in[0,T]\\|\gamma-\gamma'|\leq\eta}}\big|\phi_d(\gamma) - \phi_d(\gamma')\big| \xrightarrow[\eta\to 0]{} 0. \tag{24}$$

This lemma will be proven in Appendix D.2. Indeed, the proof is long and do not carry many elements of interest. The main idea is to start from the result of Lemma C.5 and use it in the self-consistent equation solved by ϕ_d, which we prove to be:

$$\phi_d(\gamma) + \gamma d = \frac{1}{4}\mathrm{Tr}\log\left(\left(1 + \frac{4d}{m_d}\int_0^\gamma e^{-2\phi_d(s)}\right)I_{m_d} + \frac{4d}{m_d}\int_0^\gamma e^{-2\chi(s)}\big(e^{4sd/m_d^*} - 1\big)\mathrm{d}s\; U_d^{0T}U_d^* U_d^{*T}U_d^0\right). \tag{25}$$

C.3.2. Identifying the limit

The previous lemma shows that the family $(\phi_d)_{d\in\mathbb{N}}$ is compact in $\mathcal{C}_T$, i.e., it allows us to extract converging subsequences. The following lemmas will show that those subsequences always converge to the same function.

Lemma C.7. *Let μ be defined as in equation* (10) *and $\phi : [0,T] \to \mathbb{R}$ a subsequential limit of $(\phi_d)_{d\in\mathbb{N}}$ in $\mathcal{C}_T$.*

Then, with probability one, for all $\gamma \in [0,T]$:

$$\int \log\left(1+\frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}\mathrm{d}s+\frac{4x}{\alpha}\int_0^\gamma e^{-2\phi(s)}\big(e^{4s/\alpha^*}-1\big)\mathrm{d}s\right)\mathrm{d}\mu(x)=\frac{4\gamma}{\alpha}. \tag{26}$$

Proof. For $\xi \in \mathcal{C}_T$, we set:

$$\mathcal{H}_d(\xi)(\gamma)=\frac{1}{4d}\mathrm{Tr}\log\left(\left(1+\frac{4d}{m_d}\int_0^\gamma e^{-2\xi(s)}\mathrm{d}s\right)I_{m_d}+\frac{4d}{m_d}\int_0^\gamma e^{-2\xi(s)}\big(e^{4sd/m_d^*}-1\big)\mathrm{d}s\,Y_d\right).$$

From equation (25), we have that, for all $\gamma \geq 0$:

$$\mathcal{H}_d(\phi_d)(\gamma)=\gamma+\frac{\phi_d(\gamma)}{d}\equiv u_d(\gamma). \tag{27}$$

From the bound obtained in Lemma C.5, we have that $\sup_{\gamma\in[0,T]}|\phi_d(\gamma)| \leq \kappa T\sqrt{d}$, therefore $u_d(\gamma) \xrightarrow[d\to\infty]{} \gamma \equiv u(\gamma)$ uniformly on $[0,T]$. We know that the empirical spectral distribution of Y_d converges towards μ defined in equation (10). Therefore, we also define:

$$\mathcal{H}(\xi)(\gamma)=\frac{\alpha}{4}\int\log\left(1+\frac{4}{\alpha}\int_0^\gamma e^{-2\xi(s)}\mathrm{d}s+\frac{4x}{\alpha}\int_0^\gamma e^{-2\xi(s)}\big(e^{4s/\alpha^*}-1\big)\mathrm{d}s\right)\mathrm{d}\mu(x). \tag{28}$$

It is reasonable to hope that $\mathcal{H}_d(\xi) \longrightarrow \mathcal{H}(\xi)$ from the convergence of Y_d and the fact that $m_d/d \longrightarrow \alpha$ and $m_d^*/d \longrightarrow \alpha^*$ as $d \to \infty$. To do so, we introduce the following lemma that we prove in Appendix D.3:

Lemma C.8. *For $c \in \mathbb{R}$, let $B_c = \big\{\xi \in \mathcal{C}_T,\, \xi \geq c\big\}$. Then, with probability one:*

$$\sup_{\xi\in B_c}\big\|\mathcal{H}_d(\xi)-\mathcal{H}(\xi)\big\|_T \xrightarrow[d\to\infty]{} 0. \tag{29}$$

The main ingredient of this result is the uniform convergence obtained in Lemma A.1 for the empirical spectral distribution of the matrix Y_d. However, we need to be careful since there are other quantities depending on the dimension.

Now, to prove Lemma C.7, we take an extraction $d(n)$ and suppose that $\phi_{d(n)}$ converges towards some ϕ is $\mathcal{C}_T$. From equation (27), we have that $\mathcal{H}_{d(n)}(\phi_{d(n)}) = u_{d(n)} \xrightarrow[n\to\infty]{} u$. We are going to show that with probability one, $\mathcal{H}(\phi) = u$. This will give the desired result from the definition of $\mathcal{H}$ in equation (28). We have:

$$\big\|\mathcal{H}(\phi)-u\big\|_T \leq \big\|\mathcal{H}(\phi)-\mathcal{H}(\phi_{d(n)})\big\|_T+\big\|\mathcal{H}(\phi_{d(n)})-\mathcal{H}_{d(n)}(\phi_{d(n)})\big\|_T+\big\|\mathcal{H}_{d(n)}(\phi_{d(n)})-u\big\|_T.$$

We already know that the last term goes to zero. For the first one, we need to show that $\mathcal{H}$ is continuous. Indeed, for $\xi, \varsigma \in \mathcal{C}_T$ both lower bounded by some $c \in \mathbb{R}$:

$$\begin{aligned}
\big|\mathcal{H}(\xi)(\gamma)-\mathcal{H}(\varsigma)(\gamma)\big| \leq \frac{\alpha}{4}\sup_{x\in[0,1]}\Bigg|&\log\left(1+\frac{4}{\alpha}\int_0^\gamma e^{-2\xi(s)}\mathrm{d}s+\frac{4x}{\alpha}\int_0^\gamma e^{-2\xi(s)}\big(e^{4s/\alpha^*}-1\big)\mathrm{d}s\right)\\
&-\log\left(1+\frac{4}{\alpha}\int_0^\gamma e^{-2\varsigma(s)}\mathrm{d}s+\frac{4x}{\alpha}\int_0^\gamma e^{-2\varsigma(s)}\big(e^{4s/\alpha^*}-1\big)\mathrm{d}s\right)\Bigg|\\
\leq \sup_{x\in[0,1]}&\int_0^\gamma\big|e^{-2\xi(s)}-e^{-2\varsigma(s)}\big|\mathrm{d}s+x\int_0^\gamma e^{4s/\alpha^*}\big|e^{-2\xi(s)}-e^{-2\varsigma(s)}\big|\mathrm{d}s\\
\leq\ & 2e^{-2c}T(1+e^{4T/\alpha^*})\big\|\xi-\varsigma\big\|_T.
\end{aligned}$$

Since by Lemma C.6, $(\phi_d)_{d\in\mathbb{N}}$ is uniformly bounded on $[0,T]$, it is in B_c for some $c \in \mathbb{R}$. This implies that the first term goes to zero. From equation (29), this also implies that the second term goes to zero with probability one. As a conclusion, we necessarily have that $\mathcal{H}(\phi) = u$ which proves the lemma. □

Lemma C.9. *Equation* (26) *has a unique continuous solution on* $\mathbb{R}^+$.

This lemma is technical and uses the Picard-Lindelöf theorem which gives the existence and uniqueness for the solutions of differential equations. First, equation (26) has to be interpreted as a differential equation, which can be done using equation (12). We prove this lemma in Appendix D.4.

The proof of Proposition 5.1 is now elementary. From Lemma C.6 and Lemma C.8, the family $(\phi_d)_{d\in\mathbb{N}}$ is compact in $\mathcal{C}_T$. Thus, it admits at least one subsequential limit by Arzelà-Ascoli theorem. From Lemma C.7, such a limit verifies $\mathcal{H}(\phi) = u$ with probability one. Since this equation admits a unique solution from Lemma C.9, $(\phi_d)_{d\in\mathbb{N}}$ has a single subsequential limit with probability one. Thus, outside the event of zero probability:

$$\left\{ \sup_{\xi\in B_c} \left\| \mathcal{H}_d(\xi) - \mathcal{H}(\xi) \right\|_T \not\xrightarrow[d\to\infty]{} 0 \right\},$$

$(\phi_d)_{d\in\mathbb{N}}$ uniformly converges on $[0,T]$. Here $c \in \mathbb{R}$ is such that $(\phi_d)_{d\in\mathbb{N}}$ is contained in B_c (exists and independent of the chosen extraction from Lemma C.6). To conclude, taking a sequence $T_n \xrightarrow[n\to\infty]{} \infty$, this proves that almost surely, $(\phi_d)_{d\in\mathbb{N}}$ uniformly converges on each $[0,T_n]$, thus on every compact of $\mathbb{R}^+$. From Lemma C.7, the limit function ϕ is solution of equation (26). Replacing μ from its definition in equation (10) leads to equation (12) which concludes the proof of Proposition 5.1.

C.4. Proof of Proposition 5.2

Proposition 5.2 is decomposed into two parts: first we show the uniform convergence of the overlap between the teachers and students at timescale $t = \gamma d$. Then we derive the limit of the overlap as $\gamma \to \infty$. Note that our proof method will also give access to the speed at which the convergence occurs as $\gamma \to \infty$. This will be to be compared with the convergence rates we obtained at finite dimension (see Proposition A.1).

C.4.1. High-dimensional limit for the overlap

We now prove the first part of the proposition, namely the overlap uniformly converges as $d \to \infty$. This result is a consequence of Proposition 5.1 as well as the convergence of the empirical spectral distribution of the matrix Y_d as $d \to \infty$. We first define some quantities that will be useful in the following:

$$F_d(\gamma) = 1 + \frac{4d}{m_d}\int_0^\gamma e^{-2\phi_d(s)}\mathrm{d}s, \qquad G_d(\gamma) = 1 + \frac{4d}{m_d}\int_0^\gamma e^{-2\phi_d(s)}e^{4sd/m_d^*}\mathrm{d}s, \qquad J_d(\gamma) = \frac{G_d(\gamma)}{F_d(\gamma)}.$$

Since it has been shown that ϕ_d converges as $d \to \infty$, it is easily shown that these quantities also converge as $d \to \infty$. We denote F, G and J their infinite dimensional counterparts.

Lemma C.10. *Let* $\chi_d(\gamma)$ *be defined in Proposition 5.2. Then:*

$$\chi_d(\gamma) = \sqrt{\frac{m_d}{m_d^*}}e^{4\gamma d/m_d^*}\frac{m_d^{-1}\mathrm{Tr}\Big(Y_d\big(I_{m_d} + (J_d(\gamma)-1)Y_d\big)^{-1}\Big)}{\sqrt{m_d^{-1}\mathrm{Tr}\Big(\big(I_{m_d} + (e^{4\gamma d/m_d^*}-1)Y_d\big)^2\big(I_{m_d} + (J_d(\gamma)-1)Y_d\big)^{-2}\Big)}}, \tag{30}$$

with $Y_d = U_d^{0T}U_d^*U_d^{*T}U_d^0$.

Before proving this lemma, we observe that at finite dimension, the overlap mainly depends on two quantities: the matrix Y_d (which is reasonable due to our orthonormality assumption) and the function $J_d(\gamma)$. This function also converges as $d \to \infty$, therefore this previous equation suggests that the overlap will also converge.

Proof. Using the implicit solution obtain in Proposition 3.1:

$$Z_d(t) = e^{-2\psi_d(t)}e^{2\mathrm{Tr}(Z_d^*)t}e^{2Z_d^*t}W_d^0\left(I_{m_d} + 4\int_0^t e^{-2\psi_d(t)}e^{2s\mathrm{Tr}(Z_d^*)}W_d^{0T}e^{4Z_d^*s}W_d^0\mathrm{d}s\right)^{-1}W_d^{0T}e^{2Z_d^*t}.$$

Introducing $\gamma = t/d$ and using the definitions of $\phi_d(\gamma)$, U_d^0, U_d^*:

$$Z_d(\gamma d) = \frac{1}{m_d} e^{-2\phi_d(\gamma)} \exp\left(\frac{2d\gamma}{m_d^*} U_d^* U_d^{*T}\right) U_d^0 N_d(\gamma)^{-1} U_d^{0T} \exp\left(\frac{2d\gamma}{m_d^*} U_d^* U_d^{*T}\right),$$

with:

$$\begin{aligned} N_d(\gamma) &= \left(1 + \frac{4d}{m_d}\int_0^\gamma e^{-2\phi_d(s)}\mathrm{d}s\right) I_{m_d} + \frac{4d}{m_d}\int_0^\gamma e^{-2\phi_d(s)}\big(e^{4sd/m_d^*} - 1\big)\mathrm{d}s\, Y_d \\ &\equiv F_d(\gamma) + \big(G_d(\gamma) - F_d(\gamma)\big) Y_d. \end{aligned}$$

Now using the fact that $U_d^* U_d^{*T}$ is a projection:

$$\exp\left(\frac{2d\gamma}{m_d^*} U_d^* U_d^{*T}\right) = I_d - U_d^* U_d^{*T} + \exp\left(\frac{2d\gamma}{m_d^*}\right) U_d^* U_d^{*T}.$$

And denoting $\lambda_d^* = d/m_d^*$, we obtain:

$$\begin{aligned} \mathrm{Tr}\big(Z_d^* Z_d(\gamma d)\big) &= \frac{1}{m_d m_d^*} e^{-2\phi_d(\gamma)} \mathrm{Tr}\Big(U_d^* U_d^{*T}\big(I_d + (e^{2\lambda_d^*\gamma} - 1) U_d^* U_d^{*T}\big) U_d^0 N_d(\gamma)^{-1} U_d^{0T}\big(I_d + (e^{2\lambda_d^*\gamma} - 1) U_d^* U_d^{*T}\big)\Big) \\ &= \frac{1}{m_d m_d^*} e^{-2\phi_d(\gamma)} e^{4\lambda_d^*\gamma} \mathrm{Tr}\Big(Y_d\big(F_d(\gamma) I_{m_d} + \big(G_d(\gamma) - F_d(\gamma)\big) Y_d\big)^{-1}\Big). \end{aligned} \tag{31}$$

Now, to compute the overlap, one has to compute the norms $\big\|Z_d^*\big\|_F$ and $\big\|Z_d(\gamma d)\big\|_F$. The first one is easily done:

$$\big\|Z_d^*\big\|_F^2 = \frac{1}{m_d^{*2}} \mathrm{Tr}\Big(U_d^* U_d^{*T} U_d^* U_d^{*T}\Big) = \frac{1}{m_d^*}. \tag{32}$$

For the second one, one can do a similar calculation as for the trace:

$$\big\|Z_d(\gamma d)\big\|_F^2 = \frac{1}{m_d^2} e^{-4\phi_d(\gamma)} \mathrm{Tr}\Big(\big(I_{m_d} + (e^{4\lambda_d^*\gamma} - 1) Y_d\big)^2 \big(F_d(\gamma) I_{m_d} + \big(G_d(\gamma) - F_d(\gamma)\big) Y_d\big)^{-2}\Big). \tag{33}$$

Assembling equation (31), (32) and (33) gives the desired result. □

The goal is now to prove the uniform convergence of the overlap in the high-dimensional limit. This is achieved by the following lemma:

Lemma C.11. *Let $T > 0$. Then, uniformly on $[0, T]$:*

$$\begin{aligned} \frac{1}{m_d}\mathrm{Tr}\Big(Y_d\big(I_{m_d} + (J_d(\gamma) - 1) Y_d\big)^{-1}\Big) &\xrightarrow[d\to\infty]{} \int \frac{x}{1 + (J(\gamma) - 1)x} \mathrm{d}\mu(x) \equiv A(\gamma) \\ \frac{1}{m_d}\mathrm{Tr}\Big(\big(I_{m_d} + (e^{4\lambda_d^*\gamma} - 1) Y_d\big)^2 \big(I_{m_d} + (J_d(\gamma) - 1) Y_d\big)^{-2}\Big) &\xrightarrow[d\to\infty]{} \int \left(\frac{1 + \big(e^{4\gamma/\alpha^*} - 1\big)x}{1 + (J(\gamma) - 1)x}\right)^2 \mathrm{d}\mu(x) \\ &\equiv B(\gamma). \end{aligned}$$

As a consequence, uniformly on $[0, T]$:

$$\chi_d(\gamma) \xrightarrow[d\to\infty]{} \sqrt{\frac{\alpha}{\alpha^*}} e^{4\gamma/\alpha^*} \frac{A(\gamma)}{\sqrt{B(\gamma)}} \equiv \chi(\gamma). \tag{34}$$

The proof of this technical lemma is deferred to Appendix D.5. The main ingredients are the fact that J_d uniformly converges towards J, as well as the uniform convergence result for the empirical spectral distribution of the matrix Y_d presented in Lemma A.1.

This last formula give the overlap in the high-dimensional limit. This depends on α, α^* (also through the measure

μ defined in equation (10)) and on the function J. In the following, we obtain an understanding of this function as $\gamma \to \infty$, which leads to the determination of the asymptotic behaviour of the overlap.

C.4.2. Limit and convergence rate for the overlap

We now prove the second part of Proposition 5.2 and determine in addition at which rate the convergence $\lim_{\gamma\to\infty} \chi(\gamma)$ occurs. From the result of Lemma C.11, the main challenge is to determine the behaviour of $J(\gamma)$ as $\gamma \to \infty$. This is done in the following lemma. We remind the definitions of F, G, J, depending on the function $\phi = \lim_{d\to\infty} \phi_d$:

$$F(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}\mathrm{d}s \qquad G(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}\mathrm{d}s \qquad J(\gamma) = \frac{G(\gamma)}{F(\gamma)}. \tag{35}$$

Lemma C.12. *There exists a constant $\kappa(\alpha, \alpha^*) > 0$ such that:*

$$J(\gamma) \underset{\gamma\to\infty}{\sim} \begin{cases} \kappa(\alpha,\alpha^*)e^{4\gamma/\alpha^*} & \text{for } \alpha < \alpha^* \\ \kappa(\alpha,\alpha^*)\gamma^{-1}e^{4\gamma/\alpha^*} & \text{for } \alpha \geq \alpha^*. \end{cases}$$

This technical lemma is proven in Appendix D.6. It mainly uses the implicit equation on ϕ from Proposition 5.1.

The understanding of the asymptotics of J, along with the limit we identified in Lemma C.11 allows to determine the behaviour of $\chi(\gamma)$ as $\gamma \to \infty$. This is a stronger version of Proposition 5.2 as we also determine the rate at which convergence occurs. Note that, unlike the convergence rates we determine in the finite dimensional case (see Proposition A.1), the asymptotics we obtain in the following are exact.

Before obtaining the asymptotic behaviour of $\chi(\gamma)$ as $\gamma \to \infty$, we need a last technical lemma in order to obtain the asymptotics of $A(\gamma), B(\gamma)$ as $\gamma \to \infty$ (defined in Lemma C.11).

Lemma C.13. *Let $A(\gamma)$, $B(\gamma)$ be defined in Lemma C.12. Then:*

$$A(\gamma) \underset{\gamma\to\infty}{=} \begin{cases} \dfrac{\min(\alpha,\alpha^*)}{\alpha J(\gamma)} + O\left(\dfrac{1}{J(\gamma)^2}\right) & \text{for } \alpha \neq \alpha^* \\ \dfrac{1}{J(\gamma)} + O\left(\dfrac{1}{J(\gamma)^{3/2}}\right) & \text{for } \alpha = \alpha^* \end{cases},$$

and:

$$B(\gamma) \underset{\gamma\to\infty}{=} \begin{cases} \left(1 - \dfrac{\alpha^*}{\alpha}\right)^+ + \dfrac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\left(\dfrac{\min(\alpha,\alpha^*)}{\alpha} + O\left(\dfrac{1}{J(\gamma)}\right)\right) & \text{for } \alpha \neq \alpha^* \\ \dfrac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\left(1 + O\left(\dfrac{1}{J(\gamma)^{1/2}}\right)\right) & \text{for } \alpha = \alpha^*. \end{cases}$$

This lemma is proven in Appendix D.7. Using this result, we are now ready to establish $\lim_{\gamma\to\infty} \chi(\gamma)$ as well as the convergence rate:

Lemma C.14. *The asymptotic behaviour of $\chi(\gamma)$ is given by:*

$$\chi(\gamma) \underset{\gamma\to\infty}{=} \begin{cases} \sqrt{\dfrac{\alpha}{\alpha^*}} + O\big(e^{-4\gamma/\alpha^*}\big) & \text{for } \alpha < \alpha^* \\ 1 + O\big(\sqrt{\gamma}e^{-2\gamma/\alpha^*}\big) & \text{for } \alpha = \alpha^* \\ 1 + O\big(\gamma^{-2}\big) & \text{for } \alpha > \alpha^*. \end{cases}$$

Proof. The proof is a simple consequence of the two previous lemmas and equation (34) giving the expression

of $\chi(\gamma)$. Starting with the case $\alpha < \alpha^*$, we obtain:

$$\chi(\gamma) \underset{\gamma\to\infty}{=} \sqrt{\frac{\alpha}{\alpha^*}} \frac{1 + O\big(J(\gamma)^{-1}\big)}{\sqrt{1 + O\big(J(\gamma)^{-1}\big)}} = \sqrt{\frac{\alpha}{\alpha^*}} + O\big(J(\gamma)^{-1}\big),$$

and the result from Lemma C.12. Likewise, for $\alpha = \alpha^*$:

$$\chi(\gamma) \underset{\gamma\to\infty}{=} \frac{1 + O\big(J(\gamma)^{-1/2}\big)}{\sqrt{1 + O\big(J(\gamma)^{-1/2}\big)}} = 1 + O\big(J(\gamma)^{-1/2}\big).$$

Finally, if $\alpha > \alpha^*$:

$$\chi(\gamma) \underset{\gamma\to\infty}{=} \sqrt{\frac{\alpha^*}{\alpha}} \frac{1 + O\big(J(\gamma)^{-1}\big)}{\sqrt{\left(1 - \frac{\alpha^*}{\alpha}\right) \frac{J(\gamma)^2}{e^{8\gamma/\alpha^*}} + \frac{\alpha^*}{\alpha} + O\big(J(\gamma)^{-1}\big)}} = \frac{1 + O\big(J(\gamma)^{-1}\big)}{\sqrt{1 + O\big(\gamma^{-2}\big)}} = 1 + O\big(\gamma^{-2}\big),$$

which concludes the proof. □

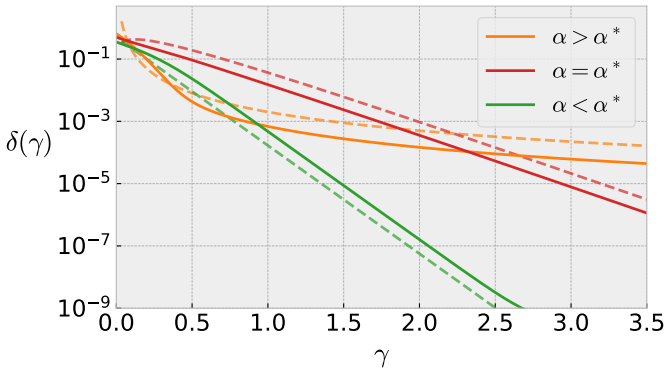

Figure 6: Evolution of $\delta(\gamma) = \chi_\infty - \chi(\gamma)$ (where $\chi_\infty = \lim_{\gamma\to\infty} \chi(\gamma)$) as a function of γ for different values of α, α^*. Dashed lines: convergence rates obtained in Lemma C.14. Simulated using equation (34), with standard discretization for the integral computation ($\alpha \neq \alpha^*$) and exact solution ($\alpha = \alpha^*$).

Figure 6 compares a simulated version of equation (34) with the bounds obtained in the previous lemma. As expected, those bounds are tight since they were obtained using an exact asymptotic development.

This last result is to be compared with the convergence rates obtained in finite dimension (see Proposition A.1). In the overparameterized case $m \geq m^*$, we obtained $Z(t) \xrightarrow[t\to\infty]{} Z^*$. Regarding the overlap:

$$\begin{aligned}\frac{\mathrm{Tr}(Z(t)Z^*)}{\|Z(t)\|_F \|Z^*\|_F} &= \frac{1}{2}\left(\frac{\|Z(t)\|_F}{\|Z^*\|_F} + \frac{\|Z^*\|_F}{\|Z(t)\|_F} - \frac{\|Z(t) - Z^*\|_F^2}{\|Z^*\|_F \|Z(t)\|_F}\right) \\ &= 1 + \Theta\big(\|Z(t) - Z^*\|_F\big)\end{aligned}$$

From the rates obtained in Proposition A.1, we obtain that the overlap is $1 + O(t^{-1})$ whenever $m > m^*$. In the high-dimensional limit, the convergence is faster for $\alpha > \alpha^*$ from Lemma C.13. As for the case $m = m^*$, the convergence at finite dimension was exponential and the speed was proportional to the smallest non-zero eigenvalue of Z^*. In our orthonormal setup, this smallest eigenvalue is given by $(m^*)^{-1}$. This leads to:

$$\frac{\mathrm{Tr}(Z(t)Z^*)}{\|Z(t)\|_F \|Z^*\|_F} = 1 + O\left(e^{-2t/m^*}\right).$$

In the high-dimensional limit where $t = \gamma d$ and $m^*/d \to \alpha^*$, we obtain a rate $O(e^{-2\gamma/\alpha^*})$ which is the one obtained in Lemma C.13 up to a correction proportional to $\sqrt{\gamma}$.

D. Proofs of the Technical Lemmas

D.1. Proof of Lemma C.3

For $k = 1, \dots, d$, we set $z_k = W^{0T} v_k^* \in \mathbb{R}^m$, so that:

$$W^{0T} E^*(t) W^0 = \sum_{k=1}^{d} \gamma_k(t) z_k z_k^T,$$

with:

$$\gamma_k(t) = \int_0^t e^{-2\psi(s)} e^{2s\mathrm{Tr}(Z^*)} e^{4\mu_k s} \mathrm{d}s \underset{t\to\infty}{\sim} \frac{1}{4\mu_k} e^{-2\psi(t)} e^{2t\mathrm{Tr}(Z^*)} e^{4\mu_k t}, \tag{36}$$

whenever $\mu_k > 0$. This is a consequence of Lemma B.3 along with the fact that $\dot{\psi}(t) = \mathrm{Tr}(W(t)W(t)^T) \longrightarrow \mathrm{Tr}(Z^*)$ as $t \to \infty$. When $\mu_k = 0$, then $\gamma_k(t) = o\big(e^{\beta t}\big)$ for all $\beta > 0$. Thus, is $\mu_k > \mu_j$, we have that $\gamma_k(t)\gamma_j(t)^{-1} \longrightarrow \infty$ as $t \to \infty$. Thus, we have some $t_0 > 0$ such that $\gamma_1(t) \geq \cdots \geq \gamma_d(t) \geq 0$ as soon as $t \geq t_0$. Using Courant-Fischer formula (Lemma B.2), we have for $j \in [\![1, \min(m, d)]\!]$:

$$\lambda_j(t) = \max_{\substack{V\subset\mathbb{R}^m \\ \dim V = j}} \min_{\substack{x\in V \\ \|x\|=1}} \left(\sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2 \right) = \min_{\substack{V\subset\mathbb{R}^m \\ \dim V = m-j+1}} \max_{\substack{x\in V \\ \|x\|=1}} \left(\sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2 \right). \tag{37}$$

We start by using the second equality and choose $V_1 = \mathrm{Span}(z_1, \dots, z_{j-1})^\perp$ which is of dimension $m - j + 1$ by assumption. Then, for $t \geq t_0$:

$$\lambda_j(t) \leq \max_{\substack{x\in V_1 \\ \|x\|=1}} \left(\sum_{k=j}^{d} \gamma_k(t)(x \cdot z_k)^2 \right) \leq \gamma_j(t) \underbrace{\max_{\substack{x\in V_1 \\ \|x\|=1}} \sum_{k=j}^{d} (x \cdot z_k)^2}_{\equiv C_j}.$$

We now use the first equality of equation (37). We choose $V_2 = \mathrm{Span}(z_1, \dots, z_j)$ which is of dimension j by assumption. Then:

$$\lambda_j(t) \geq \min_{\substack{x\in V_2 \\ \|x\|=1}} \left(\sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2 \right) \geq \min_{\substack{x\in V_2 \\ \|x\|=1}} \left(\sum_{k=1}^{j} \gamma_k(t)(x \cdot z_k)^2 \right) \geq \gamma_j(t) \underbrace{\min_{\substack{x\in V_2 \\ \|x\|=1}} \left(\sum_{k=1}^{j} (x \cdot z_k)^2 \right)}_{\equiv D_j},$$

for $t \geq t_0$. By definition of V_2, D_j cannot be zero. Equation (36) allows to conclude the proof.

D.2. Proof of Lemma C.6

We divide the proofs into two parts. We first prove that the family $(\phi_d)_{d\in\mathbb{N}}$ is uniformly bounded, then we show it is equicontinuous.

D.2.1. Uniform bound

To prove the first point, we start by determining the implicit equation solved by ϕ_d. Using equation (22) making the substitution $s \mapsto s/d$:

$$\psi_d(t) = \frac{1}{4} \mathrm{Tr} \log \left(I_{m_d} + \frac{4d}{m_d} U_d^{0T} \int_0^{t/d} e^{-2\phi_d(s)} \exp\left(\frac{4sd}{m_d^*} U_d^* U_d^{*T} \right) \mathrm{d}s \, U_d^0 \right).$$

Therefore, introducing $\gamma = t/d$, $\lambda_d = d/m_d$ and $\lambda_d^* = d/m_d^*$:

$$\phi_d(\gamma) + \gamma d = \frac{1}{4}\mathrm{Tr}\log\left(I_{m_d} + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}\left(I_{m_d} + \left(e^{4\lambda_d^* s} - 1\right)U_d^{0T}U_d^* U_d^{*T}U_d^0\right)\mathrm{d}s\right),$$

where we used that $\phi_d(\gamma) = \psi_d(\gamma d) - \gamma d$ (in our setup $\mathrm{Tr}(Z_d^*) = 1$). This is precisely equation (25). We set $Y_d = U_d^{0T}U_d^* U_d^{*T}U_d^0$ and:

$$F_d(\gamma) = 1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}\mathrm{d}s \qquad G_d(\gamma) = 1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}e^{4\lambda_d^* s}\mathrm{d}s,$$

so that:

$$\phi_d(\gamma) + \gamma d = \frac{1}{4}\mathrm{Tr}\log\Big(F_d(\gamma)(1 - Y_d) + G_d(\gamma)Y_d\Big).$$

Deriving with respect to γ:

$$\dot\phi_d(\gamma) + d = \frac{1}{4}\mathrm{Tr}\big[U_d(\gamma, Y_d)\big], \tag{38}$$

with:

$$U_d(\gamma, x) = \frac{\dot F_d(\gamma)(1-x) + \dot G_d(\gamma)x}{F_d(\gamma)(1-x) + G_d(\gamma)x}.$$

Using that $\dot G_d(\gamma) = e^{4\lambda_d^*\gamma}\dot F_d(\gamma)$ and integrating by parts:

$$G_d(\gamma) = e^{4\lambda_d^*\gamma}F_d(\gamma) - 4\lambda_d^* \int_0^\gamma e^{4\lambda_d^* s}F_d(s)\mathrm{d}s \le e^{4\lambda_d^*\gamma}F_d(\gamma).$$

Thus:

$$\dot F_d(\gamma)G_d(\gamma) - F_d(\gamma)\dot G_d(\gamma) = 4\lambda_d e^{-2\phi_d(\gamma)}\big(G_d(\gamma) - F_d(\gamma)e^{4\lambda_d^*\gamma}\big) \le 0.$$

As a consequence, for every $\gamma \ge 0$, the map $x \mapsto U_d(\gamma, x)$ is non-decreasing. Using equation (38) and the fact that the eigenvalues of Y_d are contained in $[0, 1]$:

$$\frac{m_d}{4}\frac{\dot F_d(\gamma)}{F_d(\gamma)} \le \dot\phi_d(\gamma) + d \le \frac{m_d}{4}\frac{\dot G_d(\gamma)}{G_d(\gamma)}.$$

Using the bound on $\dot\phi_d$ obtained in Lemma C.5, we obtain:

$$\frac{\dot F_d(\gamma)}{F_d(\gamma)} \le 4\kappa\frac{\sqrt{d}}{m_d} + 4\lambda_d \qquad \frac{\dot G_d(\gamma)}{G_d(\gamma)} \ge -4\kappa\frac{\sqrt{d}}{m_d} + 4\lambda_d.$$

This can be integrated to obtain a bound on ϕ_d:

$$\left(1 - \frac{\kappa}{\sqrt{d}}\right)e^{-4\lambda_d^*\gamma}\exp\left(4\lambda_d\left(1 - \frac{\kappa}{\sqrt{d}}\right)\gamma\right) \le e^{-2\phi_d(\gamma)} \le \left(1 + \frac{\kappa}{\sqrt{d}}\right)\exp\left(4\lambda_d\left(1 + \frac{\kappa}{\sqrt{d}}\right)\gamma\right).$$

Since λ_d and λ_d^* converge as $d \to \infty$, this proves the first point.

D.2.2. Equicontinuity

We now show the equicontinuity of the family $(\phi_d)_{d\in\mathbb{N}}$ on $[0, T]$. Back to equation (38), using the definition of $F_d(\gamma)$ and $G_d(\gamma)$:

$$U_d(\gamma, x) = 4\lambda_d e^{-2\phi_d(\gamma)}\underbrace{\frac{1 - x + xe^{4\lambda_d^*\gamma}}{1 + 4\lambda_d\int_0^\gamma e^{-2\phi_d(s)}(1 - x + xe^{4\lambda_d^* s})\mathrm{d}s}}_{\equiv v_d(\gamma, x)}.$$

Thus, using equation (38) and rearranging:

$$\phi_d(\gamma) = -\frac{1}{2}\log\Bigg(1 + \underbrace{\frac{\dot{\phi}_d(\gamma)}{d}}_{\equiv \epsilon_d(\gamma)}\Bigg) + \frac{1}{2}\log\Bigg(\underbrace{\frac{1}{m_d}\mathrm{Tr}\big[v_d(\gamma, Y_d)\big]}_{\equiv T_d(\gamma)}\Bigg).$$

Now, it can be shown that:

$$\left|\dot{T}_d(\gamma)\right| \leq \sup_{x\in[0,1]}\left|\frac{\partial v_d}{\partial \gamma}(\gamma, x)\right| \leq 4\lambda_d^* e^{4\lambda_d^* T} + 4\lambda_d e^{-2\phi_d(\gamma)} e^{8\lambda_d^* T}.$$

Since ϕ_d is uniformly bounded in d and on $[0,T]$ from the previous step, then this quantity is uniformly bounded in d and on $[0,T]$, say by $L > 0$. We let $\gamma, \gamma' \in [0,T]$ and look at:

$$|\phi_d(\gamma) - \phi_d(\gamma')| \leq \frac{1}{2}\left|\log\left(1 + \frac{\epsilon_d(\gamma') - \epsilon_d(\gamma)}{1 + \epsilon_d(\gamma)}\right)\right| + \frac{1}{2}\left|\log\left(1 + \frac{T_d(\gamma') - T_d(\gamma)}{T_d(\gamma)}\right)\right|. \tag{39}$$

Moreover:

$$T_d(\gamma) \geq \inf_{x\in[0,1]} v_d(\gamma, x) \geq \left(1 + \frac{\lambda_d}{\lambda_d^*} e^{-2c} e^{4\lambda_d T}\right)^{-1},$$

where $c \in \mathbb{R}$ is such that $\phi_d(\gamma) \geq c$ for all $\gamma \in [0,T]$ and $d \in \mathbb{N}$ (such a constant exists from the previous step). This proves that $T_d(\gamma)$ is lower bounded by some constant $\rho > 0$, independent from $\gamma \in [0,T]$ and d. As a consequence:

$$\left|\frac{T_d(\gamma) - T_d(\gamma')}{T_d(\gamma)}\right| \leq \frac{L}{\rho}\big|\gamma - \gamma'\big|,$$

for all $\gamma, \gamma' \in [0,T]$ and $d \in \mathbb{N}$. Take $\eta \leq \rho/(2L)$ and suppose that $|\gamma - \gamma'| \leq \eta$. Then using the inequality $\big|\log(1+x)\big| \leq 2|x|$ for $x \in [-1/2, 1/2]$, the second term of equation (39):

$$\left|\log\left(1 + \frac{T_d(\gamma') - T_d(\gamma)}{T_d(\gamma)}\right)\right| \leq \frac{2L}{\rho}\eta.$$

For the first term, we use the bound obtained in Lemma C.5, which implies that:

$$\sup_{\gamma\in[0,T]}\big|\epsilon_d(\gamma)\big| \leq \frac{\kappa}{\sqrt{d}}.$$

Thus, for $d \geq \lfloor \kappa^2 \max(25, (1+2/\eta)^2)\rfloor \equiv d_0$, we obtain:

$$\left|\log\left(1 + \frac{\epsilon_d(\gamma') - \epsilon_d(\gamma)}{1 + \epsilon_d(\gamma)}\right)\right| \leq 2\eta,$$

and finally:

$$\sup_{d\geq d_0}\sup_{\substack{\gamma,\gamma'\in[0,T]\\ |\gamma-\gamma'|\leq\eta}}\big|\phi_d(\gamma) - \phi_d(\gamma')\big| \leq \eta\left(1 + \frac{L}{\rho}\right).$$

Obviously this quantity goes to zero as $\eta \to 0$. For $d = 0, \ldots, d_0 - 1$, each ϕ_d is continuous on $[0,T]$, thus uniformly continuous, and:

$$\sup_{d\in\mathbb{N}}\sup_{\substack{\gamma,\gamma'\in[0,T]\\ |\gamma-\gamma'|\leq\eta}}\big|\phi_d(\gamma) - \phi_d(\gamma')\big| \leq \max\left[\max_{d\in[\![0,d_0-1]\!]}\sup_{\substack{\gamma,\gamma'\in[0,T]\\ |\gamma-\gamma'|\leq\eta}}\big|\phi_d(\gamma) - \phi_d(\gamma')\big|,\ \eta\left(1 + \frac{L}{\rho}\right)\right] \xrightarrow[\eta\to 0]{} 0,$$

which is the claim of Lemma C.6.

D.3. Proof of Lemma C.8

We set $\sigma_d = (d/m_d, d/m_d^*)$ which converges towards $\sigma = (\alpha^{-1}, \alpha^{*-1})$ as $d \to \infty$. Observe that:

$$\mathcal{H}_d(\xi)(\gamma) = \frac{1}{m_d}\mathrm{Tr}\big(S(\sigma_d, \xi, \gamma, Y_d)\big) \qquad \mathcal{H}(\xi)(\gamma) = \int S(\sigma, \xi, \gamma, x)\mathrm{d}\mu(x),$$

with:

$$S(\sigma, \xi, \gamma, x) = \frac{1}{4\sigma_1}\log\left(1 + 4\sigma_1\int_0^\gamma e^{-2\xi(s)}\mathrm{d}s + 4x\sigma_1\int_0^\gamma e^{-2\xi(s)}\big(e^{4\sigma_2 s} - 1\big)\mathrm{d}s\right),$$

and $\sigma = (\sigma_1, \sigma_2)$. Now, for $\xi \in B_c$ and $\gamma \in [0, T]$:

$$\begin{aligned}\big|\mathcal{H}_d(\xi)(\gamma) - \mathcal{H}(\xi)(\gamma)\big| \leq \frac{1}{m_d}\mathrm{Tr}\left|S(\sigma_d, \xi, \gamma, Y_d) - S(\sigma, \xi, \gamma, Y_d)\right| \\ + \left|\frac{1}{m_d}\mathrm{Tr}S(\sigma, \xi, \gamma, Y_d) - \int S(\sigma, \xi, \gamma, x)\mathrm{d}x\right|.\end{aligned} \tag{40}$$

We start by the first term. We have:

$$\frac{1}{m_d}\mathrm{Tr}\left|S(\sigma_d, \xi, \gamma, Y_d) - S(\sigma, \xi, \gamma, Y_d)\right| \leq \sup_{x\in[0,1]}\big|S(\sigma_d, \xi, \gamma, x) - S(\sigma, \xi, \gamma, x)\big|,$$

since the eigenvalues of Y_d are contained in $[0, 1]$. We take $\sigma_1, \tilde{\sigma}_1$ and suppose that they are lower bounded by some $\sigma_0 > 0$. Now, for $a, \tilde{a} > 0$:

$$\begin{aligned}\left|\frac{1}{4\sigma_1}\log\big(1 + 4\sigma_1 a\big) - \frac{1}{4\tilde{\sigma}_1}\log\big(1 + 4\tilde{\sigma}_1\tilde{a}\big)\right| &\leq \left|\frac{1}{4\sigma_1}\log\big(1 + 4\sigma_1 a\big) - \frac{1}{4\tilde{\sigma}_1}\log\big(1 + 4\tilde{\sigma}_1 a\big)\right| \\ &\quad + \left|\frac{1}{4\tilde{\sigma}_1}\log\big(1 + 4\tilde{\sigma}_1 a\big) - \frac{1}{4\tilde{\sigma}_1}\log\big(1 + 4\tilde{\sigma}_1\tilde{a}\big)\right| \\ &\leq \frac{2a}{\sigma_0}|\sigma_1 - \tilde{\sigma}_1| + |a - \tilde{a}|.\end{aligned}$$

We used that the map $u \mapsto \dfrac{1}{4u}\log(1 + 4ua)$ is Lipschitz with coefficient $\dfrac{2a}{u_0}$ for $u \geq u_0$ and the fact that $u \mapsto \log(1 + u)$ is $1-$Lipschitz. Applying this result with $a = \int_0^\gamma e^{-2\xi(s)}\mathrm{d}s + x\int_0^\gamma e^{-2\xi(s)}\big(e^{4\sigma_2 s} - 1\big)\mathrm{d}s$ (and the same for $\tilde{a}$ with $\tilde{\sigma}_2$):

$$\begin{aligned}\big|S(\sigma, \xi, \gamma, x) - S(\tilde{\sigma}, \xi, \gamma, x)\big| &\leq \frac{2a}{\sigma_0}|\sigma_1 - \tilde{\sigma}_1| + x\int_0^\gamma e^{-2\xi(s)}\big|e^{4s\sigma_2} - e^{4s\tilde{\sigma}_2}\big|\mathrm{d}s \\ &\leq \frac{2a}{\sigma_0}|\sigma_1 - \tilde{\sigma}_1| + 4x|\sigma_2 - \tilde{\sigma}_2|\int_0^\gamma s e^{-2\xi(s)}e^{4s\sigma_0^*}\mathrm{d}s,\end{aligned}$$

where we assumed that $\sigma_2, \tilde{\sigma}_2 \leq \sigma_0^*$. Now, using that $\gamma \in [0, T]$, $x \in [0, 1]$ and $\xi \in B_c$:

$$\big|S(\sigma, \xi, \gamma, x) - S(\tilde{\sigma}, \xi, \gamma, x)\big| \leq \frac{2}{\sigma_0}\left(Te^{-2c} + \frac{e^{-2c}}{4\sigma_0^*}e^{4\sigma_0^* T}\right)|\sigma_1 - \tilde{\sigma}_1| + \frac{Te^{-2c}}{\sigma_0^*}e^{4T\sigma_0^*}|\sigma_2 - \tilde{\sigma}_2|.$$

Now using this result with $\sigma_d = (d/m_d, d/m_d^*) \xrightarrow[d\to\infty]{} \sigma = (\alpha^{-1}, \alpha^{*-1})$, it is clear that d/m_d is bounded away from zero and d/m_d^* is lower bounded (at least for d large enough), which confirms the existence of σ_0, σ_0^*. Finally:

$$\sup_{\xi\in B_c}\sup_{\gamma\in[0,T]}\sup_{x\in[0,1]}\big|S(\sigma_d, \xi, \gamma, x) - S(\sigma, \xi, \gamma, x)\big| \xrightarrow[d\to\infty]{} 0,$$

which proves that the first term of equation (40) goes to zero uniformly in ξ and γ. We look at the second term. Using the bounds, for $\xi \in B_c$ and $\gamma \in [0, T]$:

$$1 + 4\sigma_1 \int_0^\gamma e^{-2\xi(s)} \mathrm{d}s \leq 1 + 4\sigma_1 e^{-2c} T \equiv A(\sigma),$$

$$4\sigma_1 \int_0^\gamma e^{-2\xi(s)} \big(e^{4\sigma_2 s} - 1\big) \mathrm{d}s \leq \frac{\sigma_1}{\sigma_2} e^{-2c} e^{4\sigma_2 T} \equiv B(\sigma),$$

then:

$$\sup_{\xi \in B_c} \sup_{\gamma \in [0,T]} \left| \frac{1}{m_d} \mathrm{Tr} S(\sigma, \xi, \gamma, Y_d) - \int S(\sigma, \xi, \gamma, x) \mathrm{d}x \right| \leq \frac{1}{4\sigma_1} \sup_{(a,b) \in K(\sigma)} \left| \frac{1}{m_d} \mathrm{Tr}\Big[l_{a,b}(Y_d) \Big] - \int l_{a,b}(x) \mathrm{d}\mu(x) \right|,$$

with $K(\sigma) = [1, A(\sigma)] \times [0, B(\sigma)]$ and $l_{a,b}(x) = \log(a + bx)$. Since the family $(l_{a,b})_{a,b \in K(\sigma)}$ is compact, we can apply the result of Lemma A.1 to obtain that the second term also goes to zero with probability one. Therefore, both terms of equation (40) converge uniformly to zero, which concludes the proof.

D.4. Proof of Lemma C.9

From equation (12), we have:

$$4\gamma = \alpha \log F_\phi(\gamma) + (\alpha + \alpha^* - 1)^+ \log J_\phi(\gamma) + \Theta\big(J_\phi(\gamma) - 1\big), \tag{41}$$

with:

$$F_\phi(\gamma) = 1 + \frac{4}{\alpha} \int_0^t e^{-2\phi(s)} \mathrm{d}s \qquad J_\phi(\gamma) = F_\phi(\gamma)^{-1} \left(1 + \frac{4}{\alpha} \int_0^t e^{-2\phi(s)} e^{4s/\alpha^*} \mathrm{d}s \right).$$

From the expression of Θ in Proposition 5.1:

$$\Theta(u) = \frac{1}{2\pi} \int_{r_-}^{r_+} \frac{\sqrt{(r_+ - x)(x - r_-)}}{x(1 - x)} \log(1 + ux) \mathrm{d}x,$$

it is easy to see that Θ is twice continuously differentiable on $]-1, \infty[$, and that $\Theta'(u) \geq 0$ for all $u > -1$. We now suppose that ϕ is a continuous solution of equation (41). Then, since $J_\phi(\gamma) \geq 1$, we can differentiate with respect to γ:

$$4 = \alpha \frac{\dot{F}_\phi(\gamma)}{F_\phi(\gamma)} + (\alpha + \alpha^* - 1)^+ \frac{\dot{J}_\phi(\gamma)}{J_\phi(\gamma)} + \dot{J}_\phi(\gamma) \Theta'\big(J_\phi(\gamma) - 1\big).$$

From the relationships:

$$\dot{F}_\phi(\gamma) = \frac{4}{\alpha} e^{-2\phi(\gamma)} \qquad \dot{J}_\phi(\gamma) = \frac{4}{\alpha} e^{-2\phi(\gamma)} F_\phi(\gamma)^{-1} \big(e^{4\gamma/\alpha^*} - J_\phi(\gamma)\big),$$

we finally obtain that $(F_\phi(\gamma), J_\phi(\gamma))$ is solution of the differential system:

$$\begin{pmatrix} \dot{F}(\gamma) \\ \dot{J}(\gamma) \end{pmatrix} = \frac{4}{\alpha + \Gamma(J(\gamma)) \big(e^{4\gamma/\alpha^*} - J(\gamma)\big)} \begin{pmatrix} F(\gamma) \\ e^{4\gamma/\alpha^*} - J(\gamma) \end{pmatrix}, \tag{42}$$

with initial condition $F(0) = J(0) = 1$ and:

$$\Gamma(J) = \frac{(\alpha + \alpha^* - 1)^+}{J} + \Theta'(J - 1).$$

Thus, if we manage to show that the solution of equation (42) is unique, this will imply the uniqueness of the continuous solution ϕ. Reciprocally, suppose that we have a couple $(F(\gamma), J(\gamma))$ solution of equation (42) with the initial condition, then it is easily shown that it is also solution of equation (41). Provided that $\dot{F}(\gamma) > 0$ for

all $\gamma \geq 0$, we will obtain the existence of a solution:

$$\phi(\gamma) = -\frac{1}{2}\log\left(\frac{\alpha \dot{F}(\gamma)}{4}\right).$$

The goal is now to study equation (42). We now set:

$$\Omega(F, J, \gamma) = \frac{4}{\alpha + \Gamma(J)\big(e^{4\gamma/\alpha^*} - J\big)} \begin{pmatrix} F \\ e^{4\gamma/\alpha^*} - J \end{pmatrix}.$$

Ω is well defined on the domain $D = \big\{(F, J, \gamma) \in \mathbb{R}\times]0, \infty[\times\mathbb{R}^+ \;\big|\; \alpha + \Gamma(J)\big(e^{4\gamma/\alpha^*} - J\big) \neq 0\big\}$, which is an open subset of $\mathbb{R}^2 \times \mathbb{R}^+$. Now, since Θ is twice continuously differentiable on $]-1, \infty[$, Ω is continuously differentiable on its domain. Let $u \in D$ and $\mathcal{V} \subset D$ a compact neighbourhood of u. Then by continuity:

$$\sup_{(F,J,\gamma)\in\mathcal{V}} \big\|\nabla\Omega(F, J, \gamma)\big\| < \infty.$$

Which proves that Ω is locally Lipschitz continuous on D. Thus, from Picard–Lindelöf theorem, the solution $F(\gamma), J(\gamma)$ of the differential system (42) with initial condition $F(0) = J(0) = 1$ are well defined and unique on a subset of the form $[0, \eta]$ for $\eta > 0$. We now let F_M, J_M be maximal solutions for this Cauchy-Lipschitz problem. Following the previous reasoning, they are at least defined and unique on $[0, \eta]$. We suppose by contradiction that they are defined up to $[0, T[$ with $T > 0$. From equation (42), we get that $\alpha+\Gamma(J_M(\gamma)(e^{4\gamma/\alpha^*} - J_M(\gamma)) > 0$ for $\gamma \in [0, T[$, since it is positive for $\gamma = 0$ and cannot cancel. Thus, $\dot{F}_M(\gamma)$ remains of the same sign as $F_M(\gamma)$. Since $F_M(0) = 1$, we get that $\dot{F}_M(\gamma) > 0$ and F_M is non-decreasing. Now, from the equation:

$$\dot{J}_M(\gamma) = \big(e^{4\gamma/\alpha^*} - J_M(\gamma)\big)\frac{\dot{F}_M(\gamma)}{F_M(\gamma)},$$

we get:

$$J_M(\gamma) = e^{4\gamma/\alpha^*} - \frac{4}{\alpha^* F_M(\gamma)}\int_0^\gamma F_M(s)e^{4s/\alpha^*}\mathrm{d}s < e^{4\gamma/\alpha^*},$$

for $\gamma \in [0, T[$. Thus, $\dot{J}_M(\gamma) > 0$ and J_M is non decreasing. Therefore, $\lim_{\gamma\to T^-} J_M(\gamma)$ exists and is bounded by e^{4T/α^*}. Since $J_M(\gamma)$ remains positive on $[0, T[$, this implies that:

$$\lim_{\gamma\to T^-}\Big[\alpha + \Gamma(J_M(\gamma))\big(e^{4\gamma/\alpha^*} - J_M(\gamma)\big)\Big] \geq \alpha.$$

Thus $(F_M(\gamma), J_M(\gamma), \gamma)$ remains inside D as $\gamma \to T^-$. Now, by maximality assumption, $J_M(\gamma)$ or $F_M(\gamma)$ should escape any compact as $\gamma \to T^-$. From the previous observation, this cannot be the case for $J_M(\gamma)$. Thus, we necessarily have $\lim_{\gamma\to T^-} F_M(\gamma) = +\infty$ (again this limit exists since F_M is non-decreasing on $[0, T[$). From equation (42):

$$F_M(\gamma) = \exp\left(4\int_0^\gamma \frac{\mathrm{d}s}{\alpha + \Gamma(J_M(s))\big(e^{4s/\alpha^*} - J_M(s)\big)}\right) \leq e^{4\gamma/\alpha},$$

and a contradiction. Thus necessarily $T = \infty$ and the system (42) admits a unique solution defined on all $\mathbb{R}^+$. From the equivalence between equations (41) and (42), this concludes the proof.

D.5. Proof of Lemma C.11

As proven in Proposition 5.1, ϕ_d uniformly converges on $[0, T]$ towards a function ϕ. Thus we define:

$$F(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}\mathrm{d}s \qquad G(\gamma) = 1 + \frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}\mathrm{d}s \qquad J(\gamma) = \frac{G(\gamma)}{F(\gamma)}. \tag{43}$$

It is easily seen that $F_d \xrightarrow[d\to\infty]{} F$ and $G_d \xrightarrow[d\to\infty]{} G$ uniformly on $[0,T]$. Therefore, since $F_d(\gamma), F(\gamma) \geq 1$:

$$\big|J_d(\gamma) - J(\gamma)\big| \leq \big|F(\gamma)\big|\big|G_d(\gamma) - G(\gamma)\big| + \big|G(\gamma)\big|\big|F_d(\gamma) - F(\gamma)\big|.$$

This proves that $J_d \xrightarrow[d\to\infty]{} J$ uniformly on $[0,T]$. Now, for the numerator, we set $U(x,J) = \dfrac{x}{1+(J-1)x}$, so that:

$$\left|\frac{1}{m_d}\mathrm{Tr}\Big(U(Y_d, J_d(\gamma))\Big) - \int U(x, J(\gamma))\mathrm{d}\mu(x)\right| \leq \frac{1}{m_d}\mathrm{Tr}\Big|U(Y_d, J_d(\gamma)) - U(Y_d, J(\gamma))\Big| + \left|\frac{1}{m_d}\mathrm{Tr}\Big(U(Y_d, J(\gamma))\Big) - \int U(x, J(\gamma))\mathrm{d}\mu(x)\right|.$$

Starting with the first term:

$$\frac{1}{m_d}\mathrm{Tr}\Big|U(Y_d, J_d(\gamma)) - U(Y_d, J(\gamma))\Big| \leq \sup_{x\in[0,1]}\Big|U(x, J_d(\gamma)) - U(x, J(\gamma))\Big| \leq \big|J_d(\gamma) - J(\gamma)\big|.$$

Since it is easily shown that $J \mapsto U(x,J)$ is Lipschitz on $[1,\infty[$ with constant x^2 (note that we always have $J_d(\gamma), J(\gamma) \in [1,\infty[$). Thus the first term goes uniformly to zero on $[0,T]$. For the second, J is continuous on $[0,T]$, thus it is bounded by some constant $A \geq 1$. Therefore:

$$\sup_{\gamma\in[0,T]}\left|\frac{1}{m_d}\mathrm{Tr}\Big(U(Y_d, J(\gamma))\Big) - \int U(x, J(\gamma))\mathrm{d}\mu(x)\right| \leq \sup_{a\in[1,A]}\left|\frac{1}{m_d}\mathrm{Tr}\Big(U(Y_d, a)\Big) - \int U(x,a)\mathrm{d}\mu(x)\right|,$$

which goes to zero following Lemma A.1. Therefore, the first convergence of Lemma C.11 is proven. The same can be done for the second. The same thing can be done for the denominator. Introducing:

$$V(x,J,e) = \left(\frac{1+(e-1)x}{1+(J-1)x}\right)^2,$$

as well as $e_d(\gamma) = e^{4\gamma d/m_d^*}$ and $e(\gamma) = e^{4\gamma/\alpha^*}$, we have:

$$\left|\frac{1}{m_d}\mathrm{Tr}\Big(V(Y_d, J_d(\gamma), e_d(\gamma))\Big) - \int V(x, J(\gamma), e(\gamma))\mathrm{d}\mu(x)\right| \leq \frac{1}{m_d}\mathrm{Tr}\Big|V(Y_d, J_d(\gamma), e_d(\gamma)) - V(Y_d, J(\gamma), e(\gamma))\Big| + \left|\frac{1}{m_d}\mathrm{Tr}\Big(V(Y_d, J(\gamma), e(\gamma))\Big) - \int V(x, J(\gamma), e(\gamma))\mathrm{d}\mu(x)\right|. \tag{44}$$

Now:

$$\frac{1}{m_d}\mathrm{Tr}\Big|V(Y_d, J_d(\gamma), e_d(\gamma)) - V(Y_d, J(\gamma), e(\gamma))\Big| \leq \sup_{x\in[0,1]}\Big|V(x, J_d(\gamma), e_d(\gamma)) - V(x, J(\gamma), e(\gamma))\Big|.$$

Moreover, for $e, \tilde{e} \in [1,E]$ and $J, \tilde{J} \geq 1$:

$$\begin{aligned}\Big|V(x,J,e) - V(x,\tilde{J},\tilde{e})\Big| &\leq \Big|V(x,J,e) - V(x,\tilde{J},e)\Big| + \Big|V(x,\tilde{J},e) - V(x,\tilde{J},\tilde{e})\Big| \\ &\leq 2E^2|J-\tilde{J}| + 2E|e-\tilde{e}|.\end{aligned}$$

Applying this to our case, we have $e_d(\gamma) = e^{4\gamma d/m_d^*} \leq e^{4Td/m_d^*}$ stays bounded since $d/m_d^* \xrightarrow[d\to\infty]{} \alpha^{*-1}$. Thus, again, the first term of equation (44) goes to zero uniformly on $[0,T]$. For the second, it is clear that since J and e are continuous on $[0,T]$, they both stay in a compact of the form $K = [1,A]\times[1,E]$. Therefore, by

Lemma A.1:

$$\sup_{\gamma\in[0,T]}\left|\frac{1}{m_d}\mathrm{Tr}\Big(V(Y_d,J(\gamma),e(\gamma))\Big)-\int V(x,J(\gamma),e(\gamma))\mathrm{d}\mu(x)\right|\leq\sup_{(J,e)\in K}\left|\frac{1}{m_d}\mathrm{Tr}\Big(V(Y_d,J,e)\Big)-\int V(x,J,e)\mathrm{d}\mu(x)\right|,$$

which goes to zero. As a conclusion, the second convergence of Lemma C.11 is proven. Finally, for the overlap, write from equation (30):

$$\chi_d(\gamma)=\sqrt{\frac{m_d}{m_d^*}}e^{4\gamma d/m_d^*}\frac{A_d(\gamma)}{\sqrt{B_d(\gamma)}},$$

where $A_d(\gamma)\xrightarrow[d\to\infty]{}A(\gamma)$ and $B_d(\gamma)\xrightarrow[d\to\infty]{}B(\gamma)$ are the quantity displayed in the lemma, which converge uniformly from the previous reasoning. We write:

$$\begin{aligned}\big|\chi_d(\gamma)-\chi(\gamma)\big|\leq\frac{A_d(\gamma)}{\sqrt{B_d(\gamma)}}\left|\sqrt{\frac{m_d}{m_d^*}}e^{4\gamma d/m_d^*}-\sqrt{\frac{\alpha}{\alpha^*}}e^{4\gamma/\alpha^*}\right|+\sqrt{\frac{\alpha}{\alpha^*}}e^{4\gamma/\alpha^*}\frac{1}{\sqrt{B_d(\gamma)}}\big|A_d(\gamma)-A(\gamma)\big|\\+\sqrt{\frac{\alpha}{\alpha^*}}e^{4\gamma/\alpha^*}A(\gamma)\left|\frac{1}{\sqrt{B_d(\gamma)}}-\frac{1}{\sqrt{B(\gamma)}}\right|.\end{aligned}$$

Since $B_d(\gamma)$ uniformly converges on $[0,T]$ towards $B(\gamma)$ which is strictly positive, thus it is uniformly bounded away from zero for d large enough. Thus, there are constant $k_1,k_2,k_3>0$ such that for some $d\geq d_0$:

$$\begin{aligned}\sup_{\gamma\in[0,T]}\big|\chi_d(\gamma)-\chi(\gamma)\big|\leq k_1\sup_{\gamma\in[0,T]}\left|\sqrt{\frac{m_d}{m_d^*}}e^{4\gamma d/m_d^*}-\sqrt{\frac{\alpha}{\alpha^*}}e^{4\gamma/\alpha^*}\right|+k_2\sup_{\gamma\in[0,T]}\big|A_d(\gamma)-A(\gamma)\big|\\+k_3\sup_{\gamma\in[0,T]}\big|B_d(\gamma)-B(\gamma)\big|,\end{aligned}$$

since the map $x\mapsto x^{-1/2}$ is Lipschitz on the intervals of the form $[b,\infty[$. As shown before $A_d(\gamma)\longrightarrow A(\gamma)$ and $B_d(\gamma)\longrightarrow B(\gamma)$ uniformly as $d\to\infty$, so that the two last terms go to zero. Since m_d^*/d and m_d/d respectively converge to α^*,α, it is easily checked that the first term also goes uniformly to zero. Finally, $\chi_d\xrightarrow[d\to\infty]{}\chi$ uniformly on $[0,T]$, which concludes the proof.

D.6. Proof of Lemma C.12

From the implicit equation (12) solved by ϕ, we obtain:

$$4\gamma=\min\big(\alpha,1-\alpha^*\big)\log F(\gamma)+\big(\alpha+\alpha^*-1\big)^+\log G(\gamma)+\Theta\big(J(\gamma)-1\big),\tag{45}$$

where F,G,J are defined in equation (35) (and depend on ϕ), and:

$$\Theta(u)=\frac{1}{2\pi}\int_{r_1}^{r_2}\frac{\sqrt{(r_2-x)(x-r_1)}}{x(1-x)}\log(1+ux)\mathrm{d}x.$$

Using the definition of μ (and the fact it has unit total mass), it can be shown that there exists $\Omega(\alpha,\alpha^*)$:

$$\Theta(u)\underset{u\to\infty}{=}\frac{1-|\alpha-\alpha^*|-|\alpha+\alpha^*-1|}{2}\log u+\Omega(\alpha,\alpha^*)+O\left(\frac{1}{u}\right).\tag{46}$$

The first step is to show that $F(\gamma),G(\gamma)$ and $J(\gamma)$ go to infinity as $\gamma\to\infty$. We recall that:

$$F(\gamma)=1+\frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}\mathrm{d}s\qquad G(\gamma)=1+\frac{4}{\alpha}\int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}\mathrm{d}s\qquad J(\gamma)=\frac{G(\gamma)}{F(\gamma)}.$$

We start by F and G. As they are both non-decreasing, either they converge or go to infinity. Obviously they cannot both converge as $\gamma\to\infty$, otherwise the right hand side of equation (45) would stay finite as $\gamma\to\infty$. Since $G(\gamma)\geq F(\gamma)$, at least $G(\gamma)\xrightarrow[\gamma\to\infty]{}\infty$. Suppose that $F(\gamma)\xrightarrow[\gamma\to\infty]{}F$, then we would have from equations

(45) and (46):

$$\left(\left(\alpha+\alpha^*-1\right)^+ + \frac{1-|\alpha-\alpha^*|-|\alpha+\alpha^*-1|}{2}\right)\log G(\gamma) \underset{\gamma\to\infty}{=} 4\gamma + O(1),$$

which writes:

$$G(\gamma) = e^{O(1)}\exp\left(\frac{4\gamma}{\min(\alpha,\alpha^*)}\right).$$

Therefore, integrating by parts, one has:

$$F(\gamma) = G(\gamma)e^{-4\gamma/\alpha^*} + \frac{4}{\alpha^*}\int_0^\gamma G(s)e^{-4s/\alpha^*}\mathrm{d}s \geq \frac{4}{\alpha^*}\int_0^\gamma G(s)e^{-4s/\alpha^*}\mathrm{d}s \xrightarrow[\gamma\to\infty]{} \infty. \tag{47}$$

This proves that necessarily $F(\gamma)$ also goes to infinity as $\gamma\to\infty$. We now finally show that also $J(\gamma)\longrightarrow\infty$. Differentiating:

$$\dot{J}(\gamma) = \frac{1}{G(\gamma)^2}\big(\dot{G}(\gamma)F(\gamma) - \dot{F}(\gamma)G(\gamma)\big) = \frac{\dot{G}(\gamma)}{G(\gamma)^2}\big(F(\gamma) - G(\gamma)e^{-4\gamma/\alpha^*}\big) \geq 0,$$

thus J is non-decreasing. Suppose that $J(\gamma)$ converges towards a finite value as $\gamma\to\infty$. Thus, equation (45) gives:

$$4\gamma = \min\left(\alpha, 1-\alpha^*\right)\log F(\gamma) + \left(\alpha+\alpha^*-1\right)^+\log G(\gamma) + O(1).$$

Splitting the case $\alpha+\alpha^*\leq 1$ and $\alpha+\alpha^*\geq 1$, we obtain in both case the existence of $C\in\mathbb{R}$ (depending on α,α^*) such that:

$$4\gamma \underset{\gamma\to\infty}{=} \alpha\log F(\gamma) + C + o(1).$$

From the following equality, which can be obtained by integrating by parts the relationship $\dot{G}(\gamma) = \dot{F}(\gamma)e^{4\gamma/\alpha^*}$:

$$J(\gamma) = e^{4\gamma/\alpha^*} - \frac{4}{\alpha^* F(\gamma)}\int_0^\gamma F(s)e^{4s/\alpha^*}\mathrm{d}s,$$

we have:

$$J(\gamma) \underset{\gamma\to\infty}{\sim} \frac{\alpha^*}{\alpha+\alpha^*}e^{4\gamma/\alpha^*}.$$

Therefore, we necessarily have $J(\gamma)\longrightarrow\infty$. We now prove the asymptotics on J. Using equations (45) and (46), we obtain as $\gamma\to\infty$:

$$\begin{aligned}4\gamma &\underset{\gamma\to\infty}{=} \min\left(\alpha,1-\alpha^*\right)\log F(\gamma) + \left(\alpha+\alpha^*-1\right)^+\log G(\gamma)\\ &\qquad + \frac{1-|\alpha-\alpha^*|-|\alpha+\alpha^*-1|}{2}\log J(\gamma) + \Omega(\alpha,\alpha^*) + o(1)\\ &= (\alpha-\alpha^*)^+\log F(\gamma) + \min(\alpha,\alpha^*)\log G(\gamma) + \Omega(\alpha,\alpha^*) + o(1).\end{aligned}$$

Thus, setting:

$$\lambda^* = \frac{1}{\alpha^*} \qquad \lambda_m = \frac{1}{\min(\alpha,\alpha^*)} \qquad \nu = \frac{(\alpha-\alpha^*)^+}{\min(\alpha,\alpha^*)} \qquad \zeta = \exp\left(-\frac{\Omega(\alpha,\alpha^*)}{\min(\alpha,\alpha^*)}\right),$$

we obtain that:

$$G(\gamma) = \zeta F(\gamma)^{-\nu}e^{4\lambda_m\gamma}\omega(\gamma), \tag{48}$$

where $\omega(\gamma)\xrightarrow[\gamma\to\infty]{}1$. Using the relation between F and G in equation (47):

$$F(\gamma) = \zeta F(\gamma)^{-\nu}e^{4(\lambda_m-\lambda^*)\gamma}\omega(\gamma) + 4\lambda^*\zeta\int_0^\gamma F(s)^{-\nu}e^{4(\lambda_m-\lambda^*)s}\omega(s)\mathrm{d}s. \tag{49}$$

We start by the case where $\alpha \leq \alpha^*$, so that $\nu = 0$. Therefore, using the two previous equations as well as Lemma B.3:

$$G(\gamma) \underset{\gamma\to\infty}{\sim} \zeta e^{4\gamma/\alpha} \qquad F(\gamma) \underset{\gamma\to\infty}{\sim} \begin{cases} \dfrac{4\zeta\gamma}{\alpha^*} & \text{for } \alpha = \alpha^* \\ \zeta\dfrac{\alpha^*}{\alpha^*-\alpha}\exp\left(4\gamma\dfrac{\alpha^*-\alpha}{\alpha\alpha^*}\right) & \text{for } \alpha < \alpha^*. \end{cases}$$

Therefore:

$$J(\gamma) \underset{\gamma\to\infty}{\sim} \begin{cases} \dfrac{\alpha^*}{4\gamma}e^{4\gamma/\alpha^*} & \text{for } \alpha = \alpha^* \\ \left(1-\dfrac{\alpha}{\alpha^*}\right)e^{4\gamma/\alpha^*} & \text{for } \alpha < \alpha^*. \end{cases}$$

For $\alpha > \alpha^*$, we have that $\lambda^* = \lambda_m$ and $\nu = \alpha/\alpha^* - 1 > 0$, thus $F(\gamma)^{-\nu} \xrightarrow[\gamma\to\infty]{} 0$. Thus, rewriting equation (49):

$$F(\gamma) = \underbrace{\zeta F(\gamma)^{-\nu}\omega(\gamma)}_{\equiv\,\epsilon(\gamma)} + 4\lambda^*\zeta\int_0^\gamma F(s)^{-\nu}\omega(s)\mathrm{d}s.$$

Since $\epsilon(\gamma) \xrightarrow[\gamma\to\infty]{} 0$, we introduce $\epsilon > 0$ and γ_0 such that $|\epsilon(\gamma)| \leq \epsilon$ as soon as $\gamma \geq \gamma_0$. With $K(\gamma) = \int_0^\gamma F(s)^{-\nu}\omega(s)\mathrm{d}s$, we obtain by integrating:

$$-\epsilon + \left[\left(\frac{4\zeta}{\alpha^*}K(\gamma_0)+\epsilon\right)^{\alpha/\alpha^*} + \frac{4\zeta\alpha}{\alpha^{*2}}\int_0^\gamma \omega(s)\mathrm{d}s\right]^{\alpha^*/\alpha} \leq \frac{4\zeta}{\alpha^*}K(\gamma) \leq \epsilon + \left[\left(\frac{4\zeta}{\alpha^*}K(\gamma_0)-\epsilon\right)^{\alpha/\alpha^*} + \frac{4\zeta\alpha}{\alpha^{*2}}\int_0^\gamma \omega(s)\mathrm{d}s\right]^{\alpha^*/\alpha}.$$

Therefore:

$$F(\gamma) \underset{\gamma\to\infty}{\sim} \left(\frac{4\zeta\alpha}{\alpha^{*2}}\gamma\right)^{\alpha^*/\alpha}.$$

Using equation (48), we are able to determine the behaviour of G, and the one of J:

$$G(\gamma) \underset{\gamma\to\infty}{\sim} \zeta\left(\frac{4\zeta\alpha}{\alpha^{*2}}\gamma\right)^{\alpha^*/\alpha-1} e^{4\gamma/\alpha^*} \qquad J(\gamma) \underset{\gamma\to\infty}{\sim} \frac{\alpha^{*2}}{4\alpha\gamma}e^{4\gamma/\alpha^*}.$$

D.7. Proof of Lemma C.13

From the definition of μ in equation (10) and those of $A(\gamma)$, $B(\gamma)$ in Lemma C.12, we have:

$$\begin{aligned} A(\gamma) &= \frac{1}{\alpha}(\alpha+\alpha^*-1)^+\frac{1}{J(\gamma)} + \int_{r_-}^{r_+}\frac{x}{1+(J(\gamma)-1)x}w(x)\mathrm{d}x \\ B(\gamma) &= \left(1-\frac{\alpha^*}{\alpha}\right)^+ + \frac{1}{\alpha}(\alpha+\alpha^*-1)^+\frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2} + \int_{r_-}^{r_+}\left(\frac{1+(e^{4\gamma/\alpha^*}-1)x}{1+(J(\gamma)-1)x}\right)^2 w(x)\mathrm{d}x, \end{aligned} \tag{50}$$

with:

$$w(x) = \frac{1}{2\pi\alpha}\frac{\sqrt{(r_+-x)(x-r_-)}}{x(1-x)}$$

$$r_\pm = \left(\sqrt{\alpha(1-\alpha^*)} \pm \sqrt{\alpha^*(1-\alpha)}\right)^2.$$

The challenge is to compute the asymptotics of the integrals. For the first one, we compute:

$$\left|\frac{1}{J(\gamma)}\int_{r_-}^{r_+} w(x)\mathrm{d}x - \int_{r_-}^{r_+}\frac{x}{1+(J(\gamma)-1)x}w(x)\mathrm{d}x\right| = \frac{1}{J(\gamma)^2}\int_{r_-}^{r_+}\frac{1-x}{x+(1-x)J(\gamma)^{-1}}w(x)\mathrm{d}x. \tag{51}$$

We start by supposing that $r_- > 0$. Since $\alpha, \alpha^* > 0$, this corresponds to the case $\alpha \neq \alpha^*$. Then:

$$\int_{r_-}^{r_+} \frac{1-x}{x+(1-x)J(\gamma)^{-1}} w(x)\mathrm{d}x \leq \int_{r_-}^{r_+} \frac{1-x}{x} w(x)\mathrm{d}x < \infty,$$

so that in this case:

$$\int_{r_-}^{r_+} \frac{x}{1+(J(\gamma)-1)x} w(x)\mathrm{d}x \underset{\gamma\to\infty}{=} \frac{1}{J(\gamma)} \int_{r_-}^{r_+} w(x)\mathrm{d}x + O\left(\frac{1}{J(\gamma)^2}\right).$$

The integral $\int_{r_-}^{r_+} w(x)\mathrm{d}x$ can be computed using the fact that μ has unit mass. Thus:

$$A(\gamma) \underset{\gamma\to\infty}{=} \frac{\min(\alpha,\alpha^*)}{\alpha J(\gamma)} + O\left(\frac{1}{J(\gamma)^2}\right),$$

whenever $r_- = 0$ this is not possible anymore. In this case, the right integral of equation (51) writes:

$$\frac{1}{2\pi\alpha J(\gamma)} \int_0^{r_+} \sqrt{\frac{r_+ - x}{x}} \frac{\mathrm{d}x}{xJ(\gamma)+1-x}.$$

We change variables and let $x = r_+ \dfrac{t^2}{1+t^2}$. Decomposing in partial fractions:

$$\begin{aligned}
\frac{1}{2\pi} \int_0^{r_+} \sqrt{\frac{r_+ - x}{x}} \frac{\mathrm{d}x}{xJ(\gamma)+1-x} &= \frac{r_+}{\pi} \int_0^{\infty} \frac{\mathrm{d}t}{(1+t^2)\big(1+(1-r_+ + r_+J(\gamma))t^2\big)} \\
&= \frac{1}{\pi}\left(-\frac{1}{J(\gamma)-1}\int_0^\infty \frac{\mathrm{d}t}{1+t^2} + \frac{1+r_+(J(\gamma)-1)}{J(\gamma)-1}\int_0^\infty \frac{\mathrm{d}t}{1+(1-r_++r_+J(\gamma))t^2}\right) \\
&= -\frac{1}{2}\frac{1}{J(\gamma)-1} + \frac{1}{2}\frac{\sqrt{1+r_+(J(\gamma)-1)}}{J(\gamma)-1} \\
&= O\left(J(\gamma)^{-1/2}\right),
\end{aligned}$$

so that we get:

$$\int_{r_-}^{r_+} \frac{x}{1+(J(\gamma)-1)x} w(x)\mathrm{d}x \underset{\gamma\to\infty}{=} \frac{1}{J(\gamma)} \int_{r_-}^{r_+} w(x)\mathrm{d}x + O\left(\frac{1}{J(\gamma)^{3/2}}\right),$$

which proves the result for $A(\gamma)$ using equation (50). Now for $B(\gamma)$:

$$\begin{aligned}
\int_{r_-}^{r_+} \left(\frac{1+(e^{4\gamma/\alpha^*}-1)x}{1+(J(\gamma)-1)x}\right)^2 w(x)\mathrm{d}x - \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\int_{r_-}^{r_+} w(x)\mathrm{d}x = \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\Bigg[&\left(e^{-4\gamma/\alpha^*} - J(\gamma)^{-1}\right) I_1(\gamma) \\
&+ \left(e^{-8\gamma/\alpha^*} - J(\gamma)^{-2}\right) I_2(\gamma)\Bigg],
\end{aligned} \tag{52}$$

with:

$$I_1(\gamma) = \int_{r_-}^{r_+} \frac{2x(1-x)}{\big(x+(1-x)J(\gamma)^{-1}\big)^2} w(x)\mathrm{d}x \qquad I_2(\gamma) = \int_{r_-}^{r_+} \frac{(1-x)^2}{\big(x+(1-x)J(\gamma)^{-1}\big)^2} w(x)\mathrm{d}x.$$

Clearly the last term is negligible with respect to the others. Again, for $r_- > 0$, i.e., $\alpha \neq \alpha^*$:

$$I_1(\gamma) \leq \int_{r_-}^{r_+} \frac{2(1-x)}{x} w(x)\mathrm{d}x < \infty \qquad I_2(\gamma) \leq \int_{r_-}^{r_+} \frac{(1-x)^2}{x^2} w(x)\mathrm{d}x < \infty,$$

so that:

$$\int_{r_-}^{r_+} \left(\frac{1+(e^{4\gamma/\alpha^*}-1)x}{1+(J(\gamma)-1)x}\right)^2 w(x)\mathrm{d}x = \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\left(\int_{r_-}^{r_+} w(x)\mathrm{d}x + O\big(e^{-4\gamma/\alpha^*}\big) + O\big(J(\gamma)^{-1}\big)\right).$$

From Lemma C.12, we have $e^{-4\gamma/\alpha^*} = O(J(\gamma)^{-1})$, which proves the first result. If $\alpha = \alpha^*$ thus $r_- = 0$, we apply the same trick as before. We have:

$$\begin{aligned} I_1(\gamma) &= \frac{J(\gamma)^2}{\pi\alpha}\int_0^{r_+} \sqrt{x(r_+-x)}\frac{\mathrm{d}x}{\big(1-x+xJ(\gamma)\big)^2} \\ &= \frac{2r_+^2 J(\gamma)^2}{\pi\alpha}\int_0^{\infty} \frac{t^2\mathrm{d}t}{(1+t^2)\big(1+(1-r_++r_+J(\gamma))t^2\big)^2} \\ &\underset{\gamma\to\infty}{\sim} \frac{r_+^2}{2\alpha}J(\gamma)^{1/2}. \end{aligned}$$

Since $2I_2(\gamma) \leq I_1(\gamma)$, the last term in equation (52) is negligible and from the expression of r_+:

$$\int_{r_-}^{r_+} \left(\frac{1+(e^{4\gamma/\alpha^*}-1)x}{1+(J(\gamma)-1)x}\right)^2 w(x)\mathrm{d}x = \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2}\left(\int_{r_-}^{r_+} w(x)\mathrm{d}x + O\left(e^{-4\gamma/\alpha^*}J(\gamma)^{1/2}\right) + O\left(J(\gamma)^{-1/2}\right)\right),$$

from Lemma C.12, $e^{-4\gamma/\alpha^*}J(\gamma)^{1/2} = O\left(J(\gamma)^{-1/2}\right)$, which proves the result using equation (50).

E. Numerical Experiments

All numerical experiments were carried on a professional laptop equipped with a NVIDIA GeForce GTX 1650. The code is written in Python and uses Pytorch (Paszke et al., 2019) to run the gradient descent algorithm on GPU (for finite dimensional simulations). As for the numerical integration of the high-dimensional equations, a simple Euler method was used to approximate integrals and differential equations.

Training details.

- For gradient descent (Figures 2, 3, 5), we used a stepsize $\eta = 10^{-2}$ and initialized the weights W^0 from a Gaussian distribution (Figures 3, 5) or orthonormally (Figure 2), i.e $W^0 \overset{\text{distrib}}{=} U(U^TU)^{-1/2}$ with U having i.i.d Gaussian entries.
- To simulate $\phi(\gamma)$ (Figures 2, 3) in the high-dimensional limit, we numerically solved equation (12) by interpreting it as a differential equation on the vector $(F_\phi(\gamma), G_\phi(\gamma))$, and using a stepsize $\eta = 2\times 10^{-5}$.
- The asymptotic behaviour of the overlap (Figure 6) was simulated from equation (34) using a simple discretization method and an adapted 1D grid (which helped capturing the high variations of the integration measure at the edges) to compute the integrals (with 9×10^4 integration points) in the case $\alpha \neq \alpha^*$, and an analytic formula for those integrals in the special case $\alpha = \alpha^* = 1/2$, allowing for a reduction of the numerical error.