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

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 (Safran and Shamir, 2018; Choromanska et al., 2015; Christof and Kowalczuk, 2023; Baity-Jesi et al., 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. Unlike previous related works, we do not constraint the number of neurons of the teacher and student networks. 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 than the teacher \((m \geq m^*)\), this global optimum corresponds to a perfect recovery of the teacher network. In the overparameterized case, i.e., \(m \geq \min(m^*, d)\), 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 linearly grow with the 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, prove 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 \(\|A\|_F = \left[\text{Tr}(AA^T)\right]^{1/2}\) its Frobenius norm. We also say that \(A\) is of full rank if \(\text{rank}(A) = \min(m, d)\). We respectively denote \(O_d(\mathbb{R})\), \(S_d(\mathbb{R})\), \(S_d^+(\mathbb{R})\), \(S_d^{++}(\mathbb{R})\) the spaces of \(d \times d\) orthogonal, symmetric, positive semi-definite and positive definite matrices. If \((E, \langle \cdot, \cdot \rangle)\) is a Euclidean space and \(\Phi: E \to \mathbb{R}\) is twice continuously differentiable, we denote \(\nabla \Phi(x) \in E\) its gradient and \(d^2 \Phi_x\) its Hessian at \(x\), which is a linear self-adjoint map on \(E\). We say that \(x \in E\) is a critical point of \(\Phi\) if \(\nabla \Phi(x) = 0\), and a local minimizer if in addition, \(\langle d^2 \Phi_x(h), h \rangle \geq 0\) 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:

\[
    u^{*}(x) = \frac{1}{m^{*}} \sum_{j=1}^{m^{*}} (w_{j}^{*} \cdot x)^{2} = \text{Tr}(xx^{T}W^{*}W^{*T}),
\]

\[
    u(x) = \frac{1}{m} \sum_{j=1}^{m} (w_{j} \cdot x)^{2} = \text{Tr}(xx^{T}WW^{T}),
\]

where \( m \) and \( m^{*} \) are the number of neurons of the student and the teacher networks, \((w_{j})_{1 \leq j \leq m} \in (\mathbb{R}^{d})^{m}\) and \((w_{j}^{*})_{1 \leq j \leq m^{*}} \in (\mathbb{R}^{d})^{m^{*}}\) are the corresponding weights, and \( W, W^{*} \) the associated renormalized matrices:

\[
    W^{*} = \frac{1}{\sqrt{m^{*}}} \left( w_{1}^{*} \mid \ldots \mid w_{m^{*}}^{*} \right) \in \mathbb{R}^{d \times m^{*}},
\]

\[
    W = \frac{1}{\sqrt{m}} \left( w_{1} \mid \ldots \mid w_{m} \right) \in \mathbb{R}^{d \times m}.
\]

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 [1]), 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:

\[
    \mathcal{L}(W) = \frac{1}{4} \mathbb{E}_{x} \left[ \text{Tr} \left( xx^{T} (W^{*}W^{*T} - WW^{T}) \right)^{2} \right] = \frac{1}{2} \left\| \Delta W \right\|^{2} + \frac{1}{4} \left[ \text{Tr}(\Delta W) \right]^{2},
\]

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.

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 [Gamarnik et al. 2019] who show that in the general case, an extra term appears in the loss, proportional to \( \kappa_{4} \text{diag}(\Delta W) \right\|^{2} \), where \( \kappa_{4} \) is the fourth cumulant of the distribution from which the coefficients of \( x \) are drawn (which is zero in the Gaussian case), and \( \text{diag}(\Delta W) \) is the vector of the diagonal elements of \( \Delta W \). This leads to a different flow which is not covered by our results.

Optimization on \( W \) is performed using the gradient flow, which writes:

\[
    \dot{W} = -\nabla \mathcal{L}(W) = -2\Delta W - \text{Tr}(\Delta W)W.
\]

Note that, due to the expression of \( \mathcal{L} \), we do not expect to exactly retrieve through optimization the teacher vectors \( w_{1}^{*}, \ldots, w_{m^{*}}^{*} \), 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 full rank and since \( W \in \mathbb{R}^{d \times m^{*}} \), this is only possible for \( m \geq \min(m^{*}, d) \). Only in this case, one can hope for a convergence of the flow towards optimality.

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 will assume that the teacher matrix \( W^{*} \) is of rank \( r^{*} \equiv \min(m^{*}, d) \) and denote \( \mu_{1} \geq \cdots \geq \mu_{r^{*}} > 0 \) the non-zero eigenvalues of \( Z^{*} = W^{*}W^{*T} \), and \( v_{1}^{*}, \ldots, v_{r^{*}}^{*} \in \mathbb{R}^{d} \) the associated orthonormal eigenvectors, so that:

\[
    Z^{*} = \sum_{k=1}^{r^{*}} \mu_{k} v_{k}^{*} (v_{k}^{*})^{T}.
\]

### 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 \( \mathcal{L}^{10}(W) = \frac{1}{2} \left\| \Delta W \right\|^{2} \) (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 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 $\psi$ 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 $t \geq 0$:

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

with:

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

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

$$\dot{\psi}(t) = \frac{1}{4} \text{Tr} \log \left( I_m + 4 \int_0^t M^*(s)^T M^*(s) ds \right),$$

where the log is understood as the spectral map on $\mathbb{S}^{m^2}_+(\mathbb{R})$.

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

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

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

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

with $U(t) = I_m + 4 \int_0^t M^*(s)^T M^*(s) ds$. Since the map $X \mapsto \text{Tr} \log X$ has gradient $X^{-1}$ on $\mathbb{S}^{m^2}_+(\mathbb{R})$, this leads to the result by integrating the previous equation. \qed

This proposition gives an implicit formula for the solution of the flow $W(t)$, but in terms of the matrix $Z(t) = W(t)W(t)^T$. Regarding the solution $W(t)$, it can be shown that there exists a continuous map $Q : \mathbb{R}^+ \rightarrow O_m(\mathbb{R})$ such that:

$$W(t) = M^*(t) \left( I_m + 4 \int_0^t M^*(s)^T M^*(s) ds \right)^{-1/2} Q(t).$$

However, as mentioned in Section 2.2, the signal recovery condition only involves $Z(t)$. Therefore, studying the behaviour of $W(t)$ in itself is not needed.

The first consequence of formula (5) is that if $W^0$ is of full rank, 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., $Z(t) \rightarrow Z^*$ as $t \to \infty$) whenever $m < \min(m^*, d)$.

Equation (6) defines an implicit equation on $\psi$. It can be interpreted as a first order non-linear differential equation on the matrix $U(t)$ (defined in the proof of the proposition), which allows to show that the solution $\psi$ of equation (6) exists and is uniquely defined.

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 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 $m \geq r^*$: $W_\infty W_\infty^T = Z^*$,
- if $m < r^*$ 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 $\tau$ is defined as:

$$\tau = \frac{1}{m + 2} \sum_{k=m+1}^{r^*} \mu_k. \quad (7)$$

This result ensures a strong recovery of the signal by the gradient flow when $m \geq r^* = \min(m^*, d)$, which is the minimal overparameterization one needs (remark that when $m \geq d$ the flow always recovers the teachers). Indeed, as mentioned earlier, perfect recovery is impossible when $m < r^*$, due to rank constraints. Instead we only recover the eigenspace of the teacher associated with the $m$ largest eigenvalues. The corresponding eigenvalues for $Z$ are in addition shifted by a constant value (see equation (7)) with respect to the ones of $Z^*$.

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:

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

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 $m < r^*$, the assumption on the simplicity of the eigenvalues of $Z^*$ is not mandatory, and this leads to several local minimizers of the loss (see Appendix C.1 for more detail). Overall, it is easily seen that those achieve the same value of the loss, which allows to conclude that every local minimizer of $\mathcal{L}$ is global. For $m \geq r^*$, 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 $m > r^*$, 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 [Simon Martin, Francis Bach, Giulio Biroli, 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:

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

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 $(w^0_1, \ldots, w^0_m)$ and $(w^*_1, \ldots, w^*_m)$ 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 Appendix A.2). From now on, the dependency in the dimension is made explicit when relevant. Thus, defining $U^0_d = \sqrt{m_d} W^0_d$ and $U^*_d = \sqrt{m_d} W^*_d$ (the initialization $W^0_d$ and the teacher matrix $W^*_d$ are linked to their respective vectors in equation (2)), we assume that:

$$U^0_d U^0_d^T = I_{m_d}, \quad U^*_d U^*_d^T = I_{m^*_d}. \quad (9)$$

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 $Y_d = U^0_d U^*_d U^*_d^T U^0_d \in \mathbb{R}^{m_d \times m_d}$. Indeed, we will make use of the implicit solution for the flow obtained in equation (6), which depends on the matrix $W^0_d \exp(4W^0_d W^*_d^T t) W^0_d$. Under our orthonormality assumption, we have the more convenient formula:

$$U^0_d \exp(\lambda^* U^*_d U^*_d^T) U^0_d = I_{m_d} + (e^\lambda - 1) Y_d, \quad (10)$$

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 [Kunisky, 2023; Aubrun, 2021; Hiai and Petz, 2005]).
Since $U_d^0 \in \mathbb{R}^{d \times m_d}$ and $U_d^1 \in \mathbb{R}^{d \times m_d}$, equation (9) requires that $m_d, m_d^2 \leq d$. As $d \to \infty$, we will assume that $m_d$ and $m_d^2$ diverge but keep a fixed ratio with $d$:

$$\alpha = \lim_{d \to \infty} \frac{m_d}{d} \in [0, 1], \quad \alpha^* = \lim_{d \to \infty} \frac{m_d^2}{d} \in [0, 1].$$

Thus, we let the number of teachers and students to vary with the dimension, leading to predictors parameterized by a continuum of neurons. We refer to Appendix A.2 for a formal statement of the assumptions under which the following results are valid.

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

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

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

$$d\mu(x) = \frac{\sqrt{(r_+ - x)(x - r_-)}}{2\alpha x(1 - x)} \mathbb{1}_{[r_-, r_+]}(x) dx + \left(1 - \frac{\alpha^*}{\alpha}\right) \delta_0(dx) + \frac{1}{\alpha} (\alpha + \alpha^* - 1)^+ \delta_1(dx),$$

(11)

where $\delta_y$ stands for the Dirac distribution at $y \in \mathbb{R}$, $u^+ = \max(u, 0)$ and:

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

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

![Figure 1: Probability density function of $\mu$ defined in equation (11) (dashed line) and histogram of the eigenvalues of $Y_d$ for $d = 2000$ with 30 bins, with $\alpha = 0.3$, $\alpha^* = 0.5$.](image)

This convergence result is a key ingredient to characterize the high-dimensional limit. Indeed, the relevant quantities that will be studied in the following section mainly depend on the matrix $Y_d$. Our first result identifies the high-dimensional limit of $\psi_d$ defined in Proposition 3.1. This function is solution of the implicit equation (6), and the convergence of the empirical spectral distribution of $Y_d$ will allow to derive the infinite-dimensional equivalent of this equation (see Proposition 5.1).

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 $\psi_d$. 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 $\gamma$ is fixed and accounts for a renormalized time variable. We also set:

$$\phi_d(\gamma) = \int_0^{t_d} \text{Tr}(W_d(s)W_d(s)^T - W_d^*W_d^*T) ds$$

(12)

$$= \psi_d(\gamma_d) - \gamma_d,$$

since under our assumptions, we have $\text{Tr}(Z_d^*) = 1$. Note that we cannot hope for a high-dimensional limit for $\psi_d$ 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 (1) with initial condition $W_d^0$. Then, with probability one, as $d \to \infty$, then the function $\phi_d$ defined in equation (12) uniformly converges on any compact of $\mathbb{R}_+$ towards a function $\phi$ which is solution of the equation:

$$4\gamma = \alpha \log F_\phi(\gamma) + (\alpha + \alpha^* - 1)^+ \log \left(\frac{G_\phi(\gamma)}{F_\phi(\gamma)}\right)$$

$$+ \Theta\left(\frac{G_\phi(\gamma)}{F_\phi(\gamma)} - 1\right),$$

(13)

with:

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

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

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

This result is proven in Appendix C.3. The idea is to start from equation (6) which allows to obtain an implicit equation of $\phi_d$. The key observation is that this equation only depends on the matrix $Y_d$ (this is a consequence of equation (10)). Then, using the convergence result for the empirical spectral distribution of $Y_d$, one can derive equation (13) from equation (6).
The previous equations, in which the dimension only appears through the parameters $\alpha$ and $\alpha^*$, 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 $\psi_d$: this new equation is purely deterministic and does not depend on the initialization of the flow.

Equation (13) can be interpreted as a coupled ordinary differential equation on the couple $(F_\phi(\gamma), G_\phi(\gamma)/F_\phi(\gamma))$. This allows to prove that the solution $\phi$ to this equation is unique (see Lemma C.9), and to simulate numerically the function $\phi$ in the high-dimensional limit (see Appendix D.4). 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 $\phi_d$ for increasing values of $d$, as well as the numerical solution of equation (13). 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 $\phi_d$ 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 $\phi_d$ in the different regimes can be also understood from finite dimension estimation. In Lemma C.2, 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:

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

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 $\chi_d^0$, 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, $\chi_d^0$ 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 $\chi_d(\gamma) = \chi(Z_d(\gamma d), Z_d^*)$ 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. Moreover, we are able to derive a formula for the overlap in the high-dimensional limit (involving integrals against the measure $\mu$ defined in equation (11) and an implicit function associated with the function $\phi$, see Lemma C.11, as well as tight convergence rates for the $\gamma \to \infty$ limit displayed in the previous proposition (see Lemma C.14 for more details).

Regarding the behaviour of the overlap, the overall shape is similar in all three cases, but the limit value for the underparameterized case is lower.

This last result gives 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 $Z_d^\gamma$ and the limit $Z_d^\gamma = \lim_{t \to \infty} Z_d(t)$ 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:

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

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 4 a simulation of the function $\phi_d$ (directly computed from the flow), where both teacher and students are independently drawn from a Gaussian distribution $N(0, I_d/d)$. This is compared to the approximate numerical solution of equation (13), 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 $\phi_d$ 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.
Figure 4: Simulations of $\phi_d(\gamma)$ (defined in equation (12)) with $\gamma = t/d$ for increasing values of $d$. Finite $d$: simulation using gradient descent with step-size $\eta = 10^{-2}$, and Gaussian $N(0, I_d/d)$ teachers and students at initialization. $d = \infty$: simulations using discretized version of equation (13). 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.

Further work. The assumptions we made along this paper leave 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 4.

- 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 [Mannelli et al., 2020], [Gamarnik et al., 2019], [Du and Lee, 2018]), 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


E. Massart and P.-A. Absil. Quotient geometry with simple geodesics for the manifold of fixed-


Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, it is done in subsection 2.2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
      Yes, we provide a code to reproduce the figures of the paper.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results.
      Yes, most of the proofs are done in Appendix C and D.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
      Yes.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
      Yes, we provide the training details in Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
      Not Applicable, our figures are illustrations of proven theoretical results.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
      Yes, computing infrastructure is described in Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.
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 in equation (3), 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 \( \lambda^* \) the smallest non-zero eigenvalue of \( Z^* \). Then, under Assumption 4.1 with probability one:

- If \( m, m^* \geq d \):
  \[ \mathcal{L}(W(t)) \bigg|_{t \to \infty} = O\left(e^{-8\lambda^* t}\right). \]

- If \( m = m^* < d \):
  \[ \mathcal{L}(W(t)) \bigg|_{t \to \infty} = O\left(e^{-4\lambda^* t}\right). \]

- If \( m, d > m^* \):
  \[ \mathcal{L}(W(t)) \bigg|_{t \to \infty} = O\left(\frac{1}{t^2}\right). \]

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 \( t \to \infty \) (see Proposition 4.1). 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 \( \psi(t) = \int_0^t \text{Tr}(W(s)W(s)^T) ds \) introduced in Proposition 3.1.

**Figure 5:** 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 5** displays the diagram of convergence rates in the overparameterized 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 \( \text{rank}(W(t)W(t)^T) = \text{rank}(Z^*) \) 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 \( u_*(x) = \text{Tr}(W^*W^{*T}xx^T) \) (see equation 1).

Finally, we believe that the bounds derived in Proposition A.1 are tight. As depicted in Figure 6, in the case \( m^* < m \leq d \) (left panel), the function \( t^2 \times \mathcal{L}(W(t)) \) stays bounded with time and seems to converge towards a finite value. Likewise, for the case \( m = m^* \leq d \) (right panel), the loss \( \mathcal{L}(W(t)) \) follows the line drawn by \( e^{-4\lambda^* t} \), where \( \lambda^* \) is the smallest non-zero eigenvalue of \( Z^* \) (in log-scale on the \( y \)-axis).

![Figure 6: Evolution of the loss as a function of time. Left: \( m = d/2 \) and \( m^* = d/4 \), renormalized loss \( \mathcal{L}(t) = t^2 \mathcal{L}(W(t)) \). Right: \( m = m^* = d/2 \), \( \mathcal{L}(t) = \mathcal{L}(W(t)) \) and dotted lines correspond to \( \text{cst} \times e^{-4\lambda^* t} \) where \( \lambda^* \) 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.](image)

### 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 \( Y_d = U_d^0U_d^*U_d^T U_d^0 \in \mathbb{R}^{m_d \times m_d} \), where \( U_d^0U_d^* = I_{m_d} \) and \( U_d^0U_d^* = I_{m_d^*} \). 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 \( \mathcal{M}_{d,k} = \{ U \in \mathbb{R}^{d \times k} , U^TU = I_k \} \) can be endowed with a uniform measure, and \( U^0_d, U^*_d \) 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 \( U = V(V^TV)^{-1/2} \), so that, conditionally on the probability one event \( \{ V^TV \text{ invertible} \} \), \( U \) is uniform on \( \mathcal{M}_{d,k} \) (see Götze and Sambale 2023). Thus, in all the following, we make the assumption:

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

\[
W^0_d = \frac{1}{\sqrt{m_d}} U^0_d, \quad W^*_d = \frac{1}{\sqrt{m_d^*}} U^*_d,
\]

where the matrices \( U^0_d \) and \( U^*_d \) are uniformly drawn on \( \mathcal{M}_{d,m_d} \) and \( \mathcal{M}_{d,m_d^*} \) respectively, with:

\[
\frac{m_d}{d} \xrightarrow{d \to \infty} \alpha \in [0,1], \quad \frac{m_d^*}{d} \xrightarrow{d \to \infty} \alpha^* \in [0,1].
\]

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 $X_d = U_d^0 Y_d U_d^{0T} \in S_d^+(\mathbb{R})$. Informally, they obtain that for all $f : [0, 1] \to \mathbb{R}$ continuous:

$$
\frac{1}{d} \text{Tr}(f(X_d)) \xrightarrow{d \to \infty} \int f(x)d\nu(x),
$$

where:

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

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 uniformly drawn (this corresponds to Assumption A.1).

To recover the measure defined in equation (11), note that $X_d$ and $Y_d$ have the same non-zero eigenvalues. Thus, we have for $f : [0, 1] \to \mathbb{R}$ continuous:

$$
\frac{1}{m_d} \text{Tr}(f(Y_d)) = \frac{d}{m_d} \frac{1}{d} \text{Tr}(f(X_d)) - \frac{d - m_d}{m_d} f(0) \xrightarrow{d \to \infty} \frac{1}{\alpha} \int f(x)d\nu(x) - \frac{1 - \alpha}{\alpha} f(0).
$$

Thus, the empirical spectral distribution of $Y_d$ converges as $d \to \infty$ towards a probability measure $\mu$ which satisfies $\alpha \mu = \nu - (1 - \alpha)\delta_0$, which corresponds to equation (11).

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

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

**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:

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

and:

$$
P \left( \sup_{f \in \mathcal{F}} \left| \frac{1}{m_d} \text{Tr}(f(Y_d)) - \int f(x)d\mu(x) \right| \geq \eta \right) \leq P \left( \sup_{1 \leq i \leq n} \left| \frac{1}{m_d} \text{Tr}(f_i(Y_d)) - \int f_i(x)d\mu(x) \right| \geq \frac{\eta}{2} \right). \quad (14)
$$

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

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

where $\mu$ is defined in equation (11) and $\mathcal{M}$ denotes the space of probability measures over $[0, 1]$. First, since the supremum is over a finite number of functions, $E_\eta$ is closed for the weak topology on $\mathcal{M}$. Now Hiai and Petz (2005) state that:

$$
\limsup_{d \to \infty} \frac{1}{d^2} \log P(\mu_d \in E_\eta) \leq - \inf_{\xi \in E_\eta} I(\nu) \equiv -\beta_\eta, \quad (15)
$$

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 $(\xi_p)_{p \in \mathbb{N}}$ a sequence of
$E_\eta$ such that $I(\zeta_p) \to 0$. Since $\mathcal{M}$ is compact, one can extract a converging subsequence $\zeta_{\phi(p)} \to \zeta$. Thus, by lower semi-continuity of $I$: 

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

and $\zeta = \mu$. Since $E_\eta$ is closed for the weak topology, this implies that $\mu \in E_\eta$ which leads to a contradiction by definition of $E_\eta$. Thus, from equation (14), for all $\eta > 0$:

$$\sum_{d \geq 0} \mathbb{P} \left( \sup_{f \in \mathcal{F}} \left| \frac{1}{m_d} \text{Tr}(f_i(Y_d)) - \int f_i(x) d\mu(x) \right| \geq \eta \right) \leq \exp \left( -\frac{\beta_d \eta^2}{2} \right),$$

and using equation (14), we get that, for all $\eta > 0$:

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

and the conclusion using Lemma B.4.

### A.3 Overlap

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

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

where $\| \cdot \|$ is the norm associated with the inner product on $E$. The overlap is a 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 drawn from a rotationally invariant distribution over $\mathbb{R}^d$, then:

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

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 $\mathcal{S}_d(\mathbb{R})$:

$$\chi_{\text{sym}}(X, Y) = \frac{|\text{Tr}(XY)|}{\|X\|_F \|Y\|_F}.$$

This correspond to the standard overlap on the Euclidean space $\mathcal{S}_d(\mathbb{R})$ of symmetric matrices endowed with its standard inner product. Note that we recover the $\mathbb{R}^d$ overlap with $\chi_{\text{sym}}(xx^T, yy^T) = \chi(x, y)^2$ 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.** Under Assumption A.1 the overlap between the teacher matrix and the student at initialization admits the limit, with probability one:

$$\chi_{\text{sym}}(Z_d^0, Z_d^*) \xrightarrow{d \to \infty} \sqrt{\alpha\alpha^*}.$$

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

$$\chi_{\text{sym}}(Z_d^0, Z_d^*) = \frac{\text{Tr}(U_d^{0\dagger}U_d^*U_d^{\star\dagger}U_d^{0\dagger})}{\|U_d^{0\dagger}U_d^*\|_F \|U_d^*U_d^{\star\dagger}\|_F}.$$

For the denominator, we obtain:

$$\|U_d^{0\dagger}U_d^*\|_F = \sqrt{m_d}, \quad \|U_d^*U_d^{\star\dagger}\|_F = \sqrt{m_d}.$$
Due to the rotational invariance of the distribution of $U_d^0U_d^{\top}$, one can suppose without loss of generality that $U_d^0U_d^{\top}$ is the orthogonal projection onto the $m_0^*$ first vectors of the standard basis in $\mathbb{R}^d$, denoted $\Pi_d$. Thus, one can rewrite:

$$\chi^\text{sym}(Z_d^0, Z_d^0) = \frac{1}{\sqrt{m_dm_d}} \sum_{k=1}^{m_d} \|\Pi_d(v_k)\|^2,$$

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

$$\Pr \left( \left| \chi^\text{sym}(Z_d^0, Z_d^0) - \frac{\sqrt{m_dm_d}}{d} \right| \geq \epsilon \right) \leq \sum_{k=1}^{m_d} \Pr \left( \left| \|\Pi_d(v_k)\| - \frac{m_d^*}{d} \right| \geq \epsilon \sqrt{\frac{m_d^*}{m_d}} \right).$$

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

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

which leads to:

$$\Pr \left( \left| \chi^\text{sym}(Z_d^0, Z_d^0) - \frac{\sqrt{m_dm_d}}{d} \right| \geq \epsilon \right) \leq 2 m_d \exp \left( -\frac{\epsilon^2 d^2}{4m_d} \right).$$

Using Lemma B.4 we get that with probability one:

$$\left| \chi^\text{sym}(Z_d^0, Z_d^0) - \frac{\sqrt{m_dm_d}}{d} \right| \rightarrow 0, \quad d \rightarrow \infty,$$

and the result. \(\square\)

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 $d \rightarrow \infty$. 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 $m \leq d$:

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

**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:

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

so that:

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

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

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

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

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

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

\[ \lambda_{\min}(B) \text{Tr}(A) \leq \text{Tr}(AB) \leq \lambda_{\max}(B) \text{Tr}(A). \]

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:

\[ \lambda_j = \max_{V \subset \mathbb{R}^d} \min_{\|x\|=1} x^T Ax = \min_{V \subset \mathbb{R}^d} \max_{\|x\|=1} x^T Ax. \]

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:

\[ \int_0^t e^{\lambda s} v(s) ds \sim \frac{1}{\lambda} e^{\lambda t} v(t). \]

Moreover, for all $\mu > 0$:

\[ e^{-\mu t} \int_0^t e^{\mu s} ds \xrightarrow{t \to \infty} 0. \]

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

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

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:

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

\[ \square \]

Lemma B.4 (Borel-Cantelli Lemma). Let $(X_n)_{n \in \mathbb{N}}$ and $X$ be real-valued random variables. Suppose that for all $\eta > 0$:

\[ \sum_{n \geq 0} \mathbb{P}(\{X_n - X \geq \eta\}) < \infty. \]

Then, almost surely, $X_n \xrightarrow{n \to \infty} X$.

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 use the fact that under Assumption 4.1, for almost every initializations, the gradient flow defined in equation (4) converges towards a local minimizer of the loss $L$, i.e., a point $W \in \mathbb{R}^{d \times m}$ satisfying:

\[ \nabla L(W) = 0 \quad \text{and} \quad \text{Tr}(d^2 L_W(K) K^T) \geq 0, \]

for all $K \in \mathbb{R}^{d \times m}$. We briefly prove this fact using several prior works: since the loss function $L$ is analytic in the coefficients of $W$ and has compact sublevel sets, Theorem 2.2 in [Absil et al., 2005] guarantees that the gradient flow converges towards a critical point of $L$, that is a point $W \in \mathbb{R}^{d \times m}$ such that $\nabla L(W) = 0$. Following [Bah et al., 2021], Theorem 28, who generalize the results of [Lee et al., 2019] in continuous times, the set of
initializations such that the flow converges to a strict saddle (a critical point with at least one direction of negative curvature) has zero Lebesgue measure. This allows to conclude that if the distribution of the initialization is absolutely continuous with respect to the Lebesgue measure (Assumption 4.1), then with probability one, the flow converges towards a local minimizer of the loss.

We are now ready to prove Proposition 4.1. The proof is done 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 \( \nabla L(W) = 0 \).
- 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.

Before stating the lemma, we remind the notations for the decomposition of the teacher matrix \( Z^* \):

\[
Z^* = \sum_{k=1}^{r^*} \mu_k v_k^*(v_k^*)^T,
\]

with \( r^* = \min(d,m^*) \) being its rank, \( \mu_1 \geq \cdots \geq \mu_{r^*} > 0 \) its non-zero eigenvalues and \( v_1^*, \ldots, v_{r^*}^* \in \mathbb{R}^d \) the associated eigenvectors.

**Lemma C.1.** Let \( W \in \mathbb{R}^{d \times m} \) be a critical point of \( L \). 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:

\[
\tau = \frac{1}{2 + |I|} \sum_{1 \leq k \leq r^* \atop k \notin I} \mu_k.
\]

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

\[
2(Z^* - WW^T)W + \text{Tr}(Z^* - WW^T)W = 0.
\]

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

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

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

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

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

\[
\tau = \frac{1}{2} \text{Tr}(Z^* - WW^T) = \frac{1}{2} \left( \text{Tr}(Z^*) - \sum_{k \in I} (\mu_k + \tau) \right) = \frac{1}{2} \sum_{1 \leq k \leq r^* \atop k \notin I} \mu_k - \frac{1}{2} |I| \tau,
\]

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 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 \( \text{rank}(W) < m \), \( WW^T \) 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 \( F \):

\[
\nabla L(W) = 2 \nabla F(WW^T)W \\
\text{Tr}(d^2L_W(K)K^T) = 2 \text{Tr}(\nabla F(WW^T)KK^T) + \text{Tr}(d^2F_{WW^T}(WK^T + KW^T)(WK^T + KW^T)).
\]

If \( W \in \mathbb{R}^{d \times m} \) is a local minimizer of \( L \), then \( \nabla F(WW^T)W = 0 \) and \( \text{Tr}(d^2L_W(K)K^T) \geq 0 \) for all \( K \in \mathbb{R}^{d \times m} \). If \( \text{rank}(W) < m \), 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 \( K = uu^T \), so that \( WK^T \) and \( KW^T \) are zero. Thus, we get that \( \text{Tr}(\nabla F(WW^T)uu^T) \geq 0 \) 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 \( S \in \mathcal{S}_d^+(\mathbb{R}) \):

\[
F(S) \geq F(WW^T) + \text{Tr}(\nabla F(WW^T)(S - WW^T)) = F(WW^T) + \text{Tr}(\nabla F(WW^T)S),
\]

where we used that \( \nabla F(WW^T)W = 0 \) along with the fact that the trace of a product of matrices in \( \mathcal{S}_d^+(\mathbb{R}) \) is positive. Thus \( WW^T \) is a global minimizer of \( F \) over \( \mathcal{S}_d^+(\mathbb{R}) \). \( \square \)

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

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

We now write the local minimizer condition \( \text{Tr}(d^2L_W(K)K^T) \geq 0 \):

\[
\text{Tr}((WW^T - Z^*)KK^T) - \tau \text{Tr}(KK^T) + \text{Tr}(WW^TWW^T) + \text{Tr}(WW^TWW^T) + \text{Tr}(WK^T) \geq 0,
\]

where \( \tau \) 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 \( Z^* \) (\( \mu_1 > \cdots > \mu_r > 0 \)), 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 \( |I| = \text{rank}(W) = m \) by assumption, necessarily \( I = [1, m] \). For \( m \leq r^* \), this implies the desired by the expression of \( \tau \) (obviously \( \tau = 0 \) if \( m = m^* \leq d \)). For \( m > m^* \) (in this case \( m^* < d \) since we assumed \( m \leq d \)), 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 \( WW^T = Z^* \). This concludes the proof.

When the teacher has multiple eigenvalues. When the case where the eigenvalues of \( Z^* \) are not simple anymore, the proof can be done the same way. Obviously the result is the same for \( m \geq \min(m^*, d) \). The only difference is when \( m < \min(m^*, d) \). As soon as \( j \in I \) and \( k < j \), we still have that \( \lambda_k \geq \mu_k - \mu_j \). Now if \( \mu_k = \mu_j \), this does not imply that \( k \in I \). As a consequence, the local minimizers can be understood this way: take \( \mu_m > 0 \) to be the eigenvalue at position \( m \) in the spectrum of \( Z^* \), and look at the set of the eigenvalues of \( Z^* \) equal to \( \mu_m \), which is of the form \([m - p + 1, m + q]\) for some integers \( p, q \). Using the same technique as previously, it can be shown that the local minimizers of the loss are of the form:

\[
WW^T = \sum_{k=1}^{m-p} (\mu_k + \tau)v_k^*(v_k^*)^T + (\mu_k + \tau) \sum_{k \in J} v_k^*(v_k^*)^T.
\]
where $J$ can be any subset of $\{m-p+1, m+q\}$ of size $p$. $\tau$ is defined in equation (1) and its value is independent from the choice of $J$. What we see from the previous equation is that the local minimizers have a common part corresponding to the eigenvalues of $Z^*$ which are larger that $\mu_m$, and a second contribution from the fact that there are still $p$ eigenvalues to choose to reach $m$ (the maximal rank for $W$). Overall, the number of local minimizers in this case (up to representation $WW^T$) is:

$$\binom{p+q}{p}.$$

Obviously this also depends on $m$ since the values of $p, q$ depend on the location and the multiplicity of the $m^{th}$ largest eigenvalue of $Z^*$. In the extreme case where the non-zero eigenvalues of $Z^*$ are equal (i.e. when $Z^*$ is proportional to an orthogonal projection matrix), we have $p = m$ and $q = m^* - m$, leading to the maximal number of local minimizers possible (at fixed $m, m^*$).

The main consequence of this reasoning is despite the presence of multiple local minimizers of the loss, the associated loss value is the same for all of those, and it is easily seen that they correspond to global minimizers.

### 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 $m \geq r^*$. As explained in Section 3, the understanding of the function $\psi$ defined in Proposition 3.1 is essential to the determination of the dynamics. Under the assumption that $W(t)W(t)^T \to Z^*$ as $t \to \infty$, which happens with probability one when the flow is correctly initialized, we have:

$$\psi(t) = \int_0^t \text{Tr}(W(s)W(s)^T) ds \sim t \to \infty \text{Tr}(Z^*)t. \quad (16)$$

The first challenge will be to gather more information about $\psi$, using equation (9):

$$\psi(t) = \frac{1}{4} \text{Tr} \log \left( I_m + 4W^0t \int_0^t e^{-2\psi(s)} e^{2s \text{Tr}(Z^*)} e^{4Z^*s} ds W^0 \right). \quad (17)$$

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

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

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

Our results are verified under Assumption 4.1. As mentioned before, for $m \geq r^*$, we have $W(t)W(t)^T \to Z^*$ with probability one. Moreover, almost surely, any subfamily of $(W^0v_1, \ldots, W^0v_d)$ 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 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 $\psi$, it is important to understand how the eigenvalues of $W^0tE^*(t)W^0$ 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 $\psi$. Indeed, the fact that $\psi(t) \sim t \to \infty \text{Tr}(Z^*)t$ is enough to prove the result.

In Lemma C.3 and Lemma C.4, we will assume that $m^* \leq d$ (in this case the overparameterized regime corresponds to $m \geq m^*$). Indeed, whenever $m^* > d$, the proof of convergence rates will be straightforward. In all the following we let $r = \min(m, d) \geq m^*$. 

Simon Martin, Francis Bach, Giulio Biroli
Lemma C.3. Denote \( \lambda_1(t), \ldots, \lambda_r(t) > 0 \) the non-zero eigenvalues of \( W^T E^*(t) W \). Then, for \( j \in [1, r] \):

\[
\lambda_j(t) \to \Theta \left( \int_0^t e^{-2\psi(s)} e^{2s \text{Tr}(Z^*)} e^{4\mu_j s} ds \right).
\]

The notation \( \Theta \) means that \( \lambda_j(t) \) is bounded from below and above by constants times the above integral. This lemma is proven in Appendix D.1. The goal is now to use this result along with equation (17) 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 \( m \geq m^* \):

\[
\psi(t) \to \infty \text{ Tr}(Z^*) t + \frac{1}{2} \frac{r - m^*}{r + 2} \log(t) + O(1).
\]

Proof. Set \( \xi(t) = \psi(t) - \text{Tr}(Z^*) t \) which is \( o(t) \) by equation (16). From the previous lemma, let \( C_j, D_j > 0 \) such that:

\[
C_j \leq \lambda_j(t) \left( \int_0^t e^{-2\psi(s)} e^{2s \text{Tr}(Z^*)} e^{4\mu_j s} ds \right)^{-1} \leq D_j.
\]

In the following, we only use the upper bound, the lower bound will be done similarly. By equation (17) and using the notations of the previous lemma:

\[
\psi(t) = \frac{1}{4} \sum_{j=1}^{r} \log \left( 1 + 4\lambda_j(t) \right)
\leq \frac{1}{4} \sum_{j=1}^{m^*} \log \left( 1 + 4D_j \int_0^t e^{-2\xi(s)} e^{4\mu_j s} ds \right) + \frac{1}{4} \sum_{j=m^*+1}^{r} \log \left( 1 + 4D_j \int_0^t e^{-2\xi(s)} ds \right).
\]

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

\[
\frac{1}{4} \sum_{j=1}^{m^*} \log \left( 1 + 4D_j \int_0^t e^{-2\xi(s)} e^{4\mu_j s} ds \right) = \text{Tr}(Z^*) t - \frac{m^*}{2} \xi(t) + \frac{1}{4} \sum_{j=1}^{m^*} \log \left( \frac{D_j}{h_j} \right) + o(1).
\]

For the second we need to understand how \( \int_0^t e^{-2\xi(s)} ds \) behave. As \( t \to \infty \), this integral either converge 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 it was assumed that \( \int_0^t e^{-2\xi(s)} ds \) converges at \( t \to \infty \). Thus, we obtain:

\[
\left( 1 + \frac{m^*}{2} \right) \xi(t) \lesssim \frac{r - m^*}{4} \log \left( \int_0^t e^{-2\xi(s)} ds \right) + \frac{1}{4} \sum_{j=1}^{m^*} \log \left( \frac{D_j}{h_j} \right) + \frac{1}{4} \sum_{j=m^*+1}^{r} \log(4D_j) + o(1).
\]

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

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

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

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

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

\[
\zeta + h_1(t) + \frac{\beta}{2(\beta + 1)} \log \left( \int_0^t e^{-2h_2(s)} ds \right) \leq \xi(t) \leq \zeta + h_2(t) + \frac{\beta}{2(\beta + 1)} \log \left( \int_0^t e^{-2h_1(s)} ds \right),
\]
where $\zeta$ is a constant depending only on $\beta$. We now use that $h_1(t), h_2(t)$ are bounded. 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 $\beta$ and $r = \min(m, d)$.

Now that the behaviour of $\psi$ is known, we are ready to determine the convergence rates in the case $m \geq r^*$. We show 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 \[ (1) \] (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:

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

(18)

where $Z_0 = \text{diag}(\mu_1, \ldots, \mu_r) \in \mathbb{R}^{r^* \times r^*}$ (if $m^* \geq d$, we simply assume that $Z^* = Z_0$). 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 $W(t)$ is solution of the flow then, $\dot{W}(t) = -M(t)W(t)$ with $M(t) = 2(W(t)W(t)^T - Z^*) + \text{Tr}(W(t)W(t)^T - Z^*) I_d$. Thus, using Lemma B.1

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

We now use the inequality $\text{Tr}(M(t)^2) \geq 8\mathcal{L}(W(t))$, and the fact that $\lambda_{\text{min}}(W(s)W(s)^T) \xrightarrow{t \to \infty} \lambda^* > 0$ the smallest eigenvalue of $Z^*$. Therefore:

$$\mathcal{L}(W(t)) \leq \mathcal{L}(W_0) \exp \left(-8 \int_0^t \lambda_{\text{min}}(W(s)W(s)^T) ds \right) = \mathcal{L}(W_0) \exp \left(-8\lambda^* t + o(t) \right).$$

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

$$\left| \int_0^t \lambda_{\text{min}}(W(s)W(s)^T) ds - \lambda^* t \right| \leq \int_0^t \|W(s)W(s)^T - Z^*\| ds \leq \sqrt{2} \int_0^t \sqrt{\mathcal{L}(W(s))} ds,$$

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:

$$Z(t) = e^{-2\xi(t)} e^{2Z^* t} W^0 \left( I_m + 4W^0 t \int_0^t e^{-2\xi(s)} e^{2Z^* s} \, ds \, W^0 \right)^{-1} W^0 t e^{2Z^* t},$$

with $\xi(t) = \psi(t) - \text{Tr}(Z^*) t$ 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 $\text{rank}(W_0) = \min(m, d)$ and $\text{rank}(U_0) = m^*$ (since $m \geq m^*$). In order to avoid sets of probability one, we restrict ourselves to this event. Then, from the previous equation:

$$Z(t) = e^{-2\xi(t)} \begin{pmatrix} e^{2Z^* t} U_0 N(t)^{-1} U_0^T e^{2Z^* t} \\ V_0 N(t)^{-1} U_0^T e^{2Z^* t} \end{pmatrix} V_0 N(t)^{-1} V_0^T \equiv \begin{pmatrix} A(t) \\ B(t)^T \\ C(t) \end{pmatrix},$$

(19)
with $A(t) \in S_m^+(\mathbb{R}), B(t) \in \mathbb{R}^{m \times (d-m)}$ and $C(t) \in S_{d-m}^+(\mathbb{R})$. To start, we are going to bound the bottom right term. From Lemma B.1:

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

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

$$\text{Tr}(C(t)) = \begin{cases} O(e^{-4\lambda^*t}) & \text{for } m = m^*; \\ O(t^{-1}) & \text{for } m > m^*, \end{cases}$$

(20)

where $\lambda^*$ 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 (18), this induces the evolution equations for $A(t)$ and $B(t)$:

$$\dot{A} = -2\dot{\xi} A + 2A(Z_0 - A) + 2(Z_0 - A)A - 4BB^T,$$

$$\dot{B} = -2\dot{\xi} B + 2(Z_0 - A)B - 2AB - 4BC. \quad (21)$$

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

$$\frac{d}{dt} \|B(t)\|_F^2 = -4\dot{\xi} \|B(t)\|_F^2 + 4\text{Tr}((Z_0 - A)BB^T) - 4\text{Tr}(ABB^T) - 8\text{Tr}(BCB^T)$$

$$\leq (-4\xi + 4\lambda_{\max}(Z_0 - A) - 4\lambda_{\min}(A))\|B\|_F^2.$$ 

Therefore:

$$\|B(t)\|_F^2 \leq \|B_0\|_F^2 e^{-4\lambda^*t} \exp \left(-\int_0^t \alpha(s)ds\right),$$

(22)

with:

$$\alpha(t) = 4\dot{\xi}(t) - 4\lambda_{\max}(Z_0 - A(t)) + 4\lambda_{\min}(A(t)) - 4\lambda^* \rightarrow 0 \quad \text{t} \rightarrow \infty,$$

(23)

since $A(t) \rightarrow Z_0$ and $\dot{\xi}(t) = \text{Tr}(Z(t) - Z^*) \rightarrow 0$ (note that $\lambda_{\min}(Z_0) = \lambda^*$). We now take care of $A(t)$ by introducing:

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

Differentiating and using the equation (21) on $A(t)$:

$$\dot{\phi} = -2\text{Tr}(AM^2) - 2\text{Tr}(C)^T\text{Tr}(AM) - 4\text{Tr}(BB^T M)$$

$$\leq -2\lambda_{\min}(A)\text{Tr}(M^2) + 2\text{Tr}(C)^T\text{Tr}(A^2)^{1/2}\text{Tr}(M^2)^{1/2} + 4\text{Tr}(BB^T B B^T)^{1/2}\text{Tr}(M^2)^{1/2},$$

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

$$\dot{\phi}(t) \leq -8\lambda_{\min}(A)\phi(t) + 2\sqrt{2(4 + m^*)}[\text{Tr}(C(t)^T)\text{Tr}(A(t)) + 2\text{Tr}(B(t)B(t)^T)] \sqrt{\phi(t)}.$$ 

$$\equiv \delta(t)$$

We set $A(t) = \int_0^t \lambda_{\min}(A(s))ds \sim \lambda^* t$, and obtain:

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

Using the definition of $\delta$, we have two terms to bound. From equation (22):

$$\int_0^t e^{4\lambda(s)}\text{Tr}(B(s)B(s)^T)ds \leq \|B_0\|_F^2 \int_0^t e^{4\lambda(s)} e^{-4\lambda^*s} \exp \left(-\int_0^u \alpha(u)du\right) ds.$$
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 (20) and split the cases. For $m = m^*$:

$$\int_0^t e^{4\Lambda(s)} \operatorname{Tr}(C(s)) \operatorname{Tr}(A(s)) ds \lesssim \int_0^t e^{4\Lambda(s)} e^{-4\lambda^* \cdot s} \operatorname{Tr}(A(s)) ds,$$

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

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

where $v(t) = o(t)$ as $t \to \infty$. In the case $m > m^*$, again with equation (20) and Lemma B.3

$$\int_0^t e^{4\Lambda(s)} \operatorname{Tr}(C(s)) \operatorname{Tr}(A(s)) ds \lesssim \frac{\operatorname{Tr}(Z_0) e^{4\lambda^*_t(t)}}{t},$$

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

$$\|Z(t) - Z^*\|^2_F \lesssim \|A(t) - Z_0\|^2_F + 2\|B(t)\|^2_F + \operatorname{Tr}(C(t)^2) \leq \phi(t) + 2\|B(t)\|^2_F + \left[\operatorname{Tr}(C(t))\right]^2.$$

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

$$\|Z(t) - Z^*\|^2_F = O(e^{-4\lambda^* t} \exp \left(-\int_0^t \alpha(s) ds\right)),$$

where $\alpha$ is defined in equation (23). We already have that $\int_0^t \alpha(s) ds = o(t)$ as $t \to \infty$ so that we know the main behaviour of $\|Z(t) - Z^*\|^2_F$. We now show that $\int_0^t \alpha(s) ds$ remains bounded. From the expression of $\xi$:

$$\left|\int_0^t \alpha(s) ds\right| \leq 4 \int_0^t |\operatorname{Tr}(Z(s) - Z^*)| ds + 4 \int_0^t |\lambda_{\max}(Z_0 - A(s))| |ds| + 4 \int_0^t |\lambda_{\min}(A(t)) - \lambda^*| |ds| \leq 4 \int_0^t \|Z(s) - Z^*\|_F ds + 8 \int_0^t \|A(s) - Z_0\|_F ds.$$

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

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

### C.3 Proof of Proposition 5.1

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

$$\psi_d(t) = \frac{1}{4} \operatorname{Tr} \log \left(I_{m_d} + 4 W_{d}^{0T} \int_0^t e^{-2\psi_d(s)} e^{2s \operatorname{Tr}(Z_d^s)} e^{4Z_d^s \cdot ds} W_d^0 \right).$$

In order to understand the high-dimensional limit of $\psi_d$ 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 $\psi_d$. However, this becomes easier under Assumption A.1 which states that the matrices $U_d^T U_d$ and $U_d U_d^T$ are uniformly drawn orthonormal projections (see Appendix A.2 for a more precise definition). Indeed, equation (10) shows that:

$$W_{d}^{0T} e^{4Z_d^s \cdot W_d^0} = \frac{1}{m_d} \left( I_{m_d} + (e^{4s/m_d} - 1) U_d^{0T} U_d^s U_d^s T U_d^0 \right).$$

Thus, the integrals involving $\psi_d$ 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 $\phi_d$ using the implicit equation (24). Through
those steps, the goal is to obtain sufficient information on \( \phi_d \) 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 \( C_T \) the space of continuous functions on \([0,T]\) equipped with the norm:

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

### C.3.1 Useful bounds

As mentioned earlier, the two following lemmas allow to gather sufficient information on \( \phi_d \) in order to extract converging subsequences in \( C_T \) 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 (12):

\[
\phi_d(\gamma) = \int_0^\gamma \text{Tr}(W_d(s)W_d(s)^T - W_d^*W_d^*T)ds,
\]

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:

\[
\frac{d}{dt} \mathcal{L}(W_d(t)) = -\|\nabla \mathcal{L}(W_d(t))\|^2 \leq 0.
\]

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

\[
\left[ \text{Tr}(W_d(t)W_d(t)^T - W_d^*W_d^*T) \right]^2 \leq 4 \mathcal{L}(W_d(t)) \leq 4 \mathcal{L}(W_d^0).
\]

We now use the orthonormality assumption:

\[
\mathcal{L}(W_d^0) = \frac{1}{2} \left\| \frac{1}{m_d} U_d^r U_d^{rT} - \frac{1}{m_d^*} U_d^* U_d^{*T} \right\|_F^2 + \frac{1}{4} \left( \frac{1}{m_d} \text{Tr}(U_d^r U_d^{rT}) - \frac{1}{m_d^*} \text{Tr}(U_d^* U_d^{*T}) \right)^2 = 0.
\]

\[
\leq \frac{1}{2} \left( \frac{1}{m_d} + \frac{1}{m_d^*} \right).
\]

Since \( m_d/d \xrightarrow{d \to \infty} \alpha \) and \( m_d^*/d \xrightarrow{d \to \infty} \alpha^* \), we obtain that:

\[
\left| \text{Tr}(W_d(t)W_d(t)^T - W_d^*W_d^*) \right| \leq \frac{\kappa}{\sqrt{d}},
\]

for some \( \kappa \) independent from \( t \) and \( d \). Since \( \dot{\phi}_d(\gamma) = d \text{Tr}(W_d(\gamma d)W_d(\gamma d)^T - W_d^*W_d^*T) \), we obtain the desired result.

**Lemma C.6.** For all \( T > 0 \):

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

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

\[
\sup_{\gamma \in [0,T]} \sup_{d \in \mathbb{N}} \sup_{|\gamma - \gamma'| \leq \eta} \left| \phi_d(\gamma) - \phi_d(\gamma') \right| \xrightarrow{\eta \to 0} 0.
\]

(26)
From equation (27), we have that, for all equations (11). Therefore, we also define:

\[ \phi_d(\gamma) + \gamma d = \frac{1}{4} \text{Tr} \log \left( 1 + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi(s)} ds \right) I_{m_d} + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi(s)} \left( e^{4s/m_d} - 1 \right) ds U_d^T U_d^* T U_d^T. \]  

(27)

C.3.2 Identifying the limit

The previous lemma shows that the family \( (\phi_d)_{d \in \mathbb{N}} \) is compact in \( C_T \), 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 \( \phi : [0,T] \to \mathbb{R} \) be a subsequential limit of \( (\phi_d)_{d \in \mathbb{N}} \) in \( C_T \). Then, with probability one, for all \( \gamma \in [0,T] \):

\[ \int \log \left( 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\phi(s)} ds + \frac{4x}{\alpha} \int_0^\gamma e^{-2\phi(s)} \left( e^{4s/\alpha} - 1 \right) ds \right) d\mu(x) = \frac{4\gamma}{\alpha}, \]

where \( \mu \) is defined in equation (11).

**Proof.** For \( \xi \in C_T \), we set:

\[ \mathcal{H}_d(\xi)(\gamma) = \frac{1}{4d} \text{Tr} \log \left( 1 + \frac{4d}{m_d} \int_0^\gamma e^{-2\xi(s)} ds \right) I_{m_d} + \frac{4d}{m_d} \int_0^\gamma e^{-2\xi(s)} \left( e^{4s/m_d} - 1 \right) ds U_d^T U_d^* Y_d. \]

From equation (27), we have that, for all \( \gamma \geq 0 \):

\[ \mathcal{H}_d(\phi_d)(\gamma) = \gamma + \frac{\phi_d(\gamma)}{d} \equiv u_d(\gamma). \]

(29)

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 \( \mu \) defined in equation (11). Therefore, we also define:

\[ \mathcal{H}(\xi)(\gamma) = \frac{\alpha}{4} \int \log \left( 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\xi(s)} ds + \frac{4x}{\alpha} \int_0^\gamma e^{-2\xi(s)} \left( e^{4s/\alpha} - 1 \right) ds \right) d\mu(x). \]

(30)

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

**Lemma C.8.** For \( c \in \mathbb{R} \), let \( B_c = \{ \xi \in C_T, \xi \geq c \} \). Then, with probability one:

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

(31)

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 \( \phi \) in \( C_T \). From equation (29), 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 (30). We have:

\[ \| \mathcal{H}(\phi) - u \| \leq \| \mathcal{H}(\phi) - \mathcal{H}(\phi_{d(n)}) \| + \| \mathcal{H}(\phi_{d(n)}) - \mathcal{H}_d(n)(\phi_{d(n)}) \| + \| \mathcal{H}_d(n)(\phi_{d(n)}) - u \|. \]

(32)

We already know that the last term goes to zero from equation (29). For the first one, we need to show that \( \mathcal{H} \) is continuous. Indeed, for \( \xi, \zeta \in C_T \) both lower bounded by some \( c \in \mathbb{R} \), using the fact that the map
\[ x \in \mathbb{R}^+ \mapsto \log(1 + x) \] is 1-Lipschitz:

\[
\left| \mathcal{H}(\xi)(\gamma) - \mathcal{H}(\zeta)(\gamma) \right| \leq \frac{\alpha}{4} \sup_{x \in [0,1]} \left| \log \left( 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\xi(s)}ds + \frac{4\alpha}{\alpha} \int_0^\gamma e^{-2\zeta(s)}(e^{4s/\alpha} - 1)ds \right) 
- \log \left( 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\zeta(s)}ds + \frac{4\alpha}{\alpha} \int_0^\gamma e^{-2\xi(s)}(e^{4s/\alpha} - 1)ds \right) \right| 
\leq \sup_{x \in [0,1]} \int_0^\gamma \left| e^{-2\xi(s)} - e^{-2\zeta(s)} \right| ds + x \int_0^\gamma e^{4s/\alpha^*} \left| e^{-2\xi(s)} - e^{-2\zeta(s)} \right| ds
\leq 2e^{-2cT}(1 + e^{4T/\alpha^*})\|\xi - \zeta\|_T.
\]

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 of equation (32) goes to zero. From equation (11), this also implies that the second term goes to zero with probability one. As a conclusion, we necessarily have that \( \mathcal{H}(\phi) = \gamma \) which proves the lemma. \( \square \)

**Lemma C.9.** Equation (28) 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 (28) has to be interpreted as a differential equation, which can be done using equation (13). We prove this lemma in Appendix D.4.

The proof of Proposition 5.1 is now elementary. From Lemma C.6, the family \( (\phi_d)_{d \in \mathbb{N}} \) is compact in \( C_T \). Thus, it admits at least one subsequential limit by Arzelà-Ascoli theorem. From Lemma C.7, such a limit verifies \( \mathcal{H}(\phi) = \gamma \) 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:

\[
\left\{ \sup_{\xi \in B_c} \left\| \mathcal{H}(\xi) - \mathcal{H}(\zeta) \right\|_{_{T \to \infty}} \to 0 \right\},
\]

\( (\phi_d)_{d \in \mathbb{N}} \) 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 \( T_n \to \infty \), 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 \( \phi \) is solution of equation (28). Replacing \( \mu \) from its definition in equation (11), leads to equation (13) 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 result will be compared to the convergence rates 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:

\[
F_d(\gamma) = 1 + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi_d(s)}ds, \quad G_d(\gamma) = 1 + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi_d(s)}e^{4sd/m_d^2}ds, \quad J_d(\gamma) = \frac{G_d(\gamma)}{F_d(\gamma)}.
\]

Since it has been shown that \( \phi_d \) 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:

\[
F(\gamma) = 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\phi(s)}ds, \quad G(\gamma) = 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}ds, \quad J(\gamma) = \frac{G(\gamma)}{F(\gamma)}.
\]
Introducing $\gamma$

Now, to compute the overlap, one has to compute the norms. Let the following lemma:

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

$$
\chi_d(\gamma) = \sqrt{\frac{m_d}{m_d^*}} e^{4\gamma_d/m_d^*} \frac{m_d^{-1} \text{Tr} \left( Y_d \left( I_{m_d} + (J_d(\gamma) - 1)Y_d \right)^{-1} \right)}{\sqrt{m_d^{-1} \text{Tr} \left( (I_{m_d} + (e^{4\gamma_d/m_d^*} - 1)Y_d )^2 (I_{m_d} + (J_d(\gamma) - 1)Y_d)^{-2} \right)}}. 
$$

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

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

$$
Z_d(t) = e^{-2\phi_d(t)} e^{2\text{Tr}(Z_d^*) te^{2Z_d^* t} W_d^0} \left( I_{m_d} + 4 \int_0^t e^{-2\phi_d(t)} e^{2\text{Tr}(Z_d^*) W_d^0} e^{4Z_d^* t} W_d^0 \right)^{-1} W_d^0 e^{2Z_d^* t}.
$$

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

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

with:

$$
N_d(\gamma) = \left( 1 + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi_d(s)} ds \right) I_{m_d} + \frac{4d}{m_d} \int_0^\gamma e^{-2\phi_d(s)} (e^{4sd/m_d^*} - 1) ds Y_d
$$

$$
= F_d(\gamma) I_{m_d} + (G_d(\gamma) - F_d(\gamma)) Y_d.
$$

Now using the fact that $U_d^* U_d^{0T}$ is a projection:

$$
\exp \left( \frac{2d\gamma}{m_d^*} U_d^* U_d^{0T} \right) = I_d - U_d^* U_d^{0T} + \exp \left( \frac{2d\gamma}{m_d^*} U_d^* U_d^{0T} \right).
$$

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

$$
\text{Tr} \left( Z_d^* Z_d(\gamma d) \right) = \frac{1}{m_d m_d^*} e^{-2\phi_d(\gamma)} \text{Tr} \left( U_d^{0T} U_d^* (I_d + (e^{2\lambda_d^* \gamma} - 1) U_d^* U_d^{0T}) U_d^{0T} N_d(\gamma)^{-1} U_d^{0T} (I_d + (e^{2\lambda_d^* \gamma} - 1) U_d^* U_d^{0T}) \right)
$$

$$
= \frac{1}{m_d m_d^*} e^{-2\phi_d(\gamma)} e^{4\lambda_d^* \gamma} \text{Tr} \left( Y_d (F_d(\gamma) I_{m_d} + (G_d(\gamma) - F_d(\gamma)) Y_d)^{-1} \right).
$$

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

$$
\|Z_d\|^2_F = \frac{1}{m_d^2} \text{Tr} \left( U_d U_d^{0T} U_d^* U_d^{0T} \right) = \frac{1}{m_d^2}.
$$

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

$$
\|Z_d(\gamma d)\|^2_F = \frac{1}{m_d^2} e^{-4\phi_d(\gamma)} \text{Tr} \left( (I_{m_d} + (e^{4\lambda_d^* \gamma} - 1) Y_d)^2 (F_d(\gamma) I_{m_d} + (G_d(\gamma) - F_d(\gamma)) Y_d)^{-2} \right).
$$

Assembling equation (36), (37) and (38) 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 $[0, T]$:

$$
\frac{1}{m_d} \text{Tr} \left( Y_d \left( I_{m_d} + (J_d(\gamma) - 1)Y_d \right)^{-1} \right) \xrightarrow{d \to \infty} \int \frac{x}{1 + (J(\gamma) - 1)x} d\mu(x) \equiv A(\gamma),
$$

$$
\frac{1}{m_d} \text{Tr} \left( (I_{m_d} + (e^{4\lambda_d^* \gamma} - 1) Y_d)^2 (I_{m_d} + (J_d(\gamma) - 1)Y_d)^{-2} \right) \xrightarrow{d \to \infty} \int \left( \frac{1 + (e^{4\gamma_d/m_d^*} - 1)x}{1 + (J(\gamma) - 1)x} \right)^2 d\mu(x) \equiv B(\gamma).
$$
As a consequence, uniformly on $[0, T]$:

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

(39)

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 gives the overlap in the high-dimensional limit. This depends on $\alpha, \alpha^*$ (also through the measure $\mu$ defined in equation (11)) 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.

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

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

This technical lemma is proven in [Appendix D.6]. It mainly uses the implicit equation on $\phi$ 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 the integrals $A(\gamma), B(\gamma)$ as $\gamma \to \infty$ (defined in [Lemma C.11]).

**Lemma C.13.** Let $A(\gamma), B(\gamma)$ be defined in [Lemma C.11]. Then:

\[
A(\gamma) \xrightarrow{\gamma \to \infty} \begin{cases} 
\min(\alpha, \alpha^*) + O\left(\frac{1}{J(\gamma)^2}\right) & \text{for } \alpha \neq \alpha^* \\
\frac{1}{J(\gamma)} + O\left(\frac{1}{J(\gamma)^{3/2}}\right) & \text{for } \alpha = \alpha^*.
\end{cases}
\]

and:

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

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:

\[
\chi(\gamma) \xrightarrow{\gamma \to \infty} \begin{cases} 
\sqrt{\frac{\alpha}{\alpha^*}} + O\left(e^{-4\gamma/\alpha^*}\right) & \text{for } \alpha < \alpha^* \\
1 + O\left(\sqrt{\gamma} e^{-2\gamma/\alpha^*}\right) & \text{for } \alpha = \alpha^* \\
1 + O\left(\gamma^{-2}\right) & \text{for } \alpha > \alpha^*.
\end{cases}
\]
Proof. The proof is a simple consequence of the two previous lemmas and equation \([39]\) giving the expression of \(\chi(\gamma)\). Starting with the case \(\alpha < \alpha^*\), we obtain:

\[
\chi(\gamma) = \lim_{\gamma \to \infty} \sqrt{\frac{\alpha}{\alpha^*} \frac{1 + O(J(\gamma) - 1)}{1 + O(J(\gamma) - 1)}} = \sqrt{\frac{\alpha}{\alpha^*} + O(J(\gamma) - 1)},
\]

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

\[
\chi(\gamma) = \lim_{\gamma \to \infty} \frac{1 + O(J(\gamma) - 1/2)}{\sqrt{1 + O(J(\gamma) - 1/2)}} = 1 + O(J(\gamma) - 1/2).
\]

Finally, if \(\alpha > \alpha^*\):

\[
\chi(\gamma) = \lim_{\gamma \to \infty} \sqrt{\frac{\alpha^*}{\alpha} \frac{1 + O(J(\gamma) - 1)}{\sqrt{1 - \frac{\alpha^*}{\alpha} \frac{J(\gamma)^2}{e^{\gamma^2/\alpha^*}} + \frac{\alpha^*}{\alpha} + O(J(\gamma) - 1)}}} = 1 + O(J(\gamma) - 2),
\]

which concludes the proof. \(\square\)

Figure 7 compares a simulated version of equation \([39]\) with the bounds obtained in the previous lemma. As expected, those bounds are tight since they were obtained using an exact asymptotic development.

![Figure 7](image-url)

Figure 7: Evolution of \(\delta(\gamma) = \chi = -\chi(\gamma)\) (where \(\chi = \lim_{\gamma \to \infty} \chi(\gamma)\)) as a function of \(\gamma\) for different values of \(\alpha, \alpha^*\). Dashed lines: convergence rates obtained in Lemma C.14. Simulated using equation \([39]\), with standard discretization for the integral computation \((\alpha \neq \alpha^*)\) and exact solution \((\alpha = \alpha^*)\).

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:

\[
\frac{\text{Tr}(Z(t)Z^*)}{\|Z(t)\|_F \|Z^*\|_F} = \frac{1}{2} \left( \frac{\|Z(t)\|_F}{\|Z^*\|_F} + \frac{\|Z^*\|_F}{\|Z(t)\|_F} - \frac{\|Z(t) - Z^*\|_F^2}{\|Z^*\|_F \|Z(t)\|_F} \right) = 1 + \Theta(\|Z(t) - Z^*\|_F)
\]

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:

\[
\frac{\text{Tr}(Z(t)Z^*)}{\|Z(t)\|_F \|Z^*\|_F} = 1 + O(e^{-2t/m^*}).
\]

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

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

with:

$$\gamma_k(t) = \int_0^t e^{-2s}e^{2s\text{Tr}(Z^*)}e^{4\mu_k s}ds \sim \frac{1}{4\mu_k}e^{-2s}e^{2s\text{Tr}(Z^*)}e^{4\mu_k t},$$

whenever $\mu_k > 0$. This is a consequence of Lemma B.3 along with the fact that $\psi(t) = \text{Tr}(Z(t)) \rightarrow \text{Tr}(Z^*)$ as $t \rightarrow \infty$. When $\mu_k = 0$, then $\gamma_k(t) = o(e^{-\beta t})$ for all $\beta > 0$. Using Courant-Fischer formula (Lemma B.2), we have for $j \in [1, \min(m, d)]$:

$$\lambda_j(t) = \max_{V \subset \mathbb{R}^m} \min_{x \in V} \frac{\langle \sum_{k=1}^{d} \gamma_k(t)z_kz_k^T \rangle}{\|x\|^2} = \min_{\|x\|^2=m-j+1} \max_{V \subset \mathbb{R}^m} \left( \sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2 \right).$$

We start by using the second equality and choose $V_1 = \text{Span}(z_1, \ldots, z_{j-1})$ which is of dimension $m - j + 1$ by assumption. Then:

$$\lambda_j(t) \leq \max_{x \in V_1} \frac{\sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2}{\|x\|^2}.$$

Since $\mu_1 \geq \cdots \geq \mu_m$, using the definition of the $\gamma_k(t)$ in equation (40), we have $\gamma_1(t) \geq \cdots \geq \gamma_d(t)$. Thus:

$$\lambda_j(t) \leq \gamma_j(t) \max_{x \in V_1} \frac{\sum_{k=1}^{d} (x \cdot z_k)^2}{\|x\|^2} \equiv C_j.$$

We now use the first equality of equation (41). We choose $V_2 = \text{Span}(z_1, \ldots, z_j)$ which is of dimension $j$ by assumption. Then:

$$\lambda_j(t) \geq \min_{x \in V_2} \frac{\sum_{k=1}^{d} \gamma_k(t)(x \cdot z_k)^2}{\|x\|^2} \geq \min_{x \in V_2} \frac{\sum_{k=1}^{j} \gamma_k(t)(x \cdot z_k)^2}{\|x\|^2} \geq \gamma_j(t) \min_{x \in V_2} \frac{\sum_{k=1}^{j} (x \cdot z_k)^2}{\|x\|^2} \equiv D_j.$$

By definition of $V_2$, $D_j$ cannot be zero. The definition of $\gamma_j(t)$ in equation (40) 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 $\phi_d$. Using equation (24) making the substitution $s \rightarrow s/d$:

$$\psi_d(t) = \frac{1}{4} \text{Tr} \log \left( I_{md} + \frac{4d}{m_d}U_d^T \int_0^{t/d} e^{-2\phi_d(s)} \exp \left( \frac{4sd}{m_d}U_d^*U_d^T \right) ds \right).$$
Therefore, introducing $\gamma = t/d$, $\lambda_d = d/m_d$ and $\lambda^*_d = d/m^*_d$:
\[
\phi_d(\gamma) + \gamma d = \frac{1}{4} \text{Tr} \log \left( I_{m_d} + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)} \left( I_{m_d} + (e^{4\lambda^*_d \gamma} - 1)U^T_d U_d^* T U^T_d U_d^* \right) ds \right),
\]
where we used that $\phi_d(\gamma) = \psi_d(\gamma d) - \gamma d$ (in our setup $\text{Tr}(Z_d^*) = 1$). This is precisely equation \ref{eq:phi_d_derivative}. We set $Y_d = U^T_d U_d^* T U_d^* U_d^*$ and:
\[
F_d(\gamma) = 1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)} ds, \quad G_d(\gamma) = 1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)} e^{4\lambda^*_d \gamma} ds,
\]
so that:
\[
\phi_d(\gamma) + \gamma d = \frac{1}{4} \text{Tr} \log \left( F_d(\gamma)(1 - Y_d) + G_d(\gamma)Y_d \right).
\]
Deriving with respect to $\gamma$:
\[
\dot{\phi}_d(\gamma) + d = \frac{1}{4} \text{Tr} \left[ U_d(\gamma, Y_d) \right], \quad \text{(42)}
\]
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:
\[
G_d(\gamma) = e^{4\lambda^*_d \gamma} F_d(\gamma) - 4\lambda^*_d \int_0^\gamma e^{4\lambda^*_d \gamma} F_d(s) ds \leq e^{4\lambda^*_d \gamma} F_d(\gamma).
\]
Thus:
\[
\dot{F}_d(\gamma)G_d(\gamma) - F_d(\gamma)\dot{G}_d(\gamma) = 4\lambda_d e^{-2\phi_d(\gamma)} (G_d(\gamma) - F_d(\gamma)e^{4\lambda^*_d \gamma}) \leq 0.
\]
As a consequence, for every $\gamma \geq 0$, the map $x \mapsto U_d(\gamma, x)$ is non-decreasing. Using equation \ref{eq:phi_d_derivative} and the fact that the eigenvalues of $Y_d$ are contained in $[0, 1]$:
\[
\frac{m_d \dot{F}_d(\gamma)}{4F_d(\gamma)} \leq \dot{\phi}_d(\gamma) + d \leq \frac{m_d \dot{G}_d(\gamma)}{4G_d(\gamma)}.
\]
Using the bound on $\dot{\phi}_d$ obtained in Lemma C.5 we obtain:
\[
\frac{\dot{F}_d(\gamma)}{F_d(\gamma)} \leq 4\kappa \frac{\sqrt{d}}{m_d} + 4\lambda_d, \quad \frac{\dot{G}_d(\gamma)}{G_d(\gamma)} \geq -4\kappa \frac{\sqrt{d}}{m_d} + 4\lambda_d.
\]
This can be integrated to obtain a bound on $\phi_d$:
\[
\left( 1 - \frac{\kappa}{\sqrt{d}} \right) e^{-\lambda^*_d \gamma} \exp \left( 4\lambda_d \left( 1 - \frac{\kappa}{\sqrt{d}} \right) \gamma \right) \leq e^{-2\phi_d(\gamma)} \leq \left( 1 + \frac{\kappa}{\sqrt{d}} \right) \exp \left( 4\lambda_d \left( 1 + \frac{\kappa}{\sqrt{d}} \right) \gamma \right).
\]
Since $\lambda_d$ and $\lambda^*_d$ 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 \ref{eq:U_d_def}, using the definition of $F_d(\gamma)$ and $G_d(\gamma)$:
\[
U_d(\gamma, x) = 4\lambda_d e^{-2\phi_d(\gamma)} \frac{1 - x + xe^{4\lambda^*_d \gamma}}{1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}(1 - x + xe^{4\lambda^*_d \gamma}) ds} = v_d(\gamma, x).
\]
Deriving with respect to $\gamma$, for $0 \leq x \leq 1$:
\[
\left| \frac{\partial v_d}{\partial \gamma}(\gamma, x) \right| \leq \left| \frac{4\lambda^*_d xe^{4\lambda^*_d \gamma}}{1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}(1 - x + xe^{4\lambda^*_d \gamma}) ds} \right| + \left| \frac{4\lambda_d e^{-2\phi_d(\gamma)}(1 - x + xe^{4\lambda^*_d \gamma})^2}{(1 + 4\lambda_d \int_0^\gamma e^{-2\phi_d(s)}(1 - x + xe^{4\lambda^*_d \gamma}) ds)^2} \right|
\leq 4\lambda^*_d e^{4\lambda^*_d T} + 4\lambda_d e^{-2\phi_d(\gamma)} e^{8\lambda^*_d T}.
\]

Thus, using equation (42) and rearranging:
\[
\phi_d(\gamma) = -\frac{1}{2} \log \left(1 + \frac{\dot{\phi}_d(\gamma)}{d}\right) + \frac{1}{2} \log \left(\frac{1}{m_d \text{Tr}[v_d(\gamma, Y_d)]}\right).
\]

Now:
\[
\left|\dot{T}_d(\gamma)\right| \leq \sup_{x \in [0,1]} \left|\frac{\partial v_d}{\partial \gamma}(\gamma, x)\right|.
\]

Since \(\phi_d\) is uniformly bounded in \(d\) and on \([0, T]\) from the previous step, then from equation (43), 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:
\[
|\phi_d(\gamma) - \phi_d(\gamma')| \leq \frac{1}{2} \left|\log \left(1 + \frac{\epsilon_d(\gamma') - \epsilon_d(\gamma)}{1 + \epsilon_d(\gamma)}\right)\right| + \frac{1}{2} \left|\log \left(1 + \frac{T_d(\gamma') - T_d(\gamma)}{T_d(\gamma)}\right)\right|.
\]

Moreover:
\[
T_d(\gamma) \geq \inf_{\gamma, \gamma' \in [0,1]} v_d(\gamma, x) \geq \left(1 + \frac{L_d}{\lambda_d} e^{-2e^{4\lambda_d T}}\right)^{-1},
\]

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:
\[
\left|\frac{T_d(\gamma) - T_d(\gamma')}{T_d(\gamma)}\right| \leq \frac{L}{\rho} |\gamma - \gamma'|,
\]

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 \(|\log(1 + x)| \leq 2|x|\) for \(x \in [-1/2, 1/2]\), the second term of equation (44):
\[
|\log \left(1 + \frac{T_d(\gamma') - T_d(\gamma)}{T_d(\gamma)}\right)| \leq \frac{2L}{\rho} \eta.
\]

For the first term, we use the bound obtained in Lemma C.5 which implies that:
\[
\sup_{\gamma \in [0, T]} |\epsilon_d(\gamma)| \leq \frac{\kappa}{\sqrt{d}},
\]

Thus, for \(d \geq \kappa^2 \max(25, (1 + 2/\eta)^2) = d_0(\eta)\), we obtain:
\[
|\log \left(1 + \frac{\epsilon_d(\gamma') - \epsilon_d(\gamma)}{1 + \epsilon_d(\gamma)}\right)| \leq 2\eta,
\]

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

We need to be careful since the index \(d_0\) from which the bound is valid depends on \(\eta\). To do so, let \(\delta > 0\) and \(\eta_0 = (1 + L/\rho)^{-1}\delta\). For each \(d = 1, \ldots, d_0(\eta_0)\), the map \(\phi_d\) is uniformly continuous on \([0, T]\) and we have \(\eta_d\) such that:
\[
\sup_{\gamma, \gamma' \in [0, T], |\gamma - \gamma'| \leq \eta_d} |\phi_d(\gamma) - \phi_d(\gamma')| \leq \delta.
\]

Now for \(\eta \leq \min(\eta_0, \ldots, \eta_{d_0(\eta_0)})\), we have:
\[
\sup_{d \in \mathbb{R}} \sup_{\gamma, \gamma' \in [0, T], |\gamma - \gamma'| \leq \eta} |\phi_d(\gamma) - \phi_d(\gamma')| \leq \max \left[\max_{1 \leq d \leq d_0(\eta_0)} \sup_{\gamma, \gamma' \in [0, T], |\gamma - \gamma'| \leq \eta} |\phi_d(\gamma) - \phi_d(\gamma')|, \sup_{d \geq d_0(\eta_0)} \sup_{\gamma, \gamma' \in [0, T], |\gamma - \gamma'| \leq \eta} |\phi_d(\gamma) - \phi_d(\gamma')|\right]
\]
\[
\leq \max \left[\delta, \eta_0 \left(1 + \frac{L}{\rho}\right)\right]
\]
\[
\leq \delta,
\]

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:

$$\mathcal{H}_d(\xi)(\gamma) = \frac{1}{m_d} \text{Tr}(U(\sigma_d, \xi, \gamma, Y_d)), \quad \mathcal{H}(\xi)(\gamma) = \int U(\sigma, \xi, \gamma, x) d\mu(x),$$

with:

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

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

$$|\mathcal{H}_d(\xi)(\gamma) - \mathcal{H}(\xi)(\gamma)| \leq \frac{1}{m_d} \text{Tr} \left| U(\sigma_d, \xi, \gamma, Y_d) - U(\sigma, \xi, \gamma, Y_d) \right| + \frac{1}{m_d} \text{Tr} U(\sigma, \xi, \gamma, Y_d) - \int U(\sigma, \xi, \gamma, x) d\mu(x). \quad (45)$$

We start by the first term. We have:

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

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 $a, \tilde{a} > 0$:

$$\left| \frac{1}{4\sigma_1} \log (1 + 4\sigma_1 a) - \frac{1}{4\tilde{\sigma}_1} \log (1 + 4\tilde{\sigma}_1 \tilde{a}) \right| \leq \left| \frac{1}{4\sigma_1} \log (1 + 4\sigma_1 a) - \frac{1}{4\tilde{\sigma}_1} \log (1 + 4\tilde{\sigma}_1 \tilde{a}) \right|$$

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

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

$$|U(\sigma, \xi, \gamma, x) - U(\tilde{\sigma}, \xi, \gamma, x)| \leq \frac{2a}{\sigma_0} |\sigma_1 - \tilde{\sigma}_1| + x \int_0^\gamma e^{-2\xi(s)} |e^{4\sigma_2 s} - e^{4\tilde{\sigma}_2 s}| ds$$

$$\leq \frac{2a}{\sigma_0} |\sigma_1 - \tilde{\sigma}_1| + 4x|\sigma_2 - \tilde{\sigma}_2| \int_0^\gamma e^{-2\xi(s)} e^{4\sigma_2 s} ds,$$

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

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

Now using this result with $\sigma_d = (d/m_d, d/m_d^*)$ we have $\sigma = (\alpha^{-1}, \alpha^{*-1})$, 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 $\sigma_0, \sigma_0^*$. Finally:

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

which proves that the first term of equation (45) goes to zero uniformly in $\xi$ and $\gamma$. We look at the second term. Using the bounds, for $\xi \in B_c$ and $\gamma \in [0, T]$:}

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

$$4\sigma_1 \int_0^\gamma e^{-2\xi(s)} (e^{4\sigma_2 s} - 1) ds \leq \frac{\sigma_1}{\sigma_2} e^{-2c e^{4\sigma_2 T}} \equiv B(\sigma),$$
The goal is now to study equation (47). We now set: continuous solution

Thus, if we manage to show that the solution of equation (47) is unique, this will imply the uniqueness of the initial condition, then it is easily shown that it is also solution of equation (46). Provided that

From the expression of \( \gamma \) in Proposition 5.1:

Using the definition of \( \mu \) in equation (11), it is easily shown that equations (28) and (13) are equivalent. Following the second one, we have:

with:

From the relationships:

D.4 Proof of Lemma C.9

Using the definition of \( \mu \) in equation (11), it is easily shown that equations (28) and (13) are equivalent. Following the second one, we have:

with:

From the expression of \( \Theta \) in Proposition 5.1

it is easy to see that \( \Theta \) is twice continuously differentiable on \([-1, \infty[\), and that \( \Theta'(u) \geq 0 \) for all \( u > -1 \). We now suppose that \( \phi \) is a continuous solution of equation (46). Then, since \( J_\phi(\gamma) \geq 1 \), we can differentiate with respect to \( \gamma \):

From the relationships:

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

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

Thus, if we manage to show that the solution of equation (47) is unique, this will imply the uniqueness of the continuous solution \( \phi \). Reciprocally, suppose that we have a couple \((F(\gamma), J(\gamma))\) solution of equation (47) with the initial condition, then it is easily shown that it is also solution of equation (46). Provided that \( \hat{F}(\gamma) > 0 \) for all \( \gamma \geq 0 \), we will obtain the existence of a solution:

The goal is now to study equation (47). We now set:

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

\[
\sup_{(F, J, \gamma) \in \mathcal{V}} \left\| \nabla \Omega(F, J, \gamma) \right\| < \infty.
\]

Which proves that \( \Omega \) is locally Lipschitz continuous on \( D \). Thus, from Picard–Lindelöf theorem, the solution \( F(\gamma), J(\gamma) \) of the differential system \([47]\) with initial condition \( F(0) = J(0) = 1 \) is 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 \([47]\), 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, \( F_M(\gamma) \) remains of the same sign as \( F_M(\gamma) \). Since \( F_M(0) = 1 \), we get that \( F_M(\gamma) > 0 \) and \( F_M \) is non-decreasing. Now, from the equation:

\[
\dot{J}_M(\gamma) = \left(e^{4\gamma/\alpha^*} - J_M(\gamma)\right) \frac{\dot{F}_M(\gamma)}{F_M(\gamma)},
\]

we get:

\[
J_M(\gamma) = e^{4\gamma/\alpha^*} - \frac{4}{\alpha^*F_M(\gamma)} \int_0^\gamma F_M(s)e^{4s/\alpha^*}ds < e^{4\gamma/\alpha^*},
\]

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 \( e^{4T/\alpha^*} \). Since \( J_M(\gamma) \) remains positive on \([0, T]\) and \( \Gamma(J) \geq 0 \) for \( J \geq 0 \), this implies that:

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

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 \([47]\):

\[
F_M(\gamma) = \exp \left(4 \int_0^\gamma \frac{ds}{\alpha + \Gamma(J_M(s))(e^{4s/\alpha^*} - J_M(s))}\right) \leq e^{4\gamma/\alpha^*},
\]

and a contradiction. Thus necessarily \( T = \infty \) and the system \([47]\) admits a unique solution defined on all \( \mathbb{R}^+ \). From the equivalence between equations \([46]\) and \([47]\), this concludes the proof.

**Simulating the solution of equation \([13]\)** This proof also explains how it is possible to simulate the function \( \phi(\gamma) \) from the implicit equation \([13]\). First, the solutions \( F(\gamma) \) and \( J(\gamma) \) can be integrated by discretizing equation \([47]\). Then, it is possible to deduce the value of \( \phi(\gamma) \) from equation \([48]\).

**D.5 Proof of Lemma \([C.11]\)**

As proven in **Proposition 5.1** \( \phi_d \) uniformly converges on \([0, T]\) towards a function \( \phi \). Thus we define:

\[
F(\gamma) = 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\phi(s)}ds, \quad G(\gamma) = 1 + \frac{4}{\alpha} \int_0^\gamma e^{-2\phi(s)}e^{4s/\alpha^*}ds, \quad J(\gamma) = \frac{G(\gamma)}{F(\gamma)}.
\]

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 \( F_d(\gamma), F(\gamma) \geq 1: \)

\[
|J_d(\gamma) - J(\gamma)| \leq |F(\gamma)||G_d(\gamma) - G(\gamma)| + |G(\gamma)||F_d(\gamma) - F(\gamma)|.
\]

This proves that \( J_d \xrightarrow{d \to \infty} J \) uniformly on \([0, T]\). We now have two terms to study. For the first one, we set \( U(x, J) = \frac{x}{1 + (J - 1)x} \), so that:

\[
\left| \frac{1}{m_d} \text{Tr} \left( U(Y_d, J_d(\gamma)) \right) - \int U(x, J(\gamma))d\mu(x) \right| \leq \frac{1}{m_d} \text{Tr} \left| U(Y_d, J_d(\gamma)) - U(Y_d, J(\gamma)) \right|
\]

\[
+ \frac{1}{m_d} \text{Tr} \left( U(Y_d, J(\gamma)) \right) - \int U(x, J(\gamma))d\mu(x) \right|.
\]
Starting with the first term:
\[
\frac{1}{m_d} \text{Tr} \left| U(Y_d, J_d(\gamma)) \right| - U(Y_d, J(\gamma)) \right| \leq \sup_{x \in [0,1]} \left| U(x, J_d(\gamma)) \right| - U(x, J(\gamma)) \right| \leq |J_d(\gamma) - J(\gamma)|, \\
\]
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:
\[
\sup_{\gamma \in [0, T]} \left| \frac{1}{m_d} \text{Tr} \left( U(Y_d, J_d(\gamma)) \right) - \int U(x, J(\gamma)) \, d\mu(x) \right| \leq \sup_{\alpha \in [1, A]} \left| \frac{1}{m_d} \text{Tr} \left( U(Y_d, J_d(\gamma)) \right) - \int U(x, J_d(\gamma)) \, d\mu(x) \right|, \\
\]
which goes to zero following Lemma A.1. Therefore, the first convergence of Lemma C.11 is proven. 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 \(e_d(\gamma) = e^{4\gamma d/m_d^2}\) and \(e(\gamma) = e^{4\gamma/\alpha^*}\), we have:
\[
\left| \frac{1}{m_d} \text{Tr} \left( V(Y_d, J_d(\gamma), e_d(\gamma)) \right) - \int V(x, J(\gamma), e(\gamma)) \, d\mu(x) \right| \leq \frac{1}{m_d} \text{Tr} \left| V(Y_d, J_d(\gamma), e_d(\gamma)) - V(Y_d, J(\gamma), e(\gamma)) \right| \\
+ \frac{1}{m_d} \text{Tr} \left( V(Y_d, J(\gamma), e(\gamma)) \right) - \int V(x, J(\gamma), e(\gamma)) \, d\mu(x) \right|.
\]
Now:
\[
\left| \frac{1}{m_d} \text{Tr} \left( V(Y_d, J_d(\gamma), e_d(\gamma)) - V(Y_d, J(\gamma), e(\gamma)) \right) \right| \leq \sup_{x \in [0,1]} \left| V(x, J_d(\gamma), e_d(\gamma)) - V(x, J(\gamma), e(\gamma)) \right|.
\]
Moreover, for \(e, \hat{e} \in [1, E]\) and \(J, \hat{J} \geq 1\):
\[
\left| V(x, J, e) - V(x, \hat{J}, \hat{e}) \right| \leq \left| V(x, J, e) - V(x, \hat{J}, e) \right| + \left| V(x, \hat{J}, e) - V(x, \hat{J}, \hat{e}) \right| \\
\leq 2E^2|J - \hat{J}| + 2E|e - \hat{e}|.
\]
Applying this to our case, we have \(e_d(\gamma) = e^{4\gamma d/m_d^2} \leq e^{4\gamma d/m_d^2}\) stays bounded since \(d/m_d^2 \rightarrow \alpha^{-1}\) Thus, again, the first term of equation \([50]\) 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
\[
\sup_{\gamma \in [0, T]} \left| \frac{1}{m_d} \text{Tr} \left( V(Y_d, J_d(\gamma), e(\gamma)) \right) - \int V(x, J(\gamma), e(\gamma)) \, d\mu(x) \right| \leq \sup_{(J,e) \in K} \left| \frac{1}{m_d} \text{Tr} \left( V(Y_d, J, e) \right) - \int V(x, J, e) \, d\mu(x) \right|,
\]
which goes to zero. As a conclusion, the second convergence of Lemma C.11 is proven. Finally, for the overlap, write from equation \([35]\):
\[
\chi_d(\gamma) = \sqrt{\frac{md}{m_d}} e^{4\gamma d/m_d^2} \frac{A_d(\gamma)}{\sqrt{B_d(\gamma)}}, \\
\]
where \(A_d(\gamma) \rightarrow A(\gamma)\) and \(B_d(\gamma) \rightarrow B(\gamma)\) are the quantity displayed in the lemma, which converge uniformly from the previous reasoning. We write:
\[
\left| \chi_d(\gamma) - \chi(\gamma) \right| \leq \frac{A_d(\gamma)}{\sqrt{B_d(\gamma)}} \cdot \sqrt{\frac{m_d}{m_d}} \cdot e^{4\gamma d/m_d^2} - e^{4\gamma/\alpha^*} \left| \frac{1}{\sqrt{B_d(\gamma)}} A_d(\gamma) - A(\gamma) \right| + \sqrt{\frac{\alpha}{\alpha^*}} e^{4\gamma/\alpha^*} A(\gamma) \cdot \frac{1}{\sqrt{B_d(\gamma)}} - \frac{1}{\sqrt{B(\gamma)}}.
\]
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 $d \geq d_0$:

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

since the map $x \mapsto x^{-1/2}$ is Lipschitz on the intervals of the form $[b, \infty[$. As shown before $A_d(\gamma) \xrightarrow{d \to \infty} A(\gamma)$ and $B_d(\gamma) \xrightarrow{d \to \infty} B(\gamma)$ uniformly, so that the two last terms go to zero. Since $m_d^*/d$ and $m_d/d$ respectively converge to $\alpha^*, \alpha$, 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 [13] solved by $\phi$, we obtain:

$$
4\gamma = \min(\alpha, 1 - \alpha^*) \log F(\gamma) + (\alpha + \alpha^* - 1)^+ \log G(\gamma) + \Theta(J(\gamma) - 1), \quad (51)
$$

where $F, G, J$ are defined in equation [34] (and depend on $\phi$), and:

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

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

$$
\Theta(u) \xrightarrow{u \to \infty} \frac{1 - |\alpha - \alpha^*| - |\alpha + \alpha^* - 1|}{2} \log u + \Omega(\alpha, \alpha^*) + O\left(\frac{1}{u}\right), \quad (52)
$$

The proof is decomposed into two steps: we first show that $F(\gamma), G(\gamma), J(\gamma) \xrightarrow{\gamma \to \infty} \infty$. Then we use equation (51) to determine the asymptotic behaviour of $J$.

**D.6.1 Limit**

We now show that necessarily $F(\gamma), G(\gamma)$ and $J(\gamma)$ go to infinity as $\gamma \to \infty$. We recall that:

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

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 (51) 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 (51) and (52):

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

which writes:

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

Therefore, integrating by parts, one has:

$$
F(\gamma) = G(\gamma) e^{-4\gamma / \alpha^*} + \frac{4}{\alpha^*} \int_0^{\gamma} G(s) e^{-4s / \alpha^*} ds \geq \frac{4}{\alpha^*} \int_0^{\gamma} G(s) e^{-4s / \alpha^*} ds \xrightarrow{\gamma \to \infty} \infty. \quad (53)
$$

This proves that necessarily $F(\gamma)$ also goes to infinity as $\gamma \to \infty$. We now finally show that also $J(\gamma) \xrightarrow{\gamma \to \infty} \infty$.

Differentiating:

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

Simon Martin, Francis Bach, Giulio Biroli
thus \( J \) is non-decreasing. Suppose that \( J(\gamma) \) converges towards a finite value as \( \gamma \to \infty \). Thus, equation (51) gives:

\[
4\gamma = \min (\alpha, 1 - \alpha^*) \log F(\gamma) + (\alpha + \alpha^* - 1)^+ \log G(\gamma) + O(1).
\]

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 \( \alpha, \alpha^* \)) such that:

\[
4\gamma = \alpha \log F(\gamma) + C + o(1).
\]

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

\[
J(\gamma) = e^{4\gamma/\alpha^*} - \frac{4}{\alpha^* F(\gamma)} \int_0^\gamma F(s)e^{4s/\alpha^*} ds,
\]

we obtain that:

\[
J(\gamma) \xrightarrow{\gamma \to \infty} \frac{\alpha^*}{\alpha + \alpha^*} e^{4\gamma/\alpha^*}.
\]

Therefore, we necessarily have \( J(\gamma) \xrightarrow{\gamma \to \infty} \infty \).

### D.6.2 Asymptotics

We now prove the asymptotics on \( J \). Using equations (51) and (52), we obtain as \( \gamma \to \infty \):

\[
4\gamma = \min (\alpha, 1 - \alpha^*) \log F(\gamma) + (\alpha + \alpha^* - 1)^+ \log G(\gamma)
\]

\[
= \frac{1 - |\alpha - \alpha^*| - |\alpha + \alpha^* - 1|}{2} \log J(\gamma) + \Omega(\alpha, \alpha^*) + o(1)
\]

\[
= (\alpha - \alpha^*)^+ \log F(\gamma) + \min(\alpha, \alpha^*) \log G(\gamma) + \Omega(\alpha, \alpha^*) + o(1).
\]

Thus, setting:

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

we obtain that:

\[
G(\gamma) = \zeta F(\gamma)^{-\nu} e^{4\lambda_m \gamma \omega(\gamma)},
\]

where \( \omega(\gamma) \xrightarrow{\gamma \to \infty} 1 \). Using the relation between \( F \) and \( G \) in equation (53):

\[
F(\gamma) = \zeta F(\gamma)^{-\nu} e^{4(\lambda_m - \lambda^*) \gamma \omega(\gamma)} + 4\lambda^* \zeta \int_0^\gamma F(s)^{-\nu} e^{4(\lambda_m - \lambda^*) s \omega(s)} ds.
\]

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:

\[
G(\gamma) \xrightarrow{\gamma \to \infty} \zeta e^{4\gamma/\alpha^*}, \quad F(\gamma) \xrightarrow{\gamma \to \infty} \begin{cases} 
\frac{4\zeta \gamma}{\alpha^*} & \text{for } \alpha = \alpha^* \\
\zeta \frac{\alpha^*}{\alpha^* - \alpha} \exp \left( 4\gamma \frac{\alpha^* - \alpha}{\alpha^* \alpha} \right) & \text{for } \alpha < \alpha^*.
\end{cases}
\]

Therefore:

\[
J(\gamma) \xrightarrow{\gamma \to \infty} \begin{cases} 
\frac{\alpha^*}{4\gamma} e^{4\gamma/\alpha^*} & \text{for } \alpha = \alpha^* \\
\left( 1 - \frac{\alpha}{\alpha^*} \right) e^{4\gamma/\alpha^*} & \text{for } \alpha < \alpha^*.
\end{cases}
\]

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 (55):

\[
F(\gamma) = \zeta F(\gamma)^{-\nu} \omega(\gamma) + 4\lambda^* \zeta \int_0^\gamma F(s)^{-\nu} \omega(s) ds \\
\equiv \epsilon(\gamma)
\]
Therefore:

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

Therefore:

\[F(\gamma) \sim_{\gamma \to \infty} \left( \frac{4\zeta \alpha}{\alpha^2} \right)^{\alpha/\alpha^*} \cdot \]

Using equation (54), we are able to determine the behaviour of \( G \), and the one of \( J \):

\[G(\gamma) \sim_{\gamma \to \infty} \zeta \left( \frac{4\zeta \alpha}{\alpha^2} \gamma \right)^{\alpha/\alpha^* - 1} e^{\gamma/\alpha}, \quad J(\gamma) \sim_{\gamma \to \infty} \frac{\alpha^2}{\alpha \gamma} e^{\gamma/\alpha^*}.\]

### D.7 Proof of Lemma C.13

From the definition of \( \mu \) in equation (11) and those of \( A(\gamma) \), \( B(\gamma) \) in Lemma C.11, we have:

\[A(\gamma) = \frac{1}{\alpha} (\alpha + \alpha^*) - 1 + \frac{1}{J(\gamma)} + \int_{r_-}^{r_+} \frac{x}{1 + J(\gamma) - 1} w(x)dx,\]

\[B(\gamma) = \left( 1 - \frac{\alpha^*}{\alpha} \right)^+ + \frac{1}{\alpha} (\alpha + \alpha^* - 1)^+ e^{8\gamma/\alpha^*} J(\gamma)^{-2} + \int_{r_-}^{r_+} \left( \frac{1 + (e^{4\gamma/\alpha^*} - 1)x}{1 + J(\gamma) - 1} \right)^2 w(x)dx,\]

with:

\[w(x) = \frac{1}{2\pi \alpha} \sqrt{(r_+ - x)(x - r_-)}; \quad r_\pm = \left( \sqrt{\alpha(1 - \alpha^*)} \pm \sqrt{\alpha^*(1 - \alpha)} \right)^2.\]

The challenge is to compute the asymptotics of the integrals when \( J(\gamma) \to \infty \). For the first one, we compute:

\[
\left| \frac{1}{J(\gamma)} \int_{r_-}^{r_+} w(x)dx - \int_{r_-}^{r_+} \frac{x}{1 + (J(\gamma) - 1)x} w(x)dx \right| = \frac{1}{J(\gamma)^2} \int_{r_-}^{r_+} \frac{1 - x}{x + (1 - x)J(\gamma) - 1} w(x)dx.
\]

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

\[
\int_{r_-}^{r_+} \frac{1 - x}{x + (1 - x)J(\gamma) - 1} w(x)dx \leq \int_{r_-}^{r_+} \frac{1 - x}{x} w(x)dx < \infty,
\]

so that in this case:

\[
\int_{r_-}^{r_+} \frac{x}{1 + (J(\gamma) - 1)x} w(x)dx = \frac{1}{J(\gamma)} \int_{r_-}^{r_+} w(x)dx + O\left( \frac{1}{J(\gamma)^2} \right).
\]

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

\[A(\gamma) = \min(\alpha, \alpha^*) + O\left( \frac{1}{J(\gamma)^2} \right),\]

whenever \( r_- = 0 \) (i.e. \( \alpha = \alpha^* \)), this is not possible anymore. In this case, the right integral of equation (57) writes:

\[
\frac{1}{2\pi \alpha J(\gamma)} \int_0^{r_+} \sqrt{\frac{r_+ - x}{x} x J(\gamma) + 1 - x}.\]
We change variables and let $x = r_+ \frac{t^2}{1 + t^2}$. Decomposing in partial fractions:

\[
\frac{1}{2\pi} \int_0^{r_+} \sqrt{\frac{r_+ - x}{x}} \frac{dx}{xJ(\gamma) + 1 - x} = r_+ \frac{1}{\pi} \int_0^{\infty} \frac{dt}{(1 + t^2)(1 + (1 - r_+ + r_+J(\gamma))^2)}
\]

\[
= \frac{1}{\pi} \left( -\frac{1}{J(\gamma) - 1} \int_0^{\infty} \frac{dt}{1 + t^2} + \frac{1 + r_+J(\gamma) - 1}{J(\gamma) - 1} \int_0^{\infty} \frac{dt}{(1 + (1 - r_+ + r_+J(\gamma))^2)} \right)
\]

\[
= -\frac{1}{2J(\gamma) - 1} + \frac{1}{2} \sqrt{1 + r_+J(\gamma) - 1} = O\left(J(\gamma)^{-1/2}\right),
\]

so that we get:

\[
\int_{r_-}^{r_+} \frac{x}{1 + (J(\gamma) - 1)x} w(x)dx = \frac{1}{\gamma \rightarrow \infty J(\gamma)} \int_{r_-}^{r_+} w(x)dx + O\left(\frac{1}{J(\gamma)^{3/2}}\right),
\]

which proves the result for $A(\gamma)$ using equation [56]. Now for $B(\gamma)$:

\[
\int_{r_-}^{r_+} \left( \frac{1 + (e^{4\gamma/\alpha^*} - 1)x}{1 + (J(\gamma) - 1)x} \right)^2 w(x)dx = \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2} \int_{r_-}^{r_+} w(x)dx = \frac{e^{8\gamma/\alpha^*}}{J(\gamma)^2} \left[ (e^{-4\gamma/\alpha^*} - J(\gamma)^{-1}) I_1(\gamma) \right.
\]

\[
+ \left. \left( e^{-8\gamma/\alpha^*} - J(\gamma)^{-2} \right) I_2(\gamma) \right],
\]

with:

\[
I_1(\gamma) = \int_{r_-}^{r_+} \frac{2x(1 - x)}{(x + (1 - x)J(\gamma)^{-1})^2} w(x)dx,
\]

\[
I_2(\gamma) = \int_{r_-}^{r_+} \frac{(1 - x)^2}{(1 + (1 - x)J(\gamma)^{-1})^2} w(x)dx.
\]

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

\[
I_1(\gamma) \leq \int_{r_-}^{r_+} \frac{2(1 - x)}{x} w(x)dx < \infty,
\]

\[
I_2(\gamma) \leq \int_{r_-}^{r_+} \frac{(1 - x)^2}{x^2} w(x)dx < \infty,
\]

so that:

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

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 technique as before. We have:

\[
I_1(\gamma) = \frac{J(\gamma)^2}{\pi \alpha} \int_0^{r_+} \frac{\sqrt{x(r_+ - x)} dx}{(1 - x + xJ(\gamma))^2}
\]

\[
= \frac{2r_+^2 J(\gamma)^2}{\pi \alpha} \int_0^{\infty} \frac{t^2 dt}{(1 + t^2)(1 + (1 - r_+ + r_+J(\gamma))^2)}
\]

\[
= \frac{r_+^2}{2\alpha} \sqrt{1 + 1 - r_+ + r_+J(\gamma)} \left( \frac{1}{\sqrt{1 - r_+ + r_+J(\gamma)} + 1} \right)^2
\]

\[
\sim \frac{r_+^2}{2\alpha} J(\gamma)^{1/2}.
\]

Since $2I_2(\gamma) \leq I_1(\gamma)$, the last term in equation [58] is negligible and from the expression of $r_+$:

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

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 [56].
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, 4, 6), we used a stepsize $\eta = 10^{-2}$ and initialized the weights $W^0$ from a Gaussian distribution (Figures 4, 6) 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, 4) 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 7) was simulated from equation (39) 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 $9 \times 10^4$ 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.