From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro

▶ **To cite this version:**

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. 2023. hal-04019712

HAL Id: hal-04019712
https://hal.science/hal-04019712

Preprint submitted on 8 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks

Luca Arnaboldi[1], Ludovic Stephan[1], Florent Krzakala[1], and Bruno Loureiro[2]

[1]École Polytechnique Fédérale de Lausanne (EPFL), IdePHICS Lab, CH-1015 Lausanne, Switzerland
[2]Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France
{luca.arnaboldi, ludovic.stephan, florent.krzakala}@epfl.ch, bruno.loureiro@di.ens.fr

February 14, 2023

arXiv:2302.05882v1 [stat.ML] 12 Feb 2023

Abstract

This manuscript investigates the one-pass stochastic gradient descent (SGD) dynamics of a two-layer neural network trained on Gaussian data and labels generated by a similar, though not necessarily identical, target function. We rigorously analyse the limiting dynamics via a deterministic and low-dimensional description in terms of the sufficient statistics for the population risk. Our unifying analysis bridges different regimes of interest, such as the classical gradient-flow regime of vanishing learning rate, the high-dimensional regime of large input dimension, and the overparameterised "mean-field" regime of large network width, covering as well the intermediate regimes where the limiting dynamics is determined by the interplay between these behaviours. In particular, in the high-dimensional limit, the infinite-width dynamics is found to remain close to a low-dimensional subspace spanned by the target principal directions. Our results therefore provide a unifying picture of the limiting SGD dynamics with synthetic data.

1 Introduction

A detailed understanding of the performance of stochastic gradient descent (SGD) in neural network is a major endeavour in machine learning, and significant progress was achieved in the context of large two-layers neural networks. In particular the optimisation over wide two-layer neural networks can be rigorously studied using a well defined partial differential equation (PDE) [1–4]. A consequence of these results is the global convergence of overparametrised two-layer networks towards perfect learning provided that the number of hidden neurons is large, the learning rate is sufficiently small, and enough data is at disposal. This line of work is commonly referred to as the *mean-field limit* of neural networks. The phenomenology in this regime was also studied for synthetic data with simple target functions by [5, 6].

Interestingly, the SGD dynamics of two-layer neural networks trained on synthetic Gaussian data was considered as early as in the seminal work of [7–9], and has witnessed a renewal of activity over the last few years [10–13]. However, differently from the mean-field limit, these works investigated the opposite limit of *fixed* hidden layer width and *diverging* data dimension, and studied the limiting SGD dynamics through a set of ordinary differential equations (ODEs).

Given these different limits, one may naturally wonder what is the relation, if any, between these sets of works. More generally, given data in dimension d and a two-layer network with p hidden units trained by SGD with a learning rate γ, one might inquire about the different regimes beside the mean-field ($p \to \infty$) and high-dimensional ($d \to \infty$) ones. This is the question investigated in this work. We consider a two-layer network trained on Gaussian data and labels given by a similar, though not necessarily identical, two-layer neural network target (hereafter also referred to as the *teacher*), and investigate the one-pass stochastic gradient descent (SGD) dynamics as a function of the relevant parameters d, p and γ. As summarised in Fig.1, we show that as long as $\gamma/dp \to 0^+$, a unifying deterministic description can be provided. In particular, our description recovers all the previously studied limits (mean field, high-dimensional and the classical gradient flow regime) and builds a bridge between them in a unified formalism. Namely, our **main contributions** are:

- Starting from the general approach of [7, 9], we build on the non-asymptotic results [12] and show how it allows to describe the entire phase space in Fig.1, that interpolates between the high-dimensional, classical, and mean-field limits.
- We unveil a remarkable dimension independence in the classical gradient-flow limit: once the initial conditions are given, the dynamics in terms of the sufficient statistics turns out to be *entirely* independent from the data dimension d.

- We explicitly construct the mean-field solution starting from the ODEs in the mean-field regime, bridging the hitherto different worlds of [7, 9] with the mean-field "hydrodynamic" approach. More precisely, for $p \to \infty$, we show how the ODEs simplify and give rise to a mean-field PDE, thanks to a decoupling of the learning dynamics that is found to remain close to a low-dimensional subspace spanned by the target principal directions.
- We further discuss how the SGD dynamics in the mean-field regime behave differently whether the input dimension d is small or large, leading to different finite-d and high-d dimensionless PDEs. This generalizes the recent "dimension-free" results in [6, 14].
- We provide a numerical solver for these equations. A GitHub repository with the code employed in the present work is available on [https://github.com/IdePHICS/DimensionlessDynamicsSGD]. Additionally, we discuss the interesting case of quadratic activation that allows to drastically reduce the complexity of the ODEs & PDEs.

Related work — Stochastic gradient descent was first introduced [15] as a stochastic approximation method, and later applied as an approximation to population risk minimization in [16, 17]. Its properties have been extensively studied for finite learning rate and input dimension in the strongly convex setting [18–21], and more recently in convex problems in the interpolating regime [22–25], to cite a few. High-dimensional limits of SGD were studied in [10, 26] for non-convex, single-index models. Recently, [13] has generalised and abstracted this discussion.

In the context of two-layer neural networks, the high-dimensional limit of SGD draws back from the seminal work of [7–9] and was subsequently studied by many authors under different settings [11, 27–32]. The infinite-width (a.k.a. mean-field) limit of the SGD dynamics of two-layer neural networks was studied by [1–4], who proved global convergence under certain conditions on the architecture and initialization. A bridge between these two limits was discussed by [12], who studied the joint limit where the hidden-layer width and the learning rate scale with the diverging input dimension.

Closer to us, dimension-free limits of the mean-field equations have been derived by [6] for low-dimensional target functions in the hypercube and by [14] for ReLU networks when the target is invariant under certain symmetries. [33] has proven global convergence of the gradient flow dynamics at finite width for orthogonal input data.

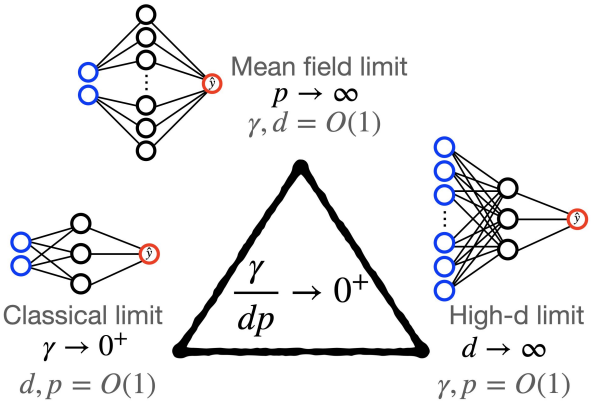


Figure 1: We discuss a unifying low-dimensional description of the one-pass SGD dynamics eq. (3) for two-layers neural networks as $\gamma/dp \to 0^+$. This includes, in particular, the mean-field ($p \to \infty$), the high-dimensional ($d \to \infty$) and the classical gradient flow ($\gamma \to 0$) limits.

2 Setting

In this manuscript we consider a supervised learning regression task where we are given n independent samples $(\boldsymbol{x}^\nu, y^\nu)_{\nu \in [n]} \in \mathbb{R}^{d+1}$ from a probability distribution ρ. We are interested in the problem of learning the training data with a (fully-connected) two-layer neural network:

$$f_\Theta(\boldsymbol{x}) = \frac{1}{p} \sum_{i=1}^{p} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}), \tag{1}$$

where $\Theta = (\boldsymbol{a}, \mathrm{W}) \in \mathbb{R}^{p(d+1)}$ denote the trainable parameters and $\sigma : \mathbb{R} \to \mathbb{R}$ the activation function. Since one of our goals is to connect with this line of work, for convenience we have adopted the mean-field normalisation [1–4]. As usual, training is performed via *empirical risk minimisation*, where the statistician chooses a *loss function* $\ell : (f_\Theta(\boldsymbol{x}), y) \in \mathbb{R}^2 \mapsto \ell(f_\Theta(\boldsymbol{x}), y) \in \mathbb{R}_+$ penalising deviations from the true labels and optimises the training parameters Θ by minimising the loss over the training data. As it is common in regression, we use the square loss $\ell(\hat{y}, y) := 1/2(\hat{y} - y)^2$, and focus on the *generalisation error* or *population risk*:

$$\mathcal{R}(\Theta) := \mathbb{E}_{(\boldsymbol{x},y)\sim\rho} \left[\frac{1}{2} (f_\Theta(\boldsymbol{x}) - y)^2 \right] \tag{2}$$

Training algorithm: This empirical risk minimisation problem being non-convex, different optimisation algorithms might reach different minima. Our goal is to characterise the training dynamics of two-layer networks under *one-pass stochastic gradient descent*:

$$\Theta^{\nu+1} = \Theta^{\nu} - \gamma \nabla_{\Theta} \ell(f_{\Theta^{\nu}}(\boldsymbol{x}^{\nu}), y^{\nu}), \qquad \nu \leq n \tag{3}$$

Note that in one-pass SGD, a fresh sample of data is used to estimate the gradient at each step, and therefore the quantity of data seen by the algorithm coincides with the number of steps. In particular, this means that at each step ν we have a random unbiased estimation of the population gradient which is uncorrelated to the previous step, defining a Markov chain. It is useful to rewrite eq. (3) by making explicit the effective noise of the process:

$$\Theta^{\nu+1} = \Theta^{\nu} - \gamma \nabla_{\Theta} \mathcal{R}(\Theta^{\nu}) + \gamma \varepsilon_{\nu}, \qquad \varepsilon_{\nu} \coloneqq \nabla_{\Theta} \left[\mathcal{R}(\Theta^{\nu}) - \ell(f_{\Theta^{\nu}}(\boldsymbol{x}^{\nu}), y^{\nu}) \right] \tag{4}$$

which is a zero mean random variable. Therefore, characterising the training dynamics of one-pass SGD translates into characterising this stochastic process.

Data model: Stochasticity in eq. (3) is induced by the draw of samples $(\boldsymbol{x}, y) \sim \rho$. In the following, we will assume that the input $\boldsymbol{x}^{\nu}$ are Gaussian $\boldsymbol{x}^{\nu} \sim \mathcal{N}\left(\mathbf{0}_d, {}^1\!/_d \mathrm{I}_d\right)$ while y^{ν} is drawn from the following generative model:

$$y^{\nu} = \frac{1}{k} \sum_{r=1}^{k} a_r^{\star} \sigma^{\star}(\boldsymbol{w}_r^{\star \top} \boldsymbol{x}^{\nu}) + \sqrt{\Delta} z^{\nu}, \qquad z^{\nu} \sim \mathcal{N}(0,1) \tag{5}$$

In other words, the target function $f_{\Theta^{\star}}$ is itself a two-layers neural network with parameters $\Theta^{\star} = (\boldsymbol{a}^{\star}, \mathrm{W}^{\star}) \in \mathbb{R}^{k(d+1)}$ and activation $\sigma^{\star}$. This setting, commonly refereed to as the *teacher-student* scenario, provides a rich data model for studying generalisation, and has been employed both in the analysis of one-pass SGD [9, 11, 12] but also more broadly in high-dimensional statistics [34]. In particular, we will be mostly interested in the realisable scenario where $k \leq p$, and therefore the minimum of the population risk in eq. (2) is achieved by perfectly learning the target.

Technical assumptions: We assume that the SGD dynamics remains in a bounded subset of $\mathbb{R}^{p \times d}$:

Assumption A1. *On an event with high probability, the SGD iterates are bounded in the following sense: for some $K > 0$, we have*

$$\forall i \in [p], \quad \|\boldsymbol{w}_i\|^2 \leq K. \tag{6}$$

This assumption can be easily checked on either the simulations or their deterministic approximations (see below), or otherwise enforced with a weight decay (as in [35]).

Assumption A2. *The student activation function σ is twice differentiable, with $\|\sigma^{(i)}\|_{\infty} \leq K$ for $i = 0, 1, 2$. The teacher activation $\sigma^{\star}$ is also upper bounded by K.*

Note that in some plots, we will sometimes use the function $\sigma(x) = x^2$, which does not satisfy this assumption. However, Assumption A1 ensures that we stay in a bounded subset of $\mathbb{R}^d$, hence we can replace σ by $\sigma \wedge K$ for K sufficiently large.

Simplifying assumptions: Since most of the interesting phenomenology happens at the hidden-layer, to lighten the discussion in the following we will focus on the case in which $a_r^{\star} = 1$ and $a_i^{\nu} = 1$ are fixed throughout learning and p is divisible by k. All of the discussion that follows can be readily generalised to the case in which $\boldsymbol{a}^{\nu}$ is learned and $p \geq k$ generically. We shall also assume that the teacher matrix $\mathrm{W}^{\star}$ is full rank; this can be avoided with a more careful definition of projections, but complicates the analysis.

3 The three limit regimes and their dimensionless description

As discussed before, the optimisation problem introduced in Sec. 2 defines a non-convex optimisation problem. Moreover, modern neural networks operate in a regime where both the data dimension d and the number of parameters in the network p are large. Therefore, characterising the evolution of the weights $\Theta^{\nu} \in \mathbb{R}^{p(d+1)}$ amounts to studying $p(d+1)$ non-linear, coupled, non-convex stochastic process - a challenging problem even for numerical methods. As motivated in the introduction Sec. 1, in this section we derive a *tractable, low-dimensional* description for SGD in different regimes of practical interest.

3.1 Main concepts

Sufficient statistics: A first observation is that the performance of the predictor $(\Theta^\nu)_{\nu\leq n}$ at iteration $\nu \leq n$ only depends on the statistics of the student and teacher pre-activations

$$\boldsymbol{\lambda}^\nu := \mathrm{W}^\nu \boldsymbol{x}^\nu \in \mathbb{R}^p, \quad \boldsymbol{\lambda}^{\star\nu} := \mathrm{W}^\star \boldsymbol{x}^\nu \in \mathbb{R}^k \tag{7}$$

Moreover, since $\boldsymbol{x}^\nu$ is Gaussian and independent from $(\mathrm{W}^\nu, \mathrm{W}^\star)$, the pre-activations are jointly Gaussian vectors $(\boldsymbol{\lambda}^\nu, \boldsymbol{\lambda}^{\star\nu}) \sim \mathcal{N}(\mathbf{0}_{p+k}, \Omega^\nu)$ with covariance:

$$\Omega^\nu := \begin{pmatrix} \mathrm{Q}^\nu & \mathrm{M}^\nu \\ \mathrm{M}^{\nu\top} & \mathrm{P} \end{pmatrix} = \begin{pmatrix} {}^1\!/_d \mathrm{W}^\nu \mathrm{W}^{\nu\top} & {}^1\!/_d \mathrm{W}^\nu \mathrm{W}^{\star\top} \\ {}^1\!/_d \mathrm{W}^\star \mathrm{W}^{\nu\top} & {}^1\!/_d \mathrm{W}^\star \mathrm{W}^{\star\top} \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)} \tag{8}$$

which is the *sufficient statistics* matrix for the population risk in eq. (2). Massaging eq. (3), we can derive a closed set of stochastic processes governing the evolution of the sufficient statistics:

$$\begin{aligned} M_{ir}^{\nu+1} - M_{ir}^\nu &= \frac{\gamma}{pd} \sigma'(\lambda_i^\nu) \lambda_r^{\star\nu}\, \mathcal{E}^\nu \\ Q_{ij}^{\nu+1} - Q_{ij}^\nu &= \frac{\gamma}{pd} \left(\sigma'(\lambda_i^\nu)\lambda_j^\nu + \sigma'(\lambda_j^\nu)\lambda_i^\nu \right) \mathcal{E}^\nu + \frac{\gamma^2\, ||\boldsymbol{x}^\nu||_2^2}{p^2 d^2} \sigma'(\lambda_i^\nu)\sigma'(\lambda_j^\nu)\, \mathcal{E}^{\nu 2} \end{aligned} \tag{9}$$

where we defined for convenience the displacement vector

$$\mathcal{E}^\nu := \frac{1}{k}\sum_{r=1}^k \sigma^\star(\lambda_r^{\star\nu}) - \frac{1}{p}\sum_{j=1}^p \sigma(\lambda_j^\nu) + \sqrt{\Delta} z^\nu. \tag{10}$$

Note that so far we have made no approximations: these equations are exact, and allow us to trade the $p(d+1)$ dimensional process for Θ^ν in Eq. (3) for a $p(k+p)$ dimensional process for $(\mathrm{M}^\nu, \mathrm{Q}^\nu)$. This can be particularly convenient if $d \gg k, p$.

The process in Eq. (9) has been previously studied in different limits and particular cases. For instance, [36] studied the stochastic dynamics in particular case of $k = p = 1$ and $\sigma(x) = x^2$, also known as *phase retrieval*, and [26] extended this discussion for arbitrary σ. More important to our work, [7, 9] has shown that this process admits a deterministic limit when $d \to \infty$ at fixed p, γ, characterized by the following ODE:

$$\frac{\mathrm{dM}}{dt} = \Psi^{(\mathrm{M})}(\Omega), \qquad \frac{\mathrm{dQ}}{\mathrm{d}t} = \Psi^{(\mathrm{GF})}(\Omega) + \frac{\gamma}{p}\Psi^{(\mathrm{Var})}(\Omega), \tag{SS-ODE}$$

where the right-hand side functions are defined by the following equations.

$$\begin{aligned} \Psi_{ir}^{(\mathrm{M})}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)} \left[\sigma'(\lambda_i)\lambda_r^\star\, \mathcal{E} \right] \\ \Psi_{ij}^{(\mathrm{GF})}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)} \left[\left(\sigma'(\lambda_i)\lambda_j + \sigma'(\lambda_j)\lambda_i \right) \mathcal{E} \right] \\ \Psi_{ij}^{(\mathrm{Var})}(\Omega) &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)} \left[\sigma'(\lambda_i)\sigma'(\lambda_j)\, \mathcal{E}^2 \right] \end{aligned} \tag{11}$$

As will be discussed later in Section 3.2, the superscript notation for the right-hand side is suggestive of their interpretation. This convergence was made rigorous by [11] and [12], who showed the following non-asymptotic result:

Theorem 3.1 ([12]). *We place ourselves under Assumptions A1-A2. Let Ω^ν be the random process of Eq. (9), and $\Omega(t)$ the solution to the ODE (SS-ODE) with starting point $\Omega(0) = \Omega^0$. Define the stepsize $\delta t = \gamma/pd$, and assume that $\gamma/p = O(1)$. Then there exists a constant $C > 0$ such that for any $\nu \geq 0$,*

$$\|\Omega^\nu - \Omega(\nu\delta t)\|_\infty \leq e^{C\nu\delta t}\sqrt{\frac{\gamma}{pd}} \tag{12}$$

[12] also described the behavior of Ω for various choices of γ and p. In particular, when $\gamma/p \ll 1$, equations (SS-ODE) reduce to the following simpler ones:

$$\frac{\mathrm{dM}}{\mathrm{d}t} = \Psi^{(\mathrm{M})}(\Omega), \qquad \frac{\mathrm{dQ}}{\mathrm{d}t} = \Psi^{(\mathrm{GF})}(\Omega) \tag{GF-ODE}$$

Our key observation is that the deterministic description above is valid beyond the high-dimensional limit $d \to \infty$ on which previous works [7, 11, 12] have focused. Indeed, Thm. 3.1 is non-asymptotic in (d, p, γ), and can thus be applied in any setting where $\gamma/pd \ll 1$. This is leveraged to provide a tractable, low-dimensional description of SGD in different scenarios of interest which we summarise in Fig. 1.

3.2 The classical regime

The first and most well-studied scenario is the classical regime in which $\gamma \to 0^+$ at fixed dimensions $d, p = O(1)$. Defining the continuous weight $\Theta^\nu = \Theta(\nu \delta t)$ via linear interpolation, a classical result from stochastic optimisation [15] is that one-pass SGD converges to gradient flow on the population risk:

$$\frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} = -\nabla_\Theta \mathcal{R}(\Theta(t)) \tag{13}$$

Or in words: the effective SGD noise ε_ν in eq. (4) is subleading in this limit. Note that this is a deterministic ordinary differential equation of dimension $p(d+1)$. Since $d, p = O(1)$, if they are small eq. (13) provides a computationally efficient description of the SGD dynamics, since it can be easily implemented and solved in a computer. Nonetheless, an alternative description can be derived from the sufficient statistics of eq. (8). Indeed, Thm. 3.1 guarantees that in the limit $\gamma \to 0^+$, the stochastic process in eq. (9) converges to the deterministic limit of eq. (GF-ODE). This ODE can be easily seen to be equivalent to the one of (13), through the following identity:

$$\frac{\mathrm{d}\left\langle \boldsymbol{w_i}, \boldsymbol{w_j} \right\rangle}{\mathrm{d}t} = \left\langle \boldsymbol{w_i}, \frac{\mathrm{d}\boldsymbol{w_j}}{\mathrm{d}t} \right\rangle + \left\langle \frac{\mathrm{d}\boldsymbol{w_i}}{\mathrm{d}t}, \boldsymbol{w_j} \right\rangle$$

This gives rise to ordinary differential equations in $p(k+p)$ parameters. Therefore, depending on the values of (d, p), it can offer a more compact description of the evolution of the performance of the predictor than eq. (13).

As it was previously hinted by the notation, equation (GF-ODE) also provides an intuitive interpretation of the right-hand side of the stochastic process (9). Indeed, the terms proportional to γ/pd in eq. (9) correspond exactly to the terms inside of the expectation in eq. (GF-ODE), and correspond to the projection of the population gradient along the weights $(\Theta, \Theta_\star)$. The remaining term, which is proportional to $\gamma^2/p^2 d$, comes from the variance of the effective noise ε, which is subleading in the limit $\gamma \to 0^+$. This agrees with the characterization of the terms given in [13], where an additional Brownian motion correction term at finer scales is also derived.

In Figure 2 (left) we plot the trajectories of individual neurons in the space spanned by the two target neurons ($k = 2$); $\mathrm{M}_{jr}/\sqrt{\mathrm{Q}_{jj}\mathrm{P}_{rr}}$ is the (normalised) scalar product between $\boldsymbol{w}_j$ and $\boldsymbol{w}_r^*$. While, initially, all neurons are pointing to the superposition of the two teacher weights ($(\sqrt{2}/2, \sqrt{2}/2)$, yellow dot), in the last phase of learning they "specialize" and split evenly to one of the two ($(1,0), (0,1)$, red squares). Note how the ODE trajectories follows closely the simulated ones.

Finally, our results imply a remarkable **dimension independence** property: Differently from (13), the alternative description (GF-ODE) turns out to be independent of the data dimension d. Given the initial conditions of the sufficient statistics (the overlaps), then the trajectories will behave **exactly in the same way** whether d is large or small (see the illustration in Figure 2 (right)). This remarkable property is a direct consequence of the Gaussianity assumption on the data. Note, of course, that dimensionality still plays a crucial role through the initialisation. Indeed, for the typically employed random initialization $\theta_i^0 \sim \mathcal{N}(0_d, \sigma^2 \mathrm{I}_d)$, the initial correlation between the hidden-units and the target, parameterised by $M_{rj}^0 \sim 1/\sqrt{d}$, explicitly depends on d. Since $M = 0$ is often a fixed-point of the dynamics, $d = O(1)$ or $d \gg 1$ will lead effectively to different behaviours.

3.3 The high-dimensional regime

Modern machine learning practice often involves high-dimensional data. As early as in [7, 9] it has motivated the study of the $d \to \infty$ limit of the SGD dynamics (4), under the assumption of fixed learning rate and model complexity $p, \gamma = O(1)$. This setting has witnessed a renewal of interest recently [10–13]. In particular, a remarkable phenomenon arises: differently from the classical limit eq. (GF-ODE) discussed above, in high-dimension the variance term induced by the SGD effective noise yields an explicit contribution to the limiting dynamics (SS-ODE). This term can yield a finite risk contribution at large times even for architectures for which the population gradient flow would otherwise converge to zero population risk (i.e. perfect learning of the target $f_{\Theta^\star}$), so that the large time dynamics plateau at finite risk (see for instance [11, 13]). This is a major difference between the classical and high-dimensional regime.

For strongly convex problems, it is known that SGD with fixed learning rate γ converges to a stationary distribution of variance $\propto \gamma$ [18, 21], leading to an asymptotic risk that closely resembles the one observed by [7] in the high-dimensional regime. This suggests a similar phenomenology in the basin of the global minima, although making this statement precise is challenging due to the non-convexity of the risk for two-layer networks. As noted by [12], this noise term is subleading in γ/p, and can be mitigated by either taking $\gamma \to 0^+$ (i.e. seeing a lot of data) or overparametrising $p \to \infty$. However, since eqs. (SS-ODE) are a system of $p(p+k)$ ordinary differential equations, they become intractable in the limit $p \to \infty$, which is a major shortcoming of this description.

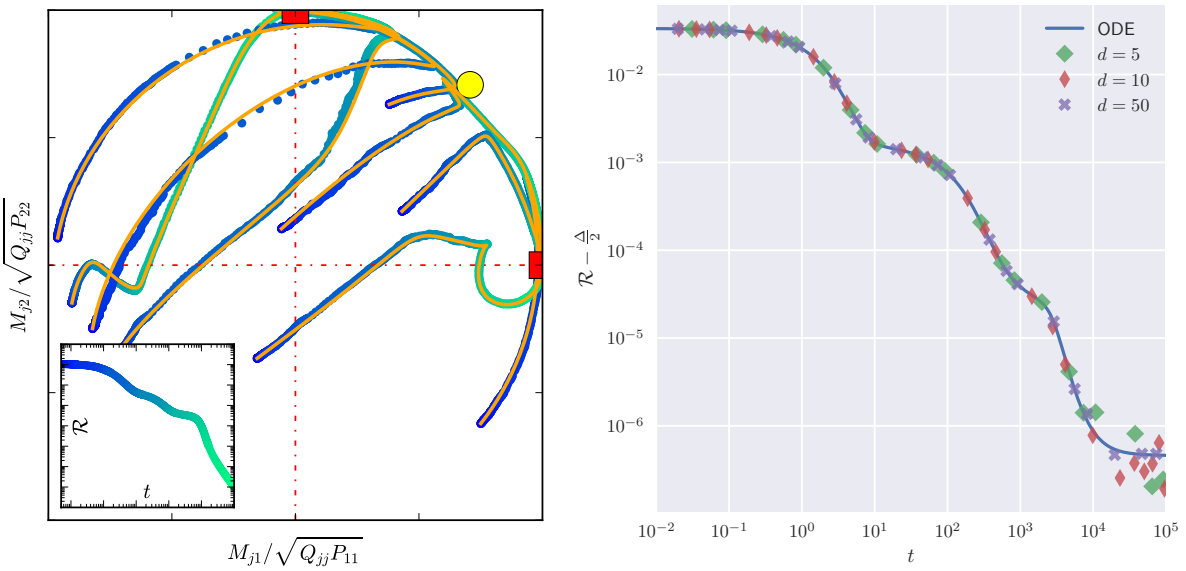


Figure 2: **Dimension independence of the dynamics in the classical regime**: Comparison between the simulated SGD dynamic and the analytical one obtained by integrating the differential equations. All the plots are using $\sigma = \mathrm{erf}(\cdot/\sqrt{2})$. (Left:) Trajectories for the cosine similarity between each network neural and the target; time evolution in simulation is blue to green, ODE trajectories in orange ($k = 2, p = 10, \gamma = 5 \times 10^{-2}, \Delta = 0.0$). (Right:) Different low-dimensional simulations, starting from the same identical initial condition ($k = 5, p = 10, \gamma = 5 \times 10^{-2}, \Delta = 10^{-3}$).

3.4 The overparametrised regime

In both the classical and high-dimensional regimes, the effective description of the SGD dynamics rely on quantities which scale with the hidden-layer width p, and therefore they are not adequate to wide models. Yet, in many scenarios of interest we need to deal with wide, overparametrised networks. The problem is finding an effective low-dimensional description of one-pass SGD for the overparametrised regime $p \to \infty$ was first addressed by [1–4]. The key idea in this line of work is to define an empirical density over the weights $\theta_i^\nu := (a_i^\nu, \boldsymbol{w}_i^\nu)$:

$$\hat{\mu}_p^\nu(\boldsymbol{\theta}) = \frac{1}{p}\sum_{i=1}^{p} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_i^\nu) \tag{14}$$

and to derive a closed-form update for the density $\hat{\mu}_p^\nu(\boldsymbol{\theta})$ from the SGD update of the weights, eq. (4). In the limit $p \to \infty$, those works have shown that the empirical density converges to an asymptotic density over $\mathbb{R}^{d+1}$, which for sufficiently small learning rate satisfies a partial differential equation (PDE) that became known in the literature as the *mean-field limit*. Drawing from the theory of PDEs and optimal transport, this description allowed for the derivation of important mathematical guarantees on the dynamics, such as the global convergence of SGD for two-layers neural networks.

However, the empirical measure $\hat{\mu}_p^\nu$ is defined on $\mathbb{R}^{d+1}$, so a problem still remains when d is large. Indeed, as remarked in [37] it remains challenging to draw quantitative results from this description except for considerably low-dimensional data. However, since in our setting the target function (5) only acts on a low-dimensional subspace of $\mathbb{R}^d$, it is possible to exploit the symmetries of the problem to derive an approximation of constant dimension. This low-dimensional equivalent stems from invariance properties of mean-field equations, which were also used in [6] and studied in depth in [14]. However, [6] only considers the $d \to \infty$ limit, while we derive a limit that is valid for any value of d. [14], on the other hand, is closer to our work (see e.g. their Lemma 4.2), but only handles the approximation of the dynamics by PDEs instead of ODEs.

Decomposing the dynamics: The starting point of this low-dimensional description is the decomposition of W as

$$W = W_{\text{proj}} + W^\perp, \tag{15}$$

where W_{proj} is the orthogonal projection on the teacher vectors $W^\star$. This projection can be expressed using the sufficient statistics defined in eq. (8):

$$W = MP^{-1}W^\star + W^\perp. \tag{16}$$

Similar to (7), we can then define the orthogonal pre-activations and and its covariance matrix:

$$\boldsymbol{\lambda}^{\perp\nu} = W^{\perp\nu}\boldsymbol{x}^{\nu} \tag{17}$$

$$Q^{\perp} = W^{\perp}(W^{\perp})^{\top} = Q - MP^{-1}M^{\top}\,. \tag{18}$$

Since we are in a regime where $\gamma/p \ll 1$, the ODE approximation of SGD corresponds to equations (GF-ODE). From there, with a little algebra (see Appendix B), we can derive the corresponding equations for $Q^{\perp}$:

$$\frac{dQ^{\perp}_{ij}}{dt} = \mathbb{E}_{(\boldsymbol{\lambda}^{\perp},\boldsymbol{\lambda})}\left[\left(\sigma'(\lambda_i)\lambda^{\perp}_j + \sigma'(\lambda_j)\lambda^{\perp}_i\right)\mathcal{E}\right] := \Psi^{\perp}_{ij}(\Omega). \tag{19}$$

Low-dimensional approximation: Informally, the interesting part of the dynamics happens in a low-dimensional space: the one spanned by the target weights $W^{\star}$. The remainder of the dynamics only depends on the student-student vector interactions, which are orthogonally invariant. We therefore make the following assumption to enforce this invariance at the start:

Assumption A3. *The initial vectors $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_p)$ are drawn i.i.d from an orthogonally invariant and K^2/d-subgaussian distribution μ_0, for some constant $K > 0$.*

We show in the appendix how we can then approximate the dynamics by only tracking the evolution of M and $\mathrm{diag}(Q^{\perp})$, and using the following ansatz:

$$\boldsymbol{w}^{\perp}_i \approx \sqrt{q^{\perp}_{ii}} \cdot \boldsymbol{g}_i, \tag{20}$$

where $\boldsymbol{g}_i$ are i.i.d uniform random variables on $\mathbb{S}^{d-k-1}$ independent from the $q^{\perp}_{ii}$. More precisely, we consider the reduced parameters $\tilde{\Theta} = (M, q) \in \mathbb{R}^{p(k+1)}$, and the following mean-field equivalent of the overlaps:

$$\tilde{\Omega} = \begin{pmatrix} \tilde{Q} & M \\ M^{\top} & P \end{pmatrix}, \quad \tilde{Q} = MP^{-1}M^{\top} + D_{\sqrt{q}}\Xi D_{\sqrt{q}} \tag{21}$$

where $D_{\sqrt{q}}$ is the diagonal matrix whose entries are the $\sqrt{q_i}$, and Ξ is a *random* matrix with independent entries such that

$$\Xi_{ii} = 1, \quad \Xi_{ij} = \langle \boldsymbol{g}, \boldsymbol{g'} \rangle \quad \text{with} \quad \boldsymbol{g}, \boldsymbol{g'} \sim \mathrm{Unif}\left(\mathbb{S}^{d-k-1}\right)$$

Then, the mean-field ODEs read

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \mathbb{E}_{\Xi}\left[\Psi^{(M)}(\tilde{\Omega})\right] \quad \frac{\mathrm{d}q_i}{\mathrm{d}t} = \mathbb{E}_{\Xi}\left[\Psi^{\perp}_{ii}(\tilde{\Omega})\right]. \tag{MF-ODE}$$

Similarly, given the parameters Θ, the risk is computed as

$$\mathcal{R}(\tilde{\Theta}) = \mathbb{E}_{\Xi}\left[\mathcal{R}(\tilde{\Omega})\right] \tag{22}$$

The consistency of this approximation is given by the following theorem:

Theorem 3.2. *Under asms. A2 and A3, let $\Omega(t)$ and $\tilde{\Theta}(t)$ denote the solutions of the ODES (GF-ODE) and (MF-ODE), respectively. Then with probability at least $1 - e^{-z^2}$ on the initialization:*

$$\sup_{t\in[0,T]} \left|\mathcal{R}(\Omega(t)) - \mathcal{R}(\tilde{\Theta}(t))\right| \leq Ce^{CT}\left(\sqrt{\log(pT)} + z\right)/\sqrt{p}\,.$$

The proof is given in Appendix C. It uses key elements from the mean-field study of [5]. Indeed, both the solutions of (GF-ODE) & (MF-ODE) can be viewed as the "particle dynamics" approximations (see e.g. [38]) of two mean-field PDEs $\mu_t, \tilde{\mu}_t$ on the space of network weights $\boldsymbol{w} \in \mathbb{R}^d$ and the space or reduced parameters $(\boldsymbol{m}, q) \in \mathbb{R}^{k+1}$, respectively. In turn, the invariance property of μ_0 extends to $\hat{\mu}_t$ for every $t \geq 0$, which implies that $\mathcal{R}(\mu_t) = \mathcal{R}(\tilde{\mu}_t)$.

Remark 3.3. We do not imply in any way, shape or form that Equation (20) represents the actual distribution of the $\boldsymbol{w}^{\perp}_i$, or that (as in (21)) then entries of $Q^{\perp}$ are independent. Rather, it is the structure of the update equations that allows for this approximation to hold.

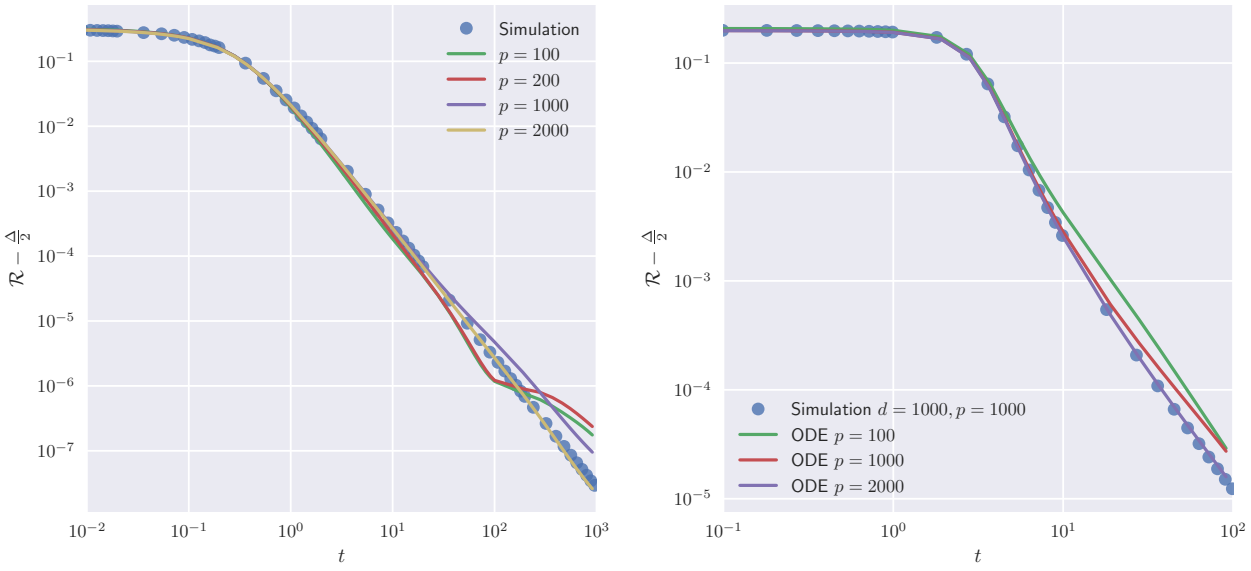


Figure 3: **The mean-field regime:** Comparison between population risks of the simulated learning dynamic and the corresponding deterministic evolution, for (left) mean-field low-dimensional regime with squared activation function ($k = 2, d = 5, \gamma = 1.0, \Delta = 0.0$) and (right:) mean-field high-dimensional regime. Apart from finite size effects, there is agreement between ODEs and simulated dynamics ($k = 5, d = 1000, \gamma = 10.0$).

The high-dimensional limit of mean-field: When $d \to \infty$, the off-diagonal entries of the matrix Ξ are of order $1/\sqrt{d}$, which suggests that they can be neglected safely. We therefore define the following equivalent of Ω_{MF}, which does not depend on auxiliary random variables:

$$\bar{\Omega} = \begin{pmatrix} \bar{Q} & M \\ M^\top & P \end{pmatrix}, \quad \bar{Q} = MP^{-1}M^\top + \mathrm{diag}(q) \tag{23}$$

The following lemma then holds:

Lemma 3.4. *For any $(M, q) \in \mathbb{R}^{p(k+1)}$, we have:*

$$\left\| \mathbb{E}_\Xi \left[\Psi^{(M)}(\tilde{\Omega}) \right] - \Psi^{(M)}(\bar{\Omega}) \right\|_\infty \leq \frac{C}{\sqrt{d}}, \text{ and } \left\| \mathbb{E}_\Xi \left[\Psi^{(\perp)}(\tilde{\Omega}) \right] - \Psi^{(\perp)}(\bar{\Omega}) \right\|_\infty \leq \frac{C}{\sqrt{d}}. \tag{24}$$

We thus define the high-dimensional equivalent of (MF-ODE):

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \Psi^{(M)}(\bar{\Omega}) \qquad \frac{\mathrm{d}q_i}{\mathrm{d}t} = \Psi^{\perp}_{ii}(\bar{\Omega}). \tag{HDMF-ODE}$$

The same propagation of perturbations arguments discussed in Appendix B of [12], and the Lipschitz property of the risk, easily imply the following theorem:

Theorem 3.5. *Under Assumptions A2 & A3, let $\Omega(t)$ and $\bar{\Theta}(t)$ denote the solutions of the ODES (GF-ODE) and (HDMF-ODE), respectively. Then with probability at least $1 - e^{-z^2}$ on the initialization:*

$$\sup_{t \in [0,T]} \left| \mathcal{R}(\Omega(t)) - \mathcal{R}(\bar{\Theta}(t)) \right| \leq Ce^{CT} \left(\frac{\sqrt{\log(pT)} + z}{\sqrt{p}} + \frac{1}{\sqrt{d}} \right)$$

This approximation eliminates the need to compute expectations in (MF-ODE). We now argue that this phenomenon is not only a consequence of rotation invariance, but also a simple concentration result. Indeed, irrespective of Assumption A3, we show the following result:

$$\left| \mathbb{E}_{\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}^\perp \sim \mathcal{N}(\mathbf{0}, \Omega)} \left[\sigma(\lambda_i) \lambda_i^\perp \mathcal{E} \right] - \mathbb{E}_{\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}^\perp \sim \mathcal{N}(\mathbf{0}, \bar{\Omega})} \left[\sigma(\lambda_i) \lambda_i^\perp \mathcal{E} \right] \right| \leq c \sqrt{Q^\perp_{ii} \cdot \frac{\|Q^\perp\|_{\mathrm{op}}}{p}}, \tag{25}$$

For a random sub-gaussian initialization, classical matrix concentration arguments (see [10], Theorem 4.4.5) imply

$$\|Q^{\perp}\|_{\text{op}} = O\left(1 + \frac{p}{d}\right)$$

and common sense arguments (the presence of an attracting force towards $Q^{\perp} = 0$) indicates that this quantity stays within the same order of magnitude. On the other hand, by Assumption A1, the Q_{ii} (and hence the $Q_{ii}^{\perp}$) remain bounded by an absolute constant during the trajectory. We therefore expect, using standard results from ODE perturbation, that

$$\|\bar{\Omega}(t) - \tilde{\Omega}(t)\|_{\infty} \leq e^{Ct}\left(\sqrt{\frac{1}{p}} + \sqrt{\frac{1}{d}}\right). \tag{26}$$

Contrary to the low-dimensional regime and the theorem above, this derivation does not require any rotation invariance of the $\boldsymbol{w}_i^{\perp}$; simply, the averaging properties of $\Psi^{(M)}$ are enough to show direct concentration properties on the dynamics.

It is instructive to look at a concrete example where the phenomenology discussed above becomes explicit. Perhaps the simplest one is given by the square activation $\sigma(x) = x^2$, for which the expectation in eq. (11) can be explicitly expressed in terms of polynomials of the covariance matrices $(\mathrm{Q}^{\nu}, \mathrm{M}^{\nu}, \mathrm{P})$ (see Appendix E for the derivation):

$$\begin{aligned}
\Psi^{(\mathrm{M})}(\Omega) &= 2\left(\frac{\mathrm{Tr}\,[\mathrm{P}]}{k} - \frac{\mathrm{Tr}\,[\mathrm{Q}]}{p}\right)\mathrm{M} + 4\left(\frac{\mathrm{PM}}{k} - \frac{\mathrm{MQ}}{p}\right) \\
\Psi^{(\mathrm{GF})}(\Omega) &= 4\left(\frac{\mathrm{Tr}\,[\mathrm{P}]}{k} - \frac{\mathrm{Tr}\,[\mathrm{Q}]}{p}\right)\mathrm{Q} + 8\left(\frac{\mathrm{M}^{\top}\mathrm{M}}{k} - \frac{\mathrm{Q}^2}{p}\right) \\
\Psi^{(\perp)}(\Omega) &= 4\left(\frac{\mathrm{Tr}\,[\mathrm{P}]}{k} - \frac{\mathrm{Tr}\,[\mathrm{Q}]}{p}\right)\mathrm{Q}^{\perp} - \frac{4}{p}\left(\mathrm{QQ}^{\perp} + \mathrm{Q}^{\perp}\mathrm{Q}\right).
\end{aligned} \tag{27}$$

Similarly, the population risk as a function of the sufficient statistics reads:

$$\mathcal{R}(\Omega) = \frac{\mathrm{Tr}\,[\mathrm{P}]^2 + 2\,\mathrm{Tr}\left[\mathrm{P}^2\right]}{2k^2} - \frac{\mathrm{Tr}\,[\mathrm{P}]\,\mathrm{Tr}\,[\mathrm{Q}] + 2\,\mathrm{Tr}\left[\mathrm{MM}^{\top}\right]}{pk} + \frac{\mathrm{Tr}\,[\mathrm{Q}]^2 + 2\,\mathrm{Tr}\left[\mathrm{Q}^2\right]}{2p^2} + \frac{\Delta}{2}. \tag{28}$$

The high-dimensional mean field limit is then particularly simple. Recalling the definition of $\bar{\Omega}$ in Equation (23), we obtain the explicit equations by replacing Ω with $\bar{\Omega}$. The situation is, however, different for the low-dimensional mean-field limit, due to the randomness introduced by matrix Ξ. The ODEs and the risk, given the parameters $\tilde{\Theta}$, are

$$\begin{aligned}
\mathbb{E}_{\Xi}\left[\Psi^{(M)}(\tilde{\Omega})\right] &= \Psi^{(M)}(\bar{\Omega}) \qquad \mathbb{E}_{\Xi}\left[\Psi^{(\text{noise})}(\tilde{\Omega})\right] = \Psi^{(\text{noise})}(\bar{\Omega}) \\
\mathbb{E}_{\Xi}\left[\Psi_{ii}^{\perp}(\tilde{\Omega})\right] &= \Psi_{ii}^{\perp}(\bar{\Omega}) - \frac{8}{p}\frac{\sum_{j=1,j\neq i}^{p} q_j}{d-k} q_i \\
\mathbb{E}_{\Xi}\left[\mathcal{R}(\tilde{\Omega})\right] &= \mathcal{R}(\bar{\Omega}) + \frac{1}{p^2}\frac{\sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p} q_i q_j}{d-k}.
\end{aligned} \tag{29}$$

In this case, the low-dimensional corrections are therefore just additive terms. As expected, these corrections vanish when $d \to \infty$, where we fall back to the high-dimensional mean-field. Conversely, when $d = k$ the correction diverges, but we don't need to track q anymore since the teacher weights are spanning the whole space $\mathbb{R}^d$, and the orthogonal space is null; hence $\mathrm{Q}^{\perp} = 0$ and this is a stable point of the dynamics. In Figure 3 we show some numerical experiments using this activation function; for a more detailed discussion see Appendix F.

4 Conclusion

Our work provides a comprehensive analysis of the one-pass SGD dynamics of two-layer neural networks. The study bridges different regimes of interest, offers a unifying picture of the limiting SGD dynamics, sheds light on the behavior of neural networks trained on synthetic data and, we believe, provides a useful tool for further investigations of the performance of these networks.

Acknowledgements

We thank Francis Bach, Gérard Ben-Arous, Lenaic Chizat, Theodor Misiakiewicz & Lenka Zdeborová for valuable discussions. We acknowledge funding from the Swiss National Science Foundation grant SNFS OperaGOST, 200021_200390 and the *Choose France - CNRS AI Rising Talents* program.

References

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

[2] Lénaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[3] Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach, 2019.

[4] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[5] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 25–28 Jun 2019.

[6] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887. PMLR, 02–05 Jul 2022.

[7] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, Oct 1995.

[8] David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995.

[9] David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In D. Touretzky, M. C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[10] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.

[11] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[12] Rodrigo Veiga, Ludovic STEPHAN, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborova. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[13] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[14] Karl Hajjar and Lenaic Chizat. On the symmetries in the dynamics of wide two-layer neural networks, 2022.

[15] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[16] Léon Bottou and Yann LeCun. Large scale online learning. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

[17] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[18] Georg Ch. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[19] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[20] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017.

[21] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020.

[22] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334. PMLR, 10–15 Jul 2018.

[23] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 16–18 Apr 2019.

[24] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4633–4635. PMLR, 15–19 Aug 2021.

[25] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21581–21591. Curran Associates, Inc., 2021.

[26] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[27] M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643–656, feb 1995.

[28] M Copelli and N Caticha. On-line learning in the committee machine. *Journal of Physics A: Mathematical and General*, 28(6):1615–1625, mar 1995.

[29] Michael Biehl, Peter Riegler, and Christian Wöhler. Transient dynamics of on-line learning in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 29(16):4769–4780, aug 1996.

[30] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020.

[31] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. In *Proceedings of Machine Learning Research*, volume 145, pages 1–46. 2nd Annual Conference on Mathematical and Scientific Machine Learning, 2021.

[32] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborova. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8936–8947. PMLR, 18–24 Jul 2021.

[33] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow reLU networks for square loss and orthogonal inputs. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[34] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151. Curran Associates, Inc., 2021.

[35] Xiaoyu Wang and Mikael Johansson. On Uniform Boundedness Properties of SGD and its Momentum Variants. Technical report, June 2022. arXiv:2201.10245 [cs, math] type: article.

[36] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval, 2019.

[37] Francis Bach and Lénaic Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. In *International Congress of Mathematicians*, 2022.

[38] A. Chertock. A Practical Guide to Deterministic Particle Methods. volume 18 of *Handbook of Numerical Methods for Hyperbolic Problems*, chapter 7, pages 177–202. Elsevier, January 2017.

[39] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, New York, February 2013.

A Effect of noise on Equations (GF-ODE)

In this appendix we present some corrective terms to Equations (GF-ODE) that allows to have better results when numerically integrating ODEs and comparing them to simulations.

In [12] the terms proportional to γ/p are neglected completely when running numerical integrations. However, we noticed that a term from $\Psi_{ij}^{(\mathrm{Var})}(\Omega)$ could be kept in order to get better agreement between ODEs and simulation in presence of label noise at small but finite $\gamma/p \ll 1$. Letting $\mathcal{E}^* := \mathcal{E} - \sqrt{\Delta}z$, we can decompose

$$\Psi_{ij}^{(\mathrm{Var})}(\Omega) = \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_i)\sigma'(\lambda_j)\,\mathcal{E}^{*2}\right] + \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_i)\sigma'(\lambda_j)\,\Delta\right],$$

allowing us to define

$$\Psi_{ij}^{(\mathrm{noise})}(\Omega) := \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_i)\sigma'(\lambda_j)\,\Delta\right]. \tag{30}$$

The asymptotic of this term at late times $t \gg 1$ was explicitly computed in [27] and [11] for different architectures, where it was found that $\Psi_{ij}^{(\mathrm{noise})} \propto \Delta$. In these cases, when the dynamics approaches the point where $\mathcal{E}^* \sim \gamma\Delta/p$, then the term $\Psi_{ij}^{(\mathrm{GF})}(\Omega)$ is of the same order of $\gamma/p\,\Psi_{ij}^{(\mathrm{noise})}(\Omega)$, while the remaining part of $\Psi_{ij}^{(\mathrm{Var})}(\Omega)$ is still negligible since is scaling as $\mathcal{E}^{*2}$. It follows that the first order correction in γ/p of equations (GF-ODE) is given by:

$$\frac{\mathrm{dM}}{\mathrm{d}t} = \Psi^{(\mathrm{M})}(\Omega), \qquad \frac{\mathrm{dQ}}{\mathrm{d}t} = \Psi^{(\mathrm{GF})}(\Omega) + \frac{\gamma}{p}\Psi^{(\mathrm{noise})}(\Omega). \tag{GF-ODE-NOISE}$$

Despite vanishing in the true $\gamma/p \to 0^+$ limit, the new term enables ODE to catch the behaviour for large times, when simulating small but finite γ/p.

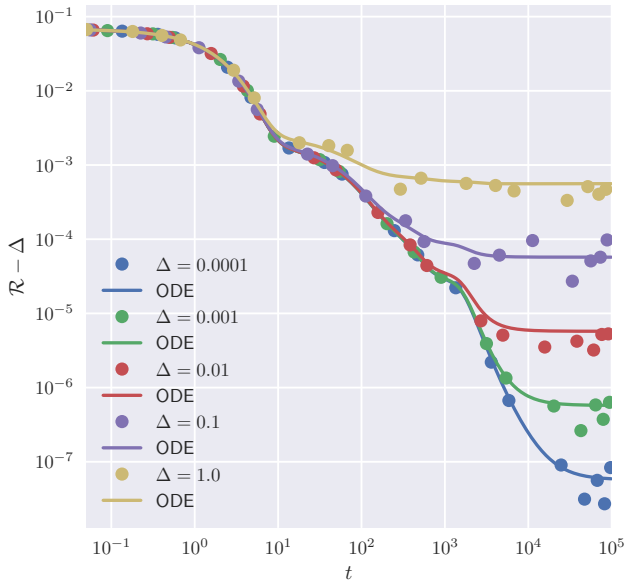


Figure 4: comparison between the simulated learning dynamic and the one obtained by integrating the differential equations, in the *classical regime*; $k = 4, p = 8, \gamma = 5 \times 10^{-2}, d = 10, \sigma = \mathrm{erf}(\cdot/\sqrt{2})$. Learning the same teacher with different level of noise.

Figure 4 shows a numerical experiment explicitly designed to show the effect of noise term. The fluctuations visible in the final plateaus result from the stochastic process in Equation (9): $\Psi^{(\mathrm{Var})}$ and consequently $\Psi^{(\mathrm{noise})}$ are proportional to $||\boldsymbol{x}^\nu||_2^2/d$. When $d = O(1)$ then this term is not concentrating to 1 and leads to a fluctuation. Let us again stress the fact that this is an effect visible only when performing numerical experiments with finite GP, whereas in the true limit the fluctuations disappear and the dynamic is described by deterministic ODEs.

B Derivation of Eq. (19) and local fields covariance

Let's start by taking the time derivative of Eq. (18)

$$\frac{\mathrm{dQ}^{\perp}}{\mathrm{d}t} = \frac{\mathrm{dQ}}{\mathrm{d}t} - \frac{\mathrm{dM}}{\mathrm{d}t}\,\mathrm{P}^{-1}\mathrm{M}^{\top} - \mathrm{M}\mathrm{P}^{-1}\,\frac{\mathrm{dM}^{\top}}{\mathrm{d}t} \coloneqq \Psi^{\perp}(\Omega),$$

and plugging in equations (GF-ODE) and (11) one after the other we get

$$\begin{aligned}\Psi^{\perp}_{ij}(\Omega) &= \mathbb{E}\left[\left(\sigma'(\lambda_i)\lambda_j + \sigma'(\lambda_j)\lambda_i\right)\mathcal{E} - \sum_{r=1}^{k}\sum_{t=1}^{k}\sigma'(\lambda_i)\lambda^{\star}_r\mathcal{E}\left[\mathrm{P}^{-1}\right]_{rt}\left[\mathrm{M}^{\top}\right]_{tj} - \sum_{r=1}^{k}\sum_{t=1}^{k}[\mathrm{M}]_{ir}\left[\mathrm{P}^{-1}\right]_{rt}\lambda^{\star}_t\sigma'(\lambda_i)\mathcal{E}\right] \\ &= \mathbb{E}\left[\mathcal{E}\sum_{r=1}^{k}\sum_{t=1}^{k}\left(\sigma'(\lambda_i)\left(\lambda_j - \lambda^{\star}_r\left[\mathrm{P}^{-1}\right]_{rt}\left[\mathrm{M}^{\top}\right]_{tj}\right) + \sigma'(\lambda_j)\left(\lambda_i - [\mathrm{M}]_{ir}\left[\mathrm{P}^{-1}\right]_{rt}\lambda^{\star}_t\right)\right)\right],\end{aligned}$$

where all the expected value are intended over $(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{\star}) \sim \mathcal{N}(\mathbf{0}_{p+k}, \Omega)$. Starting from the definition of $\boldsymbol{\lambda}^{\perp}$

$$\boldsymbol{\lambda}^{\perp} = W^{\perp}\boldsymbol{x} = \left(W - MP^{-1}W^{\star}\right)\boldsymbol{x} = \boldsymbol{\lambda} - MP^{-1}\boldsymbol{\lambda}^{\star},$$

and writing single component

$$\boldsymbol{\lambda}^{\perp}_i = \lambda_i - \sum_{r=1}^{k}\sum_{t=1}^{k}[\mathrm{M}]_{ir}\left[\mathrm{P}^{-1}\right]_{rt}\lambda^{\star}_t,$$

substituting in the expression above we finally get

$$\Psi^{\perp}_{ij}(\Omega) = \mathbb{E}\left[\left(\sigma'(\lambda_i)\lambda^{\perp}_j + \sigma'(\lambda_j)\lambda^{\perp}_i\right)\mathcal{E}\right]. \tag{31}$$

The explicit computation of the expected value depends on the particular activation function used. Even though, the final expression can only be function of the covariance matrix entries, since all the local fields are zero-mean Gaussian variables. We report the covariance matrix of local fields

$$\mathrm{Cov}\left[\boldsymbol{\lambda}, \boldsymbol{\lambda}^{\perp}, \boldsymbol{\lambda}^{\star}\right] = \begin{pmatrix} \mathrm{Q} & \mathrm{Q}^{\perp} & \mathrm{M} \\ \mathrm{Q}^{\perp} & \mathrm{Q}^{\perp} & 0 \\ \mathrm{M}^{\top} & 0 & \mathrm{P} \end{pmatrix}, \tag{32}$$

where $\mathrm{Q}^{\perp} = \mathrm{Q} - \mathrm{M}\mathrm{P}^{-1}\mathrm{M}^{\top}$ as defined above.

C Mean-field approximation: proof of Theorem 3.2

Preliminaries We begin by recalling the results of [5]. For any distribution μ on $\mathbb{R}^d$, we define the network function

$$\hat{f}_\mu(\boldsymbol{x}) = \int \sigma(\boldsymbol{w}^\top \boldsymbol{x}) d\mu(\boldsymbol{w}). \tag{33}$$

It is easy to check that when

$$\mu = \mu_p(W) := \frac{1}{p}\sum_{i=1}^{p} \delta_{\boldsymbol{w}_i},$$

then $\hat{f}_\mu = f_\Theta$, where f_Θ was defined in (1). The associated network risk is then given by

$$\hat{\mathcal{R}}(\mu) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},{}^1\!/_d I_d)}\left[\frac{1}{2}\left(f_{\Theta^\star}(\boldsymbol{x}) - \hat{f}_\mu(\boldsymbol{x})\right)^2\right]. \tag{34}$$

Similarly, we can consider the continuous equivalent of the gradient flow equation as follows:

$$\varphi(\boldsymbol{w},\mu) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},{}^1\!/_d I_d)}\left[\sigma'(\boldsymbol{w}^\top\boldsymbol{x})\boldsymbol{x}\left(f_{\Theta^\star}(\boldsymbol{x}) - \hat{f}_\mu(\boldsymbol{x})\right)\right]. \tag{35}$$

Through conservation of matter arguments, the evolution of the empirical measure for particles following the gradient flow equation (13) obeys the following partial differential equation:

$$\partial_t \mu_t = \nabla_{\boldsymbol{w}} \cdot \left(\mu_t\, \varphi(\,\cdot\,,\mu_t)\right) \tag{GF-PDE}$$

Theorem C.1 ([5], Propositions 13-16). *Let μ_0 be a K^2/d-subgaussian distribution, and consider the following two processes:*

- *the solution μ_t of (GF-PDE) with initial value μ_0,*
- *the solution $W(t)$ of the gradient flow ODE (13), with initial conditions $\boldsymbol{w_i}(0) \sim \mu_0$ i.i.d.*

Then for any $T \geq 0$ we have

$$\sup_{t\in[0,T]} \left|\mathcal{R}(W(t)) - \hat{\mathcal{R}}(\mu_t)\right| \leq \frac{Ce^{CT}\left(\sqrt{\log(pT)} + z\right)}{\sqrt{p}} \tag{36}$$

with probability at least $1 - e^{-z^2}$.

As we have already mentioned in the section devoted to the gradient flow regime, the solution $\Omega(t)$ of (GF-ODE) is exactly the overlap matrix of $W(t)$, and hence the term $\mathcal{R}(W(t))$ can be replaced by $\mathcal{R}(\Omega(t))$ in (36).

Rotation invariance The crux of Theorem 3.2 lies in the rotation invariance of μ_0. Indeed, as noticed in [6, 14], such symmetries are conserved throughout the mean-field dynamics.

Proposition C.2 ([14], Proposition 2.1). *Let U be a linear transformation, and assume that the initial measure μ_0, the teacher function $f_\star$ and the data measure ρ are all invariant under U. Let μ_t be the solution to (GF-PDE) with initial condition μ_0. Then μ_t is U-invariant for all $t \geq 0$.*

Write $\mathbb{R}^d = V^\star \oplus V^\perp$, where $V^\star$ is the span of $W^\star$ and $V^\perp$ is its orthogonal subspace. Proposition C.2 then implies immediately that μ_t is rotation-invariant on $V^\perp$. Hence, if we define the following function:

$$h(\boldsymbol{w}) = \left(\frac{W^\star \boldsymbol{w}}{d}, \|\boldsymbol{w}^\perp\|^2\right) \tag{37}$$

where $\boldsymbol{w}^\perp$ is the projection of $\boldsymbol{w}$ on $V^\perp$, then μ_t is only determined by its pushforward $\tilde{\mu}_t = h_\# \mu_t$. Conversely, for a set of reduced parameters $(\boldsymbol{m}, q) \in \mathbb{R}^d$, and $\boldsymbol{g} \in \mathbb{S}^{d-k-1}$, we define

$$\tilde{\boldsymbol{w}} = W^{\star\top} P^{-1} \boldsymbol{m} + \sqrt{q}\cdot \boldsymbol{g}, \tag{38}$$

where $\boldsymbol{g}$ is a uniform unit vector in $V^\perp$. Then μ_t is the measure of $\tilde{\boldsymbol{w}}$ when $(\boldsymbol{m}, q) \sim \tilde{\mu}_t$ and $\boldsymbol{g} \sim \mathrm{Unif}(\mathbb{S}^{d-k-1})$. This allows us to write the reduced equations for $\tilde{\mu}_t$:

$$\begin{aligned}
\partial_t \tilde{\mu}_t &= \nabla_{(\boldsymbol{m},q)} \cdot \left(\tilde{\mu}_t \tilde{\varphi}(\,\cdot\,, \tilde{\mu}_t)\right) \\
\tilde{\varphi}_{\boldsymbol{m}}((\boldsymbol{m}, q), \tilde{\mu}) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{g}}\left[\sigma'(\tilde{\boldsymbol{w}}^\top \boldsymbol{x}) W^\star \boldsymbol{x} \left(f_{\Theta^\star}(\boldsymbol{x}) - \tilde{f}_{\tilde{\mu}}(\boldsymbol{x})\right)\right] \\
\tilde{\varphi}_q((\boldsymbol{m}, q), \tilde{\mu}) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{g}}\left[\sigma'(\tilde{\boldsymbol{w}}^\top \boldsymbol{x}) \sqrt{q}\, \boldsymbol{g}^\top \boldsymbol{x} \left(f_{\Theta^\star}(\boldsymbol{x}) - \tilde{f}_{\tilde{\mu}}(\boldsymbol{x})\right)\right]
\end{aligned} \tag{DF-PDE}$$

where $\boldsymbol{x} = (\boldsymbol{z}, \boldsymbol{r}) \sim \mathcal{N}(0, {}^1\!/_d I_d)$ is a normalised Gaussian vector, $\boldsymbol{g} \sim \mathrm{Unif}(\mathbb{S}^{d-k-1})$, and

$$\tilde{f}_{\tilde{\mu}}(\boldsymbol{x}) = \int \sigma(\tilde{\boldsymbol{w}}^\top \boldsymbol{x}) d\tilde{\mu}(\boldsymbol{m}, q) d\nu(\boldsymbol{g}) \tag{39}$$

The associated population risk is now

$$\tilde{\mathcal{R}}(\tilde{\mu}) := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d I_d)}\left[\frac{1}{2}\left(f_{\Theta^\star}(\boldsymbol{x}) - \tilde{f}_{\tilde{\mu}}(\boldsymbol{x})\right)^2\right]. \tag{40}$$

We have therefore shown the following proposition:

Proposition C.3. *Assume that μ_0 is rotation-invariant on $V^\perp$. Consider these two processes:*

- *the solution μ_t of (GF-PDE) with initial value μ_0,*
- *the solution $\tilde{\mu}_t$ of (DF-PDE) with initial value $h_\# \mu_0$.*

Then, for all $t \geq 0$, we have

$$\hat{\mathcal{R}}(\mu_t) = \tilde{\mathcal{R}}(\tilde{\mu}_t) \tag{41}$$

Back to ODEs Consider a population of p particles $(\boldsymbol{m}, q)$ that evolve according to the following equations:

$$\begin{aligned}
\frac{d\boldsymbol{m_i}}{dt} &= \tilde{\varphi}_{\boldsymbol{m}}(\boldsymbol{m_i}(t), \tilde{\mu}_p(t)) \\
\frac{dq_i}{dt} &= \tilde{\varphi}_q(q_i(t), \tilde{\mu}_p(t))
\end{aligned} \tag{DF-ODE}$$

where $\tilde{\mu}_p(t)$ is the empirical distribution of the population:

$$\tilde{\mu}_p(t) = \frac{1}{p}\sum_{i=1}^{p} \delta_{\boldsymbol{m}_i(t), q_i(t)}.$$

Then, by the same arguments as Theorem C.1, we have:

Proposition C.4. *Assume that $\tilde{\mu}_0$ is a K^2/d-subgaussian distribution, and consider the two processes:*

- *the solution $\tilde{\mu}_t$ of (GF-PDE) with initial value $\tilde{\mu}_0$,*
- *the solution $\tilde{\Theta}(t) = (\boldsymbol{m_i}(t), q_i(t))_i$ of (DF-ODE) with initial conditions $\boldsymbol{m}_i(0), q_i(0) \sim \tilde{\mu}_0$ i.i.d.*

Then for all $T \geq 0$, we have

$$\sup_{t \in [0,T]} \left|\tilde{\mathcal{R}}(\tilde{\Theta}(t)) - \tilde{\mathcal{R}}(\tilde{\mu}_t)\right| \leq \frac{C e^{CT}\left(\sqrt{\log(pT)} + z\right)}{\sqrt{p}}, \tag{42}$$

with probability at least $1 - e^{-z^2}$, where $\tilde{\mathcal{R}}(\Theta(t)) = \tilde{\mathcal{R}}(\tilde{\mu}_p(t))$.

In conclusion, we have shown the following:

Theorem C.5. *Assume that conditions A2 and A3 are both satisfied, and define*

$$\mathrm{M}^0 = \frac{\mathrm{W}^0 \mathrm{W}^{\star\top}}{d}, \quad \mathrm{Q}^0 = \frac{\mathrm{W}^0 \mathrm{W}^{0\top}}{d}, \quad \boldsymbol{q}^0 = \mathrm{diag}\left(Q^0 - M^0 P^{-1} M^{0\top}\right). \tag{43}$$

Consider the two following processes:

- *the solution* $\Omega(t)$ *of* (GF-ODE) *with initial value* Ω^0,
- *the solution* $\tilde{\Theta}(t)$ *of* (DF-ODE) *with initial value* $(M^0, \boldsymbol{q}^0)$.

Then for any $T \geq 0$, *we have*

$$\sup_{t\in[0,T]} \left|\mathcal{R}(\Omega(t)) - \tilde{\mathcal{R}}(\tilde{\Theta}(t))\right| \leq \frac{Ce^{CT}\left(\sqrt{\log(pT)} + z\right)}{\sqrt{p}}, \tag{44}$$

with probability at least $1 - e^{-z^2}$.

Matching the equations To show that Theorem C.5 implies Theorem 3.2, we need to show the following:

- the update equations for (MF-ODE) and (DF-ODE) are the same
- the risk definitions for $\tilde{\mathcal{R}}(\tilde{\Theta})$ matches the one from (22).

We will only show it for $\tilde{\varphi}_{\boldsymbol{m}}$; the rest is done similarly. Expanding the definition, we have

$$\tilde{\varphi}_{\boldsymbol{m}}((\boldsymbol{m}_i, q_i), \tilde{\mu}_p)_r = \frac{1}{k}\sum_{s=1}^{k}\langle\sigma'(\tilde{\lambda}_i)\lambda_r^\star\sigma(\lambda_s^\star)\rangle - \frac{1}{p}\sum_{j=1}^{p}\langle\sigma'(\tilde{\lambda}_i)\lambda_r^\star\sigma(\tilde{\lambda}_j)\rangle \tag{45}$$

where $\tilde{\lambda}_i = \tilde{\boldsymbol{w}}_i^\top \boldsymbol{x}$ and $\langle\cdot\rangle$ denotes the expectation with respect to $\boldsymbol{x}, \boldsymbol{g}$. On the other hand,

$$\Psi^{(M)}(\Omega) = \frac{1}{k}\sum_{s=1}^{k}\langle\sigma'(\lambda_i)\lambda_r^\star\sigma(\lambda_s^\star)\rangle - \frac{1}{p}\sum_{j=1}^{p}\langle\sigma'(\lambda_i)\lambda_r^\star\sigma(\lambda_j)\rangle, \tag{46}$$

without any expectation on $\boldsymbol{g}$; hence we only need to match the expressions term by term. The first sums of (45) and (46) are actually identical (since the marginal distribution of $\lambda_i^\star$ is independent from $\boldsymbol{g}$), so we look at the second ones:

$$\langle\sigma'(\tilde{\lambda}_i)\lambda_r^\star\sigma(\tilde{\lambda}_j)\rangle = \int \mathbb{E}_{x\sim\mathcal{N}(\mathbf{0},{}^1\!/_{dI_d})}\left[\sigma'(\tilde{\lambda}_i)\lambda_r^\star\sigma(\tilde{\lambda}_j)\right] d\nu(\boldsymbol{g}_i)d\nu(\boldsymbol{g}_j) \tag{47}$$

For $i \neq j$ and a given realization of $\boldsymbol{g}_i, \boldsymbol{g}_j$, the covariance matrix of $(\tilde{\lambda}_i, \tilde{\lambda}_j, \lambda_r^\star)$ is

$$\mathrm{Cov}\left[\tilde{\lambda}_i, \tilde{\lambda}_j, \lambda_r^\star\right] = \begin{pmatrix} \boldsymbol{m}_i^\top P^{-1}\boldsymbol{m_i} + q_i & \boldsymbol{m}_i^\top P^{-1}\boldsymbol{m_j} + \sqrt{q_iq_j}\xi_{ij} & m_{ir} \\ \boldsymbol{m}_i^\top P^{-1}\boldsymbol{m_j} + \sqrt{q_iq_j}\xi_{ij} & \boldsymbol{m}_i^\top P^{-1}\boldsymbol{m_i} + q_i & m_{jr} \\ m_{ir} & m_{jr} & p_{rr} \end{pmatrix}$$

where $\xi_{ij} = \langle\boldsymbol{g_i}, \boldsymbol{g_j}\rangle$. It is easy to check that this is exactly equal to $\tilde{\Omega}^{ijr}$, where $\tilde{\Omega}$ is defined in (21). Finally, by linearity of expectation, since we never have a three-way correlation term between $\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_\ell$, we can consider all the ξ_{ij} to be independent. This ends the proof of Theorem 3.2.

D Proofs for the high-dimensional mean-field approximation

D.1 Preliminaries

We first provide several bounds on expectations of functions of Gaussians, that will be used throughout this section. We begin by recalling the classical Gaussian Poincaré inequality (see e.g. [39], Theorem 3.20):

Lemma D.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and $X \sim \mathcal{N}(0, I)$. Then*

$$\mathrm{Var}(f(X)) \le \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

If $X \sim \mathcal{N}(0, \Sigma)$, the following bounds hold instead:

$$\mathrm{Var}(f(X)) \le \mathbb{E}\left[\langle \nabla f(X), \Sigma \nabla f(X)\rangle\right] \le \|\Sigma\|_{\mathrm{op}} \cdot \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

Now, we provide some concentration bounds for expectations of random variables. Let $h : \mathbb{R} \to \mathbb{R}$ be a function, and z a random variable; our goal is to bound

$$\left|\mathbb{E}[h(z)] - h(\mathbb{E}[z])\right|$$

under different assumptions on f, z.

Lemma D.2. *Assume that h is L-Lipschitz, and that z has a second moment. Then*

$$\left|\mathbb{E}[h(z)] - h(\mathbb{E}[z])\right| \le L\sqrt{\mathrm{Var}(z)} \tag{48}$$

Proof. By Jensen's inequality, we have

$$\begin{aligned}\left|\mathbb{E}[h(z)] - h(\mathbb{E}[z])\right| &\le \mathbb{E}\left[|h(z) - h(\mathbb{E}[z])|\right] \\ &\le L\mathbb{E}[|z - \mathbb{E}[z]|] \\ &\le L\sqrt{\mathrm{Var}(z)}\end{aligned}$$

where the last line uses the Cauchy-Schwarz inequality. □

Lemma D.3. *Assume that h is twice differentiable, with $\left\|h^{(k)}\right\|_\infty \le K$, and that z has a fourth moment. Then*

$$\left|\mathbb{E}[h(z)] - h(\mathbb{E}[z])\right| \le \frac{K}{2}\sqrt{\mathbb{E}\left[(z-s)^4\right]} \tag{49}$$

Proof. For simplicity, let $s = \mathbb{E}[z]$. By the Lagrange formula for Taylor series, we can write

$$h(x) = h(s) + (x-s)h'(s) + (x-s)^2 R(x) \tag{50}$$

where R is bounded by $K/2$. Plugging z into this equation,

$$\left|\mathbb{E}[h(z)] - h(s)\right| = \left|\mathbb{E}\left[(z-s)^2 R(z)\right]\right| \le \sqrt{\mathbb{E}\left[(z-s)^4\right]\mathbb{E}\left[R(z)^2\right]} \tag{51}$$

via the Cauchy-Schwarz inequality, and the result ensues. □

D.2 Proof of Lemma 3.4

We only show the result for $\Psi^{(M)}$; the one for $\Psi^{\perp}$ ensues from similar methods. Recalling the expansion in (46) as a sum of 3-point correlation functions, we define the following function of 3×3 matrices:

$$f(\Sigma) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\Sigma)}\left[\sigma'(z_1)z_2\sigma(z_3).\right] \tag{52}$$

Then $\Psi^{(M)}(\Omega)$ is simply an average of $f(\Sigma)$ for Σ submatrices of Ω. We first show the following bound:

Lemma D.4. *Under Assumption A2, the function f satisfies the following inequality: for any Σ, Σ' such that $\left\|\Sigma' - \Sigma\right\|_\infty \le K$,*

$$\left|f(\Sigma') - f(\Sigma)\right| \le C(1 + \sqrt{\Sigma_{22}})\left\|\Sigma' - \Sigma\right\|_\infty \tag{53}$$

Proof. Let Σ and Σ' be two positive semidefinite matrices, and $\delta\Sigma = \Sigma' - \Sigma$. We can write $\delta\Sigma = \delta\Sigma_1 - \delta\Sigma_2$ where both $\delta\Sigma_i$ are positive, and

$$\|\delta\Sigma_i\|_\infty \leq C\|\delta\Sigma_i\|_{\mathrm{op}} \leq C\|\delta\Sigma\|_{\mathrm{op}} \leq C'\|\delta\Sigma\|_\infty$$

using norm equivalence. Hence, we shall assume from now on that $\delta\Sigma$ is positive semidefinite, and write

$$z' = z + \delta z$$

where $\delta z \sim \mathcal{N}(\mathbf{0}, \delta\Sigma)$ is independent from z. Define

$$\delta\sigma = \sigma(z_3') - \sigma(z_3) \quad \text{and} \quad \delta\sigma' = \sigma'(z_1') - \sigma'(z_1)$$

Then

$$\begin{aligned} f(\Sigma') - f(\Sigma) = \; & \mathbb{E}\left[\sigma'(z_1)z_2\delta\sigma + \sigma'(z_1)\delta z_2\sigma(z_3) + \delta\sigma' z_2\sigma(z_3)\right] \\ & + \mathbb{E}\left[\delta\sigma'\delta z_2\sigma(z_3) + \delta\sigma' z_2\delta\sigma + \sigma'(z_1)\delta z_2\delta\sigma\right] \\ & + \mathbb{E}[\delta\sigma'\delta z_2\delta\sigma] \end{aligned}$$

Since δz is independent from z, and σ satisfies Assumption A2, we can apply Lemma D.3 to get

$$|\mathbb{E}_{\delta z_3}[\delta\sigma]| \leq C\delta\Sigma_{33},$$

and hence by the Cauchy-Schwarz inequality

$$\left|\mathbb{E}\left[\sigma'(z_1)z_2\delta\sigma\right]\right| \leq C'\sqrt{\Sigma_{22}}\,\delta\Sigma_{33}.$$

A similar bound holds for the two other terms of the first line. For the second line, another application of Cauchy-Schwarz yields

$$\begin{aligned} \mathbb{E}_{\delta z}\left[\delta\sigma' z_2\delta\sigma\right] &\leq z_2\sqrt{\mathbb{E}[(\delta\sigma)^2]}\sqrt{\mathbb{E}[(\delta\sigma')^2]} \\ &\leq z_2\sqrt{\delta\Sigma_{11}\delta\Sigma_{33}} \end{aligned}$$

by a combination of Lemmas D.1 and D.2. A similar bound holds for the third line, and it is easily checked that all of the obtained bounds are lower than the one of Lemma D.4. □

Now, the expansion of $\Psi^{(M)}(\tilde{\Omega}) - \Psi^{(M)}(\bar{\Omega})$ is an average of terms of the form $f(\tilde{\Sigma}) - f(\bar{\Sigma})$, where $(\tilde{\Sigma} - \bar{\Sigma})_{ij}$ is either 0 or $\sqrt{q_{ii}^\perp q_{jj}^\perp}\xi_{ij}$. Hence, we can write each of those terms as $h(\xi_{ij}) - h(0)$, and by Assumption A1 and Lemma D.4, the function h is Lipschitz. Lemma 3.4 then ensues from an application of Lemma D.2.

D.3 Proof of Eq. (25)

We begin by showing the following lemma. Recall the definition of $\mathcal{E}$:

$$\mathcal{E} = \frac{1}{k}\sum_{r=1}^{k}\sigma(\lambda_r^\star) - \frac{1}{p}\sum_{i=1}^{p}\sigma(\lambda_i) \tag{54}$$

Lemma D.5. *There exists a constant $c \geq 0$ such that for any choice of $\boldsymbol{\lambda}^\star$,*

$$\mathrm{Var}_{\boldsymbol{\lambda}^\perp}(\mathcal{E}) \leq \frac{c\,\|Q^\perp\|_{\mathrm{op}}}{p}$$

Proof. We apply the Gauss-Poincaré inequality of Lemma D.1 to

$$f(\boldsymbol{\lambda}^\perp) = \sum_{i=1}^{p}\sigma([M\boldsymbol{\lambda}^\star]_i + \lambda_i^\perp), \quad \text{which implies} \quad [\nabla f]_i = \sigma'(\lambda_i)$$

Whenever σ is Lipschitz, we thus have $\|\nabla f(\boldsymbol{\lambda}^\perp)\|^2 \leq cp$, and the lemma ensues. □

We are now in a position to show Eq. (25). For brevity, we denote by $\mathbb{E}$ (resp. $\mathbb{E}_{\text{MF}}$) the expectations with respect to $\boldsymbol{\lambda}^\perp \sim \mathcal{N}(\mathbf{0}, Q^\perp)$ (resp. $\boldsymbol{\lambda}^\perp \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(Q^\perp))$). Since the marginals of both distributions are the same, and by linearity,

$$\mathbb{E}\left[f(\lambda_i^\perp)\right] = \mathbb{E}_{\text{MF}}\left[f(\lambda_i^\perp)\right] \quad \text{and} \quad \mathbb{E}\left[\mathcal{E}\right] = \mathbb{E}_{\text{MF}}\left[\mathcal{E}\right].$$

Under the distribution $\mathcal{N}(\mathbf{0}, \operatorname{diag}(Q^\perp))$, $\lambda_i^\perp$ is almost independent from $\mathcal{E}$, except for the term containing $\sigma(\lambda_i)$. We can thus write, for any $\boldsymbol{\lambda}^\star$

$$\mathbb{E}_{\text{MF}}\left[\mathcal{E}_i \lambda_i^\perp\right] = \mathbb{E}\left[\sigma'(\lambda_i)\lambda_i^\perp\right]\mathbb{E}\left[\mathcal{E}\right] - \frac{1}{p}\mathbb{E}_{\text{MF}}\left[\sigma'(\lambda_i)\lambda_i^\perp\sigma(\lambda_i)\right] + \frac{1}{p}\mathbb{E}\left[\sigma'(\lambda_i)\lambda_i^\perp\right]\mathbb{E}\left[\sigma(\lambda_i)\right]$$

Hence,

$$\mathbb{E}\left[\mathcal{E}_i \lambda_i^\perp\right] - \mathbb{E}_{\text{MF}}\left[\mathcal{E}_i \lambda_i^\perp\right] = \mathbb{E}\left[\sigma'(\lambda_i)\lambda_i^\perp\left(\mathcal{E} - \mathbb{E}\left[\mathcal{E}\right]\right)\right] - \frac{1}{p}\mathbb{E}\left[\sigma'(\lambda_i)\lambda_i^\perp\left(\sigma(\lambda_i) - \mathbb{E}[\sigma(\lambda_i)]\right)\right] \tag{55}$$

Now, using the Cauchy-Schwarz inequality,

$$\left|\mathbb{E}\left[\sigma'(\lambda_i)\lambda_i^\perp\left(\mathcal{E} - \mathbb{E}\left[\mathcal{E}\right]\right)\right]\right| \leq \sqrt{\mathbb{E}\left[\left(\sigma'(\lambda_i)\lambda_i^\perp\right)^2\right]\mathbb{E}\left[\left(\mathcal{E} - \mathbb{E}\left[\mathcal{E}\right]\right)^2\right]}.$$

For Lipschitz σ, the first term is easily bounded by $c\,Q_{ii}^\perp$, and the second is exactly the variance computed in Lemma D.5. The second term in (55) being clearly negligible before the first, we finally get

$$\left|\mathbb{E}\left[\mathcal{E}_i \lambda_i^\perp\right] - \mathbb{E}_{\text{MF}}\left[\mathcal{E}_i \lambda_i^\perp\right]\right| \leq c\sqrt{Q_{ii}^\perp \cdot \frac{\|Q^\perp\|_{\text{op}}}{p}}, \tag{56}$$

and Eq. (25) ensues by taking the expectation w.r.t $\boldsymbol{\lambda}^\star$ on both sides.

E Derivation of explicit expression for the squared activation

In this appendix we show how to derive the differential equations for the dynamics when both σ and σ^* are the square function. We will not present the $\Psi^{(\mathrm{Var})}$ term since we never use the square activation in the high-dimensional regime.

The starting points are Equations (11) and the fact that $\mathcal{R} = \mathcal{E}^2/2$. Due to the linearity of the expected value, we can reduce the expectation on products of $\mathcal{E}$ and λ. Let's start with the population risk. The expected values we need can be expanded to

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\mathcal{E}^2\right] =& \frac{1}{k^2}\sum_{r,s=1}^{k}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma(\lambda_r^*)\sigma(\lambda_s^*)\right] + \\
& \frac{1}{p^2}\sum_{j,l=1}^{k}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma(\lambda_j)\sigma(\lambda_l)\right] \\
& -\frac{2}{pk}\sum_{j=1}^{p}\sum_{r=1}^{k}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma(\lambda_j)\sigma(\lambda_r^*)\right], \\
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_l\mathcal{E}\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_{r'}^*)\right] \\
& -\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_{l'})\right], \\
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_r^*\mathcal{E}\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_r^*\sigma(\lambda_{r'}^*)\right] \\
& -\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda_j)\lambda_r^*\sigma(\lambda_{l'})\right].
\end{aligned}$$

These expansions are still valid for any generic activation function. Before specializing in $\sigma(x) = x^2$, we introduce a shorthand in the notation. We will use

$$\omega_{\alpha\beta} \coloneqq [\Omega]_{\alpha\beta},$$

where the indices α and β can discriminate between teacher and student local fields, as well as the numerical index. Actually, this notation allow us to compute also $\Psi^{\perp}(\Omega)$ by expanding Equation (19) as above, and using the matrix in Equation (32) as covariance for the normal distribution.

With this consideration, there are only 2 types of expected values to be computed. Let us write them explicitly, using our specific activation function

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma(\lambda^\alpha)\sigma(\lambda^\beta)\right] &= \mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[(\lambda^\alpha)^2(\lambda^\beta)^2\right], \\
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\sigma'(\lambda^\alpha)\lambda^\beta\sigma(\lambda^\gamma)\right] &= 2\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\lambda^\alpha\lambda^\beta(\lambda^\gamma)^2\right].
\end{aligned}$$

We are left with some expected values of polynomials of Gaussian variables. Since the local fields all have zero mean, these are nothing but moments of a Gaussian distribution with multiple variables. The standard result used to calculate these is the Isserlis' Theorem:

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[(\lambda^\alpha)^2(\lambda^\beta)^2\right] &= \omega_{\alpha\alpha}\omega_{\beta\beta} + 2\omega_{\alpha\beta}^2, \\
2\mathbb{E}_{(\boldsymbol{\lambda},\boldsymbol{\lambda}^\star)\sim\mathcal{N}(\mathbf{0}_{p+k},\Omega)}\left[\lambda^\alpha\lambda^\beta(\lambda^\gamma)^2\right] &= 2\omega_{\alpha\beta}\omega_{\gamma\gamma} + 4\omega_{\alpha\gamma}\omega_{\beta\gamma}.
\end{aligned}$$

By retracing all steps backward and making the necessary substitutions, we can arrive at an explicit form of the Equations (GF-ODE). In order to obtain a matrix form such as in the Equations (27), we have to write in a closed form all the summations appeared during the derivation and use the fact that Q and P are symmetric matrices.

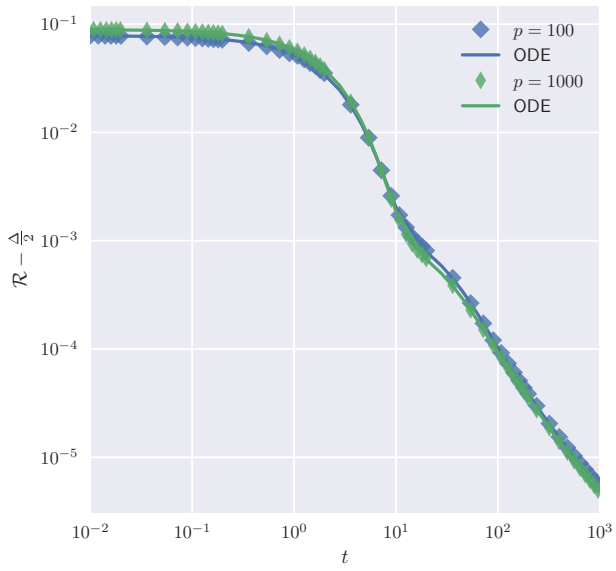
F Numerical experiments in the mean-field limit

In this appendix we show the result of some numerical experiments we performed to verify the statements exposed in Section 3.4. We also refer to our GitHub repository on [https://github.com/IdePHICS/DimensionlessDynamicsSGD].

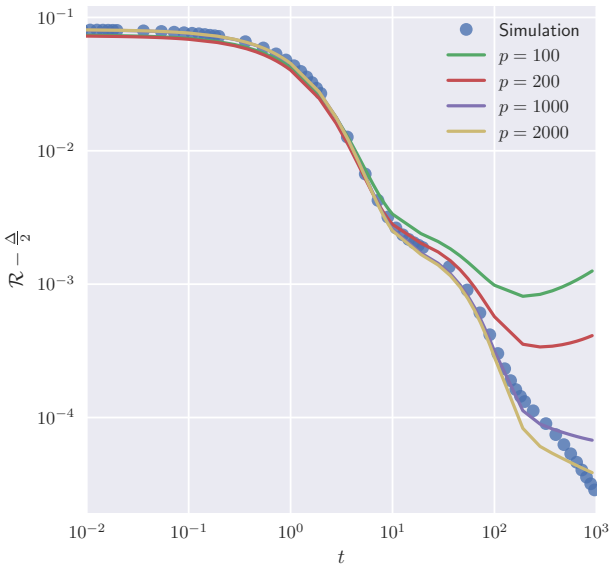
First, we checked that the point $\mathrm{Q}^{\perp} = 0$ is a fixed point of the dynamic. We compare a simulation of this case with an integration of just the matrix M. Since when $\mathrm{Q}^{\perp} = 0$, M is a *sufficient statistic* by itself, we expect it to be the only parameter to be evolved to characterize the dynamics. Figure 5 (a) shows agreement between simulation and ODE integration.

Secondary, we would like to test is the actual need to calculate expected value on the matrix Ξ when integrating the mean field in low dimension. In Figure 5 (b) we present the results for erf activation function: not taking into account the off-diagonal terms, the integration of the ODEs do not match the simulated dynamics even for large p. The effect becomes even more evident by comparing Figure 5 (c) with Figure 3 (a). In fact, we can see that in the case of quadratic activation, neglecting the matrix Ξ leads to a mismatch of population risk even at the initial instant, as it follows naturally from Equation (29). We do not see such a marked difference in the case of $\sigma = \mathrm{erf}(\cdot/\sqrt{2})$ since the latter is an odd function and non-diagonal terms of Ξ are symmetric random variables.

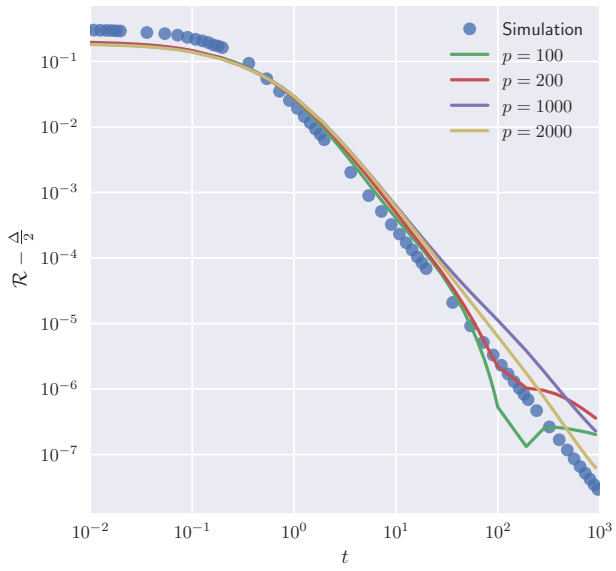
Finally, we looked at the evolution of the distribution of student weights. We use use $k = 1$ to have just one teacher vector $\boldsymbol{w}^{\star} \in \mathbb{R}^d$, while $p = 5000$ to have enough sample to understand how the distribution is evolving. In particular, we are looking at the distribution of the cosines of the angles between student weights and $\boldsymbol{w}^{\star}$; in Figure 5 (d) we plot this distribution a 3 different times of the evolution. At $t = 0$, the distribution is uniform, as expected since we sample weights from a spherical invariant distribution with $d = 3$. During the evolution, the student's weight moves in the direction of $\boldsymbol{w}^{\star}$ and $-\boldsymbol{w}^{\star}$ (since σ is an even function the sign of the weights does not count). At large time, the distribution is essentially $1/2\,(\delta_{\boldsymbol{w}^{\star}} + \delta_{-\boldsymbol{w}^{\star}})$, which represents perfect learning.



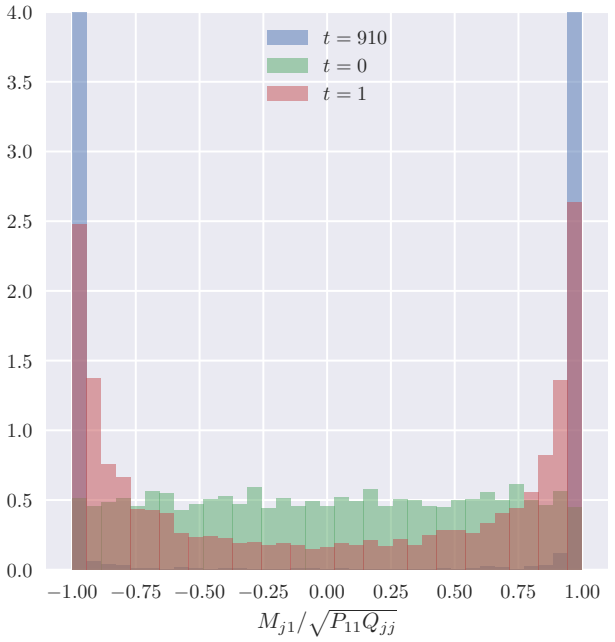
(a) $k = 2, d = 10, \gamma = 1.0, \Delta = 0.0, \sigma(x) = \text{erf}(x/\sqrt{2})$



(b) $k = 2, d = 5, \gamma = 1.0, \Delta = 0.0, \sigma(x) = \text{erf}(x/\sqrt{2})$



(c) $k = 2, d = 5, \gamma = 1.0, \Delta = 0.0, \sigma(x) = x^2$



(d) $k = 1, d = 3, \gamma = 1.0, \Delta = 0.0, \sigma(x) = x^2$

Figure 5: (a) Evolution from starting point $Q^{\perp} = 0$, tracking only M. (b) - (c) Low-dimension simulations and the corresponding ODE trajectory, without the off-diagonal entries of matrix Ξ (d) Histogram for the distribution of teacher student-teacher weights overlappings, at different times.