Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová

Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks

Rodrigo Veiga[1,3], Ludovic Stephan[1], Bruno Loureiro[1], Florent Krzakala[1], and Lenka Zdeborová[2]

[1]Ecole Polytechnique Fédérale de Lausanne (EPFL). Information, Learning and Physics (IdePHICS) lab. CH-1015 Lausanne, Switzerland.
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL). Statistical Physics of Computation (SPOC) lab. CH-1015 Lausanne, Switzerland.
[3]Universidade de São Paulo. Instituto de Física. São Paulo, SP, Brazil.

Abstract

Despite the non-convex optimization landscape, over-parametrized shallow networks are able to achieve global convergence under gradient descent. The picture can be radically different for narrow networks, which tend to get stuck in badly-generalizing local minima. Here we investigate the cross-over between these two regimes in the high-dimensional setting, and in particular investigate the connection between the so-called mean-field/hydrodynamic regime and the seminal approach of Saad & Solla. Focusing on the case of Gaussian data, we study the interplay between the learning rate, the time scale, and the number of hidden units in the high-dimensional dynamics of stochastic gradient descent (SGD). Our work builds on a deterministic description of SGD in high-dimensions from statistical physics, which we extend and for which we provide rigorous convergence rates.

Contents

1 Introduction

Descent-based algorithms such as stochastic gradient descent (SGD) and its variants are the workhorse of modern machine learning. They are simple to implement, efficient to run and most importantly: they work well in practice. A detailed understanding of the performance of SGD is a major topic in machine learning. Quite recently, significant progress was achieved in the context of learning in shallow neural networks. In a series of works, it was shown that the optimisation of wide two-layer neural networks can be mapped to a convex problem in the space of probability distributions over the weights [1, 2, 3, 4]. This remarkable result implies global convergence of 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 disposition. This line of work is commonly referred to as the mean-field or the hydrodynamic limit of neural networks. Mathematically, these works showed that one could describe the entire dynamics using a partial differential equation (PDE) in d dimensions.

In a different, and older, line of work one-pass SGD for two-layer neural networks with a *finite number* p of hidden units, synthetic Gaussian input data and teacher-generated labels has been widely studied starting with the seminal work of [5]. These works consider the limit of high-dimensional data and show, in particular, that the stochastic process driven by gradient updates converge to a set of p^2 deterministic ordinary differential equations (ODEs) as the input dimension $d \to \infty$ and the learning rate is proportional to $1/d$. The validity of these ODEs in this limit was proven by [6]. However, the picture drawn from the analysis of these ODEs is slightly different from the mean-field/hydrodynamic picture: in this case SGD can get stuck for long time in minima associated to no specialization of the hidden units to the teacher hidden units, and even when it converges to specializing minima, it fails to perfectly learn (i.e. to achieve zero population risk). In fact, in this analysis, the interplay between the limit of the learning rate going to zero and $d \to \infty$ appeared to be fundamental.

One should naturally wonder about the link between these two sets of works with, on the one hand a d-dimensional PDE (with large p), and on the other a p^2-dimensional ODE (with large d). In this work we aim to build a bridge between these two approaches for studying one-pass SGD.

Our starting point is the framework from [5], which we build upon and expand to a much broader range of choices of learning rate, time scales, and hidden layer width. This allows us to provide a sharp characterisation of the performance of SGD for two-layer neural networks in high-dimensions. We show it depends on the precise way in which the limit is taken, and in particular on how the quantity of data, the hidden layer width, and the learning rate scale as $d \to \infty$. For different choices of scaling, we can observe scenarios such as perfect learning, imperfect learning with an unavoidable error, or even no learning at all.

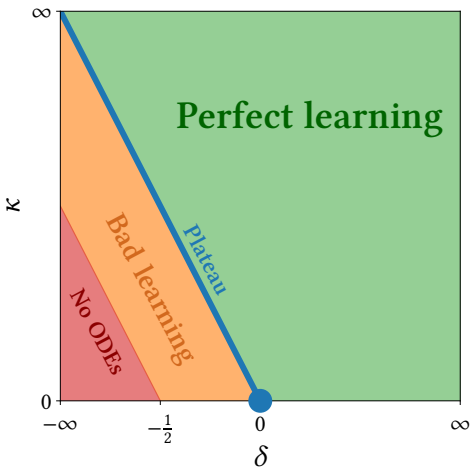
As a consequence of our analysis, we provide a phase diagram (see Figure 1a) describing the possible scenarios arising in the high-dimensional setting. Our **main contributions** are as follow:

C1 We rigorously show that the dynamics of SGD can be captured by a set of deterministic ODEs, considerably extending the proof of [6] to accommodate for general time scalings defined by an arbitrary learning rate, and a general range of hidden layer width. We provide much finer non-asymptotic guarantees which are crucial for our subsequent analysis.

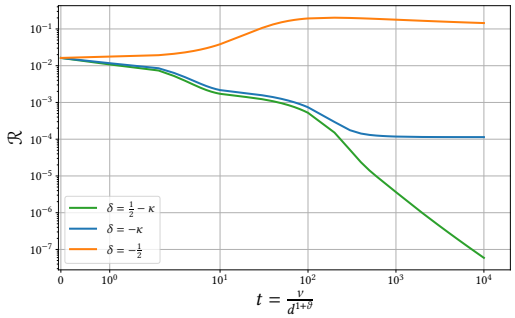
C2 From the analysis of the ODEs, we derive a phase diagram of SGD for two-layer neural networks in the high-dimensional input layer limit $d \to \infty$. In particular, scaling both the learning rate γ and hidden layer width p with the input dimension d as

$$\gamma \propto d^{-\delta} \,, \tag{1a}$$

$$p \propto d^{\kappa} \,, \tag{1b}$$



(a) The phase diagram of SGD learning regimes for two-layer neural networks in the high-dimensional input layer limit $d \to \infty$. Eqs. (1) define proper time scalings for each of the regions. Perfect learning region: $\kappa + \delta > 0$. Plateau line: $\kappa + \delta = 0$. Bad learning region: $-1/2 < \kappa + \delta < 0$. No ODEs region: $\kappa + \delta < -1/2$.



(b) A solution of the ODEs in all regions of Figure 1a, with matching colors. Parameters $\kappa = 0.301$, $p = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Noise: $\Delta = 10^{-3}$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbb{1})$. The time scaling is not uniform through the phase diagram: $\vartheta = \kappa + \delta$ on green and blue regimes and $\vartheta = 2(\kappa + \delta)$ on the orange region. The green curve decays as a power law to zero excess error.

Figure 1: Phase diagram (left) and typical behavior of the ODE in each regions (right).

we identify four different learning regimes which are summarized in Figure 1a:

- Perfect learning (green region, $\kappa > -\delta$): we show that perfect learning (zero population risk) can be asymptotically achieved with $n \sim d^{1+\kappa+\delta}$ samples even for tasks with additive noise.
- Plateau (blue line $\kappa = -\delta$): learning reaches a plateau related to the noise strength. The point $\kappa = \delta = 0$ goes back to the classical work of [5].
- Bad learning (orange region $-1/2 < \kappa + \delta < 0$): here the noise dominates the learning process.
- No ODEs (red region $\kappa + \delta < -1/2$): the stochastic process associated to SGD is not guaranteed to converge to a set of deterministic ODEs. This region is thus outside the scope of our analysis.

To better illustrate this phase diagram we present in Figure 1b a solution of the ODEs in all three regimes.

Relation to previous work – Deterministic dynamical descriptions of one-pass stochastic gradient descent in high-dimensions have a long tradition in the statistical physics community, starting with single- and two-layer neural networks with few hidden units [7, 8, 9, 10, 11]. The seminal work by [5] overcame previous limitations by constructing a set of deterministic ODEs for two-layer networks with any finite number of hidden units, paving the way for a series of important contributions [12, 13, 14, 6]. This line of work corresponds to the $\kappa = \delta = 0$ case of Figure 1a. One of our goal is to generalize this picture beyond fixed hidden layer size and learning rate.

A more recent line of work investigating the dynamics of SGD is the so-called *mean-field limit* [1, 15, 2, 3, 4], which connects the SGD dynamics of large-width two-layer neural networks to a diffusion equation

in the hidden layer weight density. In particular, [15] provide non-asymptotic convergence bounds for sufficiently small learning rates, corresponding to the green region of Figure 1a (with $p \to \infty$). The mean-field approach computes the empirical distribution (in $\mathbb{R}^d$) of the hidden layer weights, while we focus on the macroscopic overlaps between the teacher and student weights.

Reproducibility A code is provided at https://github.com/rodsveiga/phdiag_sgd.

2 Setting

Consider a supervised learning regression task. The data set is composed of n pairs $(\boldsymbol{x}^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^{d+1}$ identically and independently sampled from $\mathbb{P}(\boldsymbol{x}, y)$. The probability $\mathbb{P}(\boldsymbol{x})$ is assumed to be known and $\mathbb{P}(y|\boldsymbol{x})$ is modelled by a two layer neural network called the *teacher*. Given a feature vector $\boldsymbol{x}^\nu \in \mathbb{R}^d$, the respective label $y^\nu \in \mathbb{R}$ is defined as the output of a network with k hidden units, fixed weights $\boldsymbol{W}^* \in \mathbb{R}^{k\times d}$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$y^\nu = f(\boldsymbol{x}^\nu, \boldsymbol{W}^*) + \sqrt{\Delta}\zeta^\nu \,, \tag{2}$$

where

$$f(\boldsymbol{x}^\nu, \boldsymbol{W}^*) = \frac{1}{k}\sum_{r=1}^{k} \sigma\left(\frac{\boldsymbol{w}_r^{*\top}\boldsymbol{x}^\nu}{\sqrt{d}}\right) = \frac{1}{k}\sum_{r=1}^{k} \sigma(\lambda_r^{*\nu}) \,, \tag{3}$$

with $\boldsymbol{w}_r^* \equiv [\boldsymbol{W}^*]_r \in \mathbb{R}^d$ as the r-th row of the matrix $\boldsymbol{W}^*$ and $\lambda_r^{*\nu} \equiv \boldsymbol{w}_r^{*\top}\boldsymbol{x}^\nu/\sqrt{d} \in \mathbb{R}$ as the r-th component of the teacher *local field* vector $\boldsymbol{\lambda}^{*\nu} \in \mathbb{R}^k$. The parameter $\Delta \geq 0$ controls the strength of additive label noise: $\zeta^\nu \sim \mathbb{P}(\zeta^\nu)$ such that $\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}[\zeta] = 0$ and $\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}[\zeta^2] = 1$.

Given a new sample $\boldsymbol{x} \sim \mathbb{P}(\boldsymbol{x})$ outside the training data, the goal is to obtain an estimation $\hat{f}(\boldsymbol{x})$ for the respective label y. The error is quantified by a loss function $\mathcal{L}(y, \hat{f}(\boldsymbol{x}, \boldsymbol{\Theta}))$, where $\boldsymbol{\Theta}$ is an arbitrary set of parameters to be learned from data.

In this manuscript we are interested in the problem of estimating $\boldsymbol{W}^*$ with another two-layer neural network with the same activation function, which we will refer to as the *student*. The student network has p hidden units and a matrix of weights $\boldsymbol{W} \in \mathbb{R}^{p\times d}$ to be *learned* from the data. Given a feature vector $\boldsymbol{x} \sim \mathbb{P}(\boldsymbol{x})$ the student prediction for the respective label is given as

$$\hat{f}(\boldsymbol{x}, \boldsymbol{W}) = \frac{1}{p}\sum_{j=1}^{p} \sigma\left(\frac{\boldsymbol{w}_j^\top \boldsymbol{x}}{\sqrt{d}}\right) = \frac{1}{p}\sum_{j=1}^{p} \sigma(\lambda_j^\nu) \,, \tag{4}$$

where $\boldsymbol{w}_j \equiv [\boldsymbol{W}]_j \in \mathbb{R}^d$ is the j-th row of the matrix $\boldsymbol{W}$ and $\lambda_j \equiv \boldsymbol{w}_j^\top \boldsymbol{x}/\sqrt{d} \in \mathbb{R}$ is defined as j-th component of the student *local field* vector $\boldsymbol{\lambda} \in \mathbb{R}^p$.

One-pass gradient descent – Typically, one minimizes the *empirical risk* over the full data set. Instead, learning with *one-pass* gradient descent minimizes directly the *population risk*:

$$\mathcal{R}(\boldsymbol{W}, \boldsymbol{W}^*) \equiv \mathbb{E}_{\boldsymbol{x},y\sim\mathbb{P}(\boldsymbol{x},y)}\left[\mathcal{L}\left(f(\boldsymbol{x}, \boldsymbol{W}^*), \hat{f}(\boldsymbol{x}, \boldsymbol{W})\right)\right] \,. \tag{5}$$

Given a *single* sample $(\boldsymbol{x}^\nu, y^\nu)$ the weights are updated sequentially by the gradient descent rule:

$$\boldsymbol{w}_j^{\nu+1} = \boldsymbol{w}_j^\nu - \gamma \boldsymbol{\nabla}_{\boldsymbol{w}_j} \mathcal{L}\left(y^\nu, \hat{f}(\boldsymbol{x}^\nu, \boldsymbol{W})\right) \,, \tag{6}$$

with $\nu \in [n]$ and $j \in [p]$. The parameter $\gamma > 0$ is the learning rate. Despite being a simplification with respect to batch learning, one-pass gradient descent is an amenable surrogate for the theoretical analysis of non-convex optimization, since at each step the gradient is computed with a fresh data sample, which is equivalent to performing SGD directly on the population risk.

In particular, in this manuscript we assume realizability $p \geq k$, and focus our analysis on the square loss $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$, leading to

$$\boldsymbol{w}_j^{\nu+1} = \boldsymbol{w}_j^{\nu} + \frac{\gamma}{p\sqrt{d}} \sigma'(\lambda_j^{\nu})\, \mathcal{E}^{\nu}\, \boldsymbol{x}^{\nu} \,, \tag{7}$$

where

$$\mathcal{E}^{\nu} \equiv \frac{1}{k} \sum_{r=1}^{k} \sigma(\lambda_r^{*\nu}) - \frac{1}{p} \sum_{l=1}^{p} \sigma(\lambda_l^{\nu}) + \sqrt{\Delta}\zeta^{\nu} \,. \tag{8}$$

with population risk given by

$$\mathcal{R}(\boldsymbol{W}, \boldsymbol{W}^*) = \frac{1}{2}\, \mathbb{E}_{\boldsymbol{x}, y \sim \mathbb{P}(\boldsymbol{x}, y)} \left[\left(\hat{f}(\boldsymbol{x}, \boldsymbol{W}) - f(\boldsymbol{x},\ \boldsymbol{W}^*) \right)^2 \right] \,. \tag{9}$$

Therefore, from the above expression we can see that to monitor the population risk along the learning dynamics it is sufficient to track the joint distribution of the local fields $(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$. For Gaussian data $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \mathbb{1})$, one can replace the expectation $\mathbb{E}_{\boldsymbol{x}, y \sim \mathbb{P}(\boldsymbol{x}, y)}[\cdot]$ by $\mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \Omega)}[\cdot]$ and fully describe the dynamics through the following sufficient statistics, known in the statistical physics literature as *macroscopic variables*:

$$\boldsymbol{Q}^{\nu} \equiv \mathbb{E}_{\boldsymbol{x}, y \sim \mathbb{P}(\boldsymbol{x}, y)} \left[\boldsymbol{\lambda}^{\nu} \boldsymbol{\lambda}^{\nu \top} \right] = \frac{1}{d} \boldsymbol{W}^{\nu} \boldsymbol{W}^{\nu \top} \,, \tag{10a}$$

$$\boldsymbol{M}^{\nu} \equiv \mathbb{E}_{\boldsymbol{x}, y \sim \mathbb{P}(\boldsymbol{x}, y)} \left[\boldsymbol{\lambda}^{\nu} \boldsymbol{\lambda}^{*\nu \top} \right] = \frac{1}{d} \boldsymbol{W}^{\nu} \boldsymbol{W}^{* \top} \,, \tag{10b}$$

$$\boldsymbol{P} \equiv \mathbb{E}_{\boldsymbol{x}, y \sim \mathbb{P}(\boldsymbol{x}, y)} \left[\boldsymbol{\lambda}^{*\nu} \boldsymbol{\lambda}^{*\nu \top} \right] = \frac{1}{d} \boldsymbol{W}^{*} \boldsymbol{W}^{* \top} \,. \tag{10c}$$

with matrix elements, called *order parameters* in the statistical physics literature, denoted by $q_{jl}^{\nu} \equiv [\boldsymbol{Q}^{\nu}]_{jl}$, $m_{jr}^{\nu} \equiv [\boldsymbol{M}^{\nu}]_{jr}$ and $\rho_{rs} \equiv [\boldsymbol{P}]_{rs}$. The *macroscopic state* of the system at the learning step ν is given by the *overlap matrix* $\Omega^{\nu} \in \mathbb{R}^{(p+k)\times(p+k)}$:

$$\Omega^{\nu} = \begin{bmatrix} \boldsymbol{Q}^{\nu} & \boldsymbol{M}^{\nu} \\ \boldsymbol{M}^{\nu \top} & \boldsymbol{P} \end{bmatrix} \,, \tag{11}$$

and the population risk is completely determined by the macroscopic state:

$$\mathcal{R}(\Omega) = \frac{1}{2}\, \mathbb{E}_{\boldsymbol{\lambda}, \boldsymbol{\lambda}^* \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\lambda}^* | \boldsymbol{0}, \Omega)}\, \mathbb{E}_{\zeta \sim \mathbb{P}(\zeta)} \left[\left(\hat{f}(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda}^*) \right)^2 \right] \,. \tag{12}$$

The training dynamics (6) defines a discrete-time stochastic process for the evolution of the overlap matrix

$$\left\{ \Omega^{\nu} \in \mathbb{R}^{(p+k)\times(p+k)} \,, \nu \in [n] \right\} \,, \tag{13}$$

with $\boldsymbol{P}$ fixed and $\boldsymbol{Q}^{\nu}$ and $\boldsymbol{M}^{\nu}$ updated as:

$$q_{jl}^{\nu+1} - q_{jl}^{\nu} = \frac{\gamma}{pd} \underbrace{\left(\mathcal{E}_j^{\nu}\, \lambda_l^{\nu} + \mathcal{E}_l^{\nu}\, \lambda_j^{\nu} \right)}_{\text{learning}} + \frac{\gamma^2 \|\boldsymbol{x}\|^2}{p^2 d^2} \underbrace{\mathcal{E}_j^{\nu}\, \mathcal{E}_l^{\nu}}_{\text{variance}} \,, \tag{14a}$$

$$m_{jr}^{v+1} - m_{jr}^{v} = \frac{\gamma}{pd} \underbrace{\mathcal{E}_j^v \, \lambda_r^{*v}}_{\text{learning}} , \tag{14b}$$

with $v \in [n]$, $j, l \in [p]$, $r \in [k]$ and $\mathcal{E}_j^v \equiv \sigma'(\lambda_j^v)\, \mathcal{E}^v$. In what follows, we will make the concentration assumption $\|\boldsymbol{x}\|^2 = d$; this will be justified in the proof of Theorem 3.1.

We emphasize in (14) the specific role played by each term in the right hand-side. The "learning" terms are the fundamental ones, that actually drive the learning of the teacher by the student. We show in Appendix C.3 that these "learning" terms are identical to those obtained in the gradient flow approximation of SGD, whose performance is the topic of many works [1, 2, 3, 4]. Those are *precisely* the terms that draw the population risk towards zero. However, in our setting there is an additional variance term (so that this flow approximation is incomplete) that corresponds to the fluctuations of $\mathcal{L}(\boldsymbol{x}, \boldsymbol{W}, \boldsymbol{W}^*)$ around its expected value $\mathcal{R}(\boldsymbol{W}, \boldsymbol{W}^*)$. In particular, this is where the effects of the noise ζ can be felt. These terms were sometimes denoted as (I_2) and (I_4) in [16]. We shall see that the additional "variance" term is the one responsible for the plateau in the critical (blue) region of Figure 1a, while its contribution vanishes in the perfect learning (green) region.

Additionally, albeit our work particularizes to Gaussian input data, we believe our conclusion, and the phase diagram discussed in Figure 1a, to hold beyond this restricted case. Indeed, while the Gaussian assumption is crucial to reach a particular set of ODEs and their analytic expression, the approach can be applied to more complex data distribution, as long as one can track the sufficient statistics required to have a closed set of equations. For instance, [17] obtained very similar equations for an arbitrary mixture of Gaussians – that would obey the same scaling analysis as ours – while [18, 19, 20] proved that many complex distributions behave as Gaussians in high-dimensional setting, including, e.g. realistic GAN-generated data. We thus expect our conclusions to be robust in this respect.

3 Main results

Although $t_0 = v/d$ seems to be the most natural time scaling in the high-dimensional limit $d \to \infty$, if γ and p are allowed to vary with d the right-hand side (RHS) of Eqs. (14) can diverge and render the ODE approximation obsolete. Instead, for a given time scaling δt, we can rewrite Eqs. (14) as

$$\frac{q_{jl}^{v+1} - q_{jl}^{v}}{\delta t} = \frac{\gamma}{pd\,\delta t} \left(\mathcal{E}_j^v \, \lambda_l^v + \mathcal{E}_l^v \, \lambda_j^v \right) + \frac{\gamma^2}{p^2\, d\, \delta t} \, \mathcal{E}_j^v \, \mathcal{E}_l^v \, , \tag{15a}$$

$$\frac{m_{jr}^{v+1} - m_{jr}^{v}}{\delta t} = \frac{\gamma}{pd\,\delta t} \, \mathcal{E}_j^v \, \lambda_r^{*v} \, . \tag{15b}$$

In Theorem 3.1 we prove that as $d \to \infty$, Ω^v converges to the solution of the ODE:

$$\frac{d}{dt} \bar{\Omega}(t) = \psi\left(\bar{\Omega}(t)\right) \, , \tag{16}$$

where $\psi : \mathbb{R}^{(p+k)\times(p+k)} \to \mathbb{R}^{(p+k)\times(p+k)}$ is the expected value of the RHS of Eqs. (15), provided that this solution stays bounded. This enhances the result of [6] by providing convergence rates to the ODEs encompassing all scalings adopted hereafter:

Theorem 3.1 (Deterministic scaling limit of stochastic processes)**.** *Let $\tau \in \mathbb{R}$ be the continuous time horizon and $\delta t = \delta t(d)$ be a time scaling factor such that the following assumptions hold:*

1. *the time scaling* δt *satisfies for some constant* c,

$$\delta t \geq c \max\left(\frac{\gamma}{pd}, \frac{\gamma^2}{p^2 d}\right) \tag{17}$$

2. *the activation function* σ *is* L*-Lipschitz,*

3. *the function* $\psi : \mathbb{R}^{(p+k)\times(p+k)} \to \mathbb{R}^{(p+k)\times(p+k)}$ *is* L'*-Lipschitz.*

Then, there exists a constant $C > 0$ *(depending on* c, L, L'*) such that for any* $0 \leq v \leq \lfloor \tau/\delta t \rfloor$*, the following inequality holds:*

$$\mathbb{E}\left\|\Omega^v - \bar{\Omega}\left(v\delta t\right)\right\|_\infty \leq e^{C\tau} \log(p)\sqrt{\delta t}\ . \tag{18}$$

Our proof is based on techniques introduced in [21] (namely, their Lemma 2) which studies a different problem with related proof techniques. The proof involves decomposing Ω^{v+1} as

$$\Omega^{v+1} = \Omega^v + \delta t \psi(\Omega^v) + \left(\Omega^{v+1} - \Omega^v - \delta t \psi(\Omega^v)\right)\ , \tag{19}$$

where the two first terms can be considered as a deterministic discrete process, and the last term is a martingale increment. The main challenge lies in showing that the martingale contribution stays bounded throughout the considered time period.

Although the method is similar to [6], there are a number of differences between the two approaches. First, our proof fixes a number of holes in [6], in particular bounding q^v_{jj} by a sufficiently slowly diverging function of v. Additionally, the techniques used in this paper yield a dependency in p that is nearly negligible, while the previous methods imply bounds that are much too coarse for our needs.

The function ψ can be computed explicitly for various choices of σ, which allows to check Assumption 3 directly. We provide in Appendix C the necessary computations for $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$; those for the ReLU unit can be found in [22]. It can be checked that in the ReLU case, the function ψ is not Lipschitz around the matrices Ω satisfying

$$\Omega_{jl} = \sqrt{\Omega_{jj}\Omega_{ll}}$$

for some $j \neq l$. However, in every case we have a weaker square-root-Lipschitz property: there exists $C \in \mathbb{R}$ such that

$$\|\psi(\Omega) - \psi(\Omega')\| \leq C\left\|\sqrt{\Omega} - \sqrt{\Omega'}\right\|$$

for any Ω, Ω'. Since the square root function is Lipschitz whenever the eigenvalues of Ω are bounded away from zero (see e.g. [23]), Assumption 3 is implied by the condition

$$\Omega^v \geq \epsilon I_{p+k}\ ;$$

however, this assumption is much stronger, and becomes unrealistic in the specialization phase (as well as when $p \gg d$).

Theorem 3.1 allows us to safely navigate through Figure 1a by keeping track of convergence rates of the discrete process to a set ODEs. The interplay between learning rate and hidden layer width defines the time scaling δt and the trade-off between the linear contribution on $\mathcal{E}_j$ and the quadratic one, playing a central role on whether the network achieves perfect learning or not. Specifically, consider the following learning rate and hidden layer width scaling with d:

$$\gamma = \frac{\gamma_0}{d^\delta}\ , \tag{20a}$$

$$p = p_0 d^\kappa \,, \tag{20b}$$

where $\gamma_0 \in \mathbb{R}^+$ and $p_0 \in \mathbb{N}$ are constants. The exponent $\delta \in \mathbb{R}$ can be either greater or smaller than zero, while $\kappa \in \mathbb{R}^+$. Replacing these scalings on Eqs. (14), we find:

$$q_{jl}^{v+1} - q_{jl}^v = \frac{1}{d^{1+\kappa+\delta}} \underbrace{\left(\mathcal{E}_j^{\,v} \, \lambda_l^v + \mathcal{E}_l^{\,v} \, \lambda_j^v \right)}_{\text{learning}} + \frac{1}{d^{1+2(\kappa+\delta)}} \underbrace{\mathcal{E}_j^{\,v} \, \mathcal{E}_l^{\,v}}_{\text{noise}} \,, \tag{21a}$$

$$m_{jr}^{v+1} - m_{jr}^v = \frac{1}{d^{1+\kappa+\delta}} \underbrace{\mathcal{E}_j^{\,v} \, \lambda_r^{*v}}_{\text{learning}} \,, \tag{21b}$$

where we have chosen $\gamma_0 = p_0$ without loss of generality.

Since the distribution of the label noise $\mathbb{P}(\zeta)$ is such that $\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}[\zeta] = 0$, the linear contribution in $\mathcal{E}_j$ is noiseless in the high-dimensional limit $d \to \infty$, and therefore we will refer to it as the *learning term*. The noise enters in the equations through the variance computed on the quadratic contribution $\mathcal{E}_j \, \mathcal{E}_l$, which we will refer to as the *noise term*; intuitively, it is a high-dimensional variance correction which hinders learning. In order to satisfy (17), we shall take

$$\delta t = \max\left(\frac{1}{d^{1+\kappa+\delta}}, \frac{1}{d^{1+2(\kappa+\delta)}} \right) \,. \tag{22}$$

When $\kappa + \delta \neq 0$, this implies that either the learning term or the noise term scale like a negative power of d, and is negligible with respect to the other term. It is then easy to check that at a finite time horizon τ, the resulting ODEs behave as if the negligible term was not present. We refer to Theorem B.1 in the appendix for a quantitative proof of this phenomenon. Let us now describe the different regimes depicted in Figure 1a.

Blue line (plateau) – When γ and p are scaled such that $\kappa = -\delta$, Eqs. (21) converge to

$$\frac{dq_{jl}}{dt_0} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\mathcal{E}_j \, \lambda_l + \mathcal{E}_l \, \lambda_j \right] + \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \, \mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)} \left[\mathcal{E}_j \, \mathcal{E}_l \right] \,, \tag{23a}$$

$$\frac{dm_{jr}}{dt_0} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\mathcal{E}_j \, \lambda_r^* \right] \,, \tag{23b}$$

with $\delta t_0 \equiv 1/d$. This regime is an extension of [5] for which $\kappa = \delta = 0$. The convergence rate to the ODEs scales with $d^{-1/2}\log(d)$, and the phenomenology we observe for $\kappa = \delta = 0$ is consistent with previous works studying the setting $\kappa = \delta = 0$; namely the existence of an asymptotic plateau proportional to the noise level. For instance, the asymptotic population risk $\mathcal{R}_\infty$ is known to be proportional to $\gamma\Delta$ [6] when $\kappa = \delta = 0$ and the dynamics is driven by a rescaled version of Eqs. (23). Since the noise term does not vanish under this scaling, perfect learning to zero population risk is not possible. There is always an asymptotic plateau related to the noise level Δ, and the learning rate γ.

Green region (perfect learning) – If $\kappa > -\delta$ we can define the time scaling $\delta t_{\kappa+\delta} \equiv 1/d^{1+\kappa+\delta}$. By Theorem 3.1, Eqs. (21) converge to the following deterministic set of ODEs:

$$\frac{dq_{jl}}{dt_{\kappa+\delta}} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\mathcal{E}_j \, \lambda_l + \mathcal{E}_l \, \lambda_j \right] + \mathcal{O}\left(\frac{\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \, \mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)} \left[\mathcal{E}_j \, \mathcal{E}_l \right]}{d^{\kappa+\delta}} \right) \,, \tag{24a}$$

$$\frac{dm_{jr}}{dt_{\kappa+\delta}} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\mathcal{E}_j\,\lambda_r^*\right]\,, \tag{24b}$$

at a rate proportional to $d^{-(1+\kappa+\delta)/2}\log(d)$, where we have highlighted that the noise term vanishes with $d^{-(\kappa+\delta)}$. Hence, as long as $\kappa > -\delta$ the noise does not play any role on the dynamics. This setting could be understood by taking an effect learning rate $\gamma_{\text{eff}} \propto d^{-\kappa-\delta}$ on $\mathcal{R}_\infty \propto \gamma\Delta$, which leads to zero population risk, i.e. perfect learning, in the high dimensional limit $d \to \infty$. We validate this claim by a finite size analysis in the next section.

As discussed, the time scaling determines the number of data samples required to complete one learning step on the continuous scale. The bigger $\kappa+\delta$, the more attenuated the noise term, thus the closer to perfect learning. The trade-off is that the bigger $\kappa+\delta$, the larger the number of samples needed is, since $n = \tau d^{1+\kappa+\delta}$. Given a realizable learning task, one would thus rather choose the parameters to attain the perfect learning region, but being as close as possible to the plateau line for not increasing too much the needed number of samples. We remark that [15] provides an alternative deterministic approximation in this regime, with non-asymptotic bounds, whenever $p \gg 1$; this is the so-called mean-field approximation, with known convergence guarantees [2].

Orange region (bad learning) – We now step in the unusual situation where the learning rate grows faster with d than the hidden layer width: $\kappa < -\delta$. In this case, by (22) the noise term dominates over the dynamics. Defining the time scaling $\delta t_{2(\kappa+\delta)} \equiv 1/d^{1+2(\kappa+\delta)}$, we have

$$\frac{dq_{jl}}{dt_{2(\kappa+\delta)}} = \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\mathcal{E}_l\right] + \mathcal{O}\left(\frac{\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\mathcal{E}_j\,\lambda_l + \mathcal{E}_l\,\lambda_j\right]}{d^{-(\kappa+\delta)}}\right)\,, \tag{25a}$$

$$\frac{dm_{jr}}{dt_{2(\kappa+\delta)}} = \mathcal{O}\left(\frac{\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\mathcal{E}_j\,\lambda_r^*\right]}{d^{-(\kappa+\delta)}}\right)\,. \tag{25b}$$

According to Theorem 3.1 the convergence rate of Eqs. (21) to Eqs. (25) scales with $d^{-(1/2+\kappa+\delta)}\log(d)$. Therefore the existence of the noisy ODEs above is circumscribed to the region

$$-1/2 < \kappa + \delta < 0\,, \tag{26}$$

and presents a convergence trade-off absent in the other regimes: the faster one of the contributions of Eqs. (21) goes to zero, the worse is the convergence rate. In the present case, the more the learning term is attenuated, i.e. the more negative is $\kappa+\delta$, the worse the dynamics is described by Eqs. (25). Although the weights are updated, the correlation between the teacher and the student weights parametrized by the overlap matrix $\boldsymbol{M}$ remains fixed on its initial value $\boldsymbol{M}^0$, which is a fixed point of the dynamics under this scaling. Unsurprisingly, this leads to poor generalization capacity.

Red region (no ODEs) – If $\kappa+\delta < -1/2$, the stochastic process driven by the weight dynamics does not converge to deterministic ODEs under the assumptions of Theorem 3.1. We are then not able to state any claim about this regime.

Initialization and convergence – There are two additional features worth commenting on the high-dimensional dynamics and its connection to the mean-field/hydrodynamic approach, regarding initialization and the specialization transition.

In the ODE approach we discuss here, we always observe a first plateau where the teacher-student overlaps are all the same. This means all the hidden layer neurons learned the same linear separator. At this point, the two-layer network is essentially linear. This is called a *unspecialized* network in [16, 5]. In fact, this is a perfectly normal phenomenon, as with few samples even the Bayes-optimal solution would be unspecialized [24]. Only by running the dynamics long enough the student hidden neurons start to *specialize*, each of them learning a different sub-function so that the two-layer network can learn the non-trivial teacher.

Let us make two comments on this phenomenon: (i) while the "linear" learning in the unspecialized regime may remind the reader of the linear learning in the lazy regime [25, 26] of neural nets, the two phenomena are *completely* different. In lazy training, the learning is linear because weights change very little, so that the effective network is a linear approximation of the initial one. Here, instead, the weights are changing *considerably*, but each hidden neuron learns essentially the same function. (ii) If the ODEs are initialized with weights uncorrelated with the teacher, then the unspecialized regime is a fixed point of the ODEs: the student thus never specializes, at any time. Strikingly, such condition arises as well in the analysis of mean-field equations (see e.g. Theorem 2 in [27] that discusses the need to have *spread* initial conditions with a non-zero overlap with the teacher) to guarantee global convergence.

This raises the question about the precise dependence of the learning on the initialization condition in the high-dimensional regime, where a random start gets a vanishing $(1/\sqrt{d})$ overlap. This is a challenging problem that only recently has been studied (though in a simpler setting) in [28, 29, 30] who showed it yields an additional $\log(d)$ time-dependence. Generalizing these results for high-dimensional two-layer nets is an open question which we leave for future work.

4 Discussion, special cases, and simulations

To illustrate the phase diagram of Figure 1a, we present now several special cases for which we can perform simulations or numerically solve the set of ODEs. Henceforth, we take $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$, for which the expectations of the ODEs and of the population risk, Eq. (12), can be calculated analytically [5]. The explicit expressions are presented in Appendix C. Teacher weights are such that $\rho_{rs} = \delta_{rs}$. The initial student weights are chosen such that the dimension d can be varied without changing the initial conditions $\boldsymbol{Q}^0$, $\boldsymbol{M}^0$, $\boldsymbol{P}$ and consequently the initial population risk $\mathcal{R}_0$. A detailed discussion can be found in Appendix D.

4.1 Saad & Solla scaling $\kappa = \delta = 0$

We start by recalling the well-known setting characterized by the point $\kappa = \delta = 0$. The convergence of the stochastic process for fixed learning rate and hidden layer width to Eqs. (23) was first obtained heuristically by [5]. In Figure 2 we recall this classical result by plotting the population risk dynamics for different noise levels. Dots represent simulations, while solid lines are obtained by integration of the ODEs, Eq. (23).

Learning is characterized by two phases after the initial decay. The first is the unspecialized plateau where all the teacher-student overlaps are approximately the same: $m_{jr} \approx m$. Waiting long enough, the dynamics reaches the *specialization* phase, where the student neurons start to *specialize*, i.e., their overlaps with one of the teacher neurons increase and consequently the population risk decreases. This specialization is discussed extensively in [5]. If $\Delta = 0$, the population risk goes asymptotically to zero. Instead, if $\Delta \neq 0$, the specialization phase presents a second plateau related to the noise Δ.

The asymptotic population risk $\mathcal{R}_\infty$ related to the second plateau is proportional to $\gamma\Delta$ [6] in the high-dimensional limit $d \to \infty$ with p finite. As mentioned in the previous section, the expectation over $\mathcal{E}_j\,\mathcal{E}_l$ in Eq. (23a) prevents one from obtaining zero population risk for a noisy teacher.

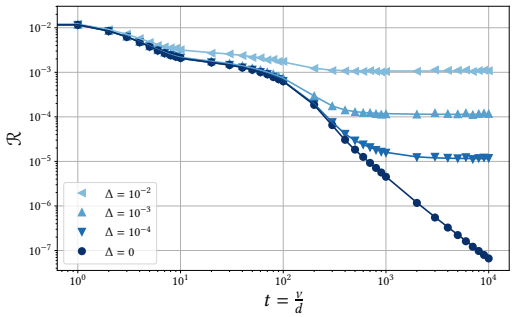
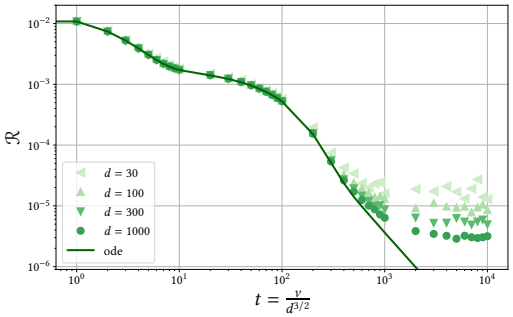
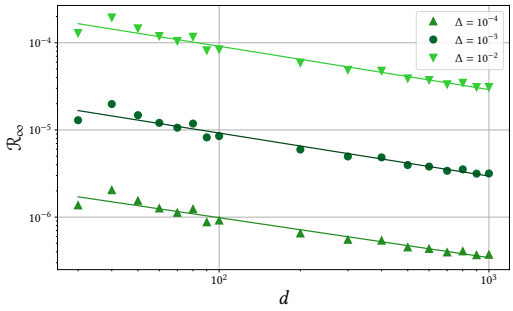


Figure 2: Population risk dynamics for $\kappa = \delta = 0$ (Saad & Solla scaling) : $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Activation: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbb{1})$. Dots represent simulations ($d = 1000$), while solid lines are obtained by integration of the ODEs given by Eqs. (23).



(a) Population risk dynamics for $\kappa = 0$ and $\delta = 1/2$. Fixed noise $\Delta = 10^{-3}$ and varying d. Dots represent simulations, while the solid line is obtained by integration of the ODEs given by Eqs. (24). The data are compatible with the claim that as $d \to \infty$ the curve converges to zero population risk.



(b) Asymptotic population risk $\mathcal{R}_\infty$ from simulations (dots) as a function of d for different noise levels under the scaling $\kappa = 0$ and $\delta = 1/2$. The fitted straight lines have slopes -0.458, -0.494, -0.497, for $\Delta = 10^{-4}, 10^{-3}, 10^{-2}$, respectively.

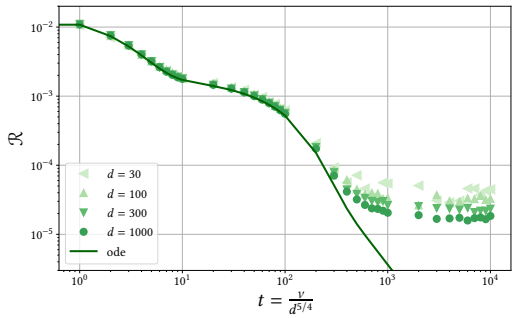
Figure 3: Network parameters: $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbb{1})$.

4.2 Perfect learning for $\kappa = 0$

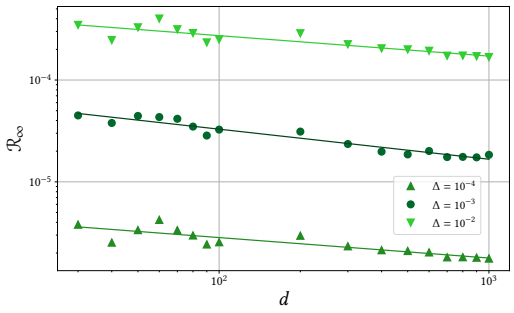
In this section we study the line $\kappa = 0$ with $\delta > 0$ of Figure 1a, for which Eqs. (24) with $\kappa = 0$ hold. We show that perfect learning can be asymptotically achieved in the realizable setting for any finite hidden layer width $p = p_0$. Keeping δ and Δ fixed, we have done simulations increasing the input layer dimension d. In Figure 3a we set $\delta = 1/2$, $\Delta = 10^{-3}$ and vary the input layer dimension. The bigger d is, the closer we are to the ODE-derived noiseless result.

Gathering the asymptotic population risk from simulations for varying d and Δ we perform a finite-size analysis to study the dependence of $\mathcal{R}_\infty$ with d. This shows that the noise term goes to zero under this setting. In Figure 3b we plot $\mathcal{R}_\infty$ versus d from simulations (dots) for different noise levels. We fit lines under the log-log scale showing that $\mathcal{R}_\infty \propto d^{-\delta}$, as expected. Figure 4 draws the same conclusion for $\delta = 1/4$.

As already stated, the interplay between the exponents directly affects the time scale. We end this subsection by graphically illustrating this fact through simulations. Setting the noise to $\Delta = 10^{-3}$ we

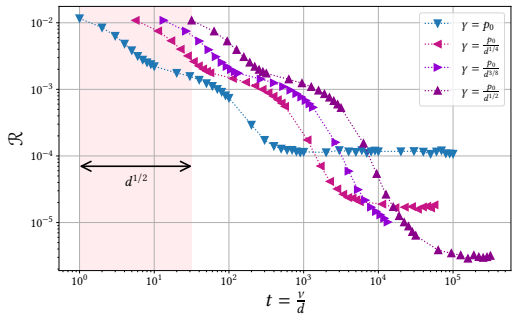


(a) Population risk dynamics for $\kappa = 0$ and $\delta = 1/4$. Fixed noise $\Delta = 10^{-3}$ and varying d. Dots represent simulations, while the solid line is obtained by integration of the ODEs given by Eqs. (24). The data are compatible with the claim that as $d \to \infty$ the curve converges to zero population risk.

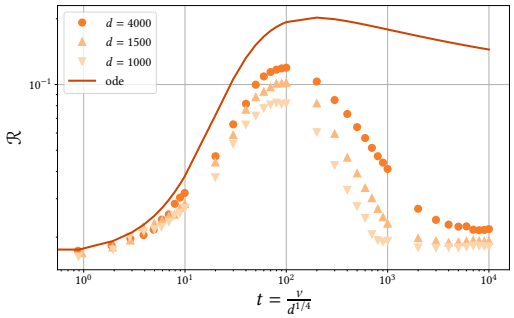


(b) Asymptotic population risk $\mathcal{R}_\infty$ from simulations (dots) as a function of d for different noise levels under the scaling $\kappa = 0$ and $\delta = 1/4$.The fitted straight lines have slopes -0.201, -0.295, -0.201, for $\Delta = 10^{-4}, 10^{-3}, 10^{-2}$, respectively.

Figure 4: Network parameters: $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \mathbb{1})$.



(a) Simulations ($d = 1000$) for $\kappa = 0$ comparing different choices of the exponent δ. The final plateau is proportional to learning rate: $\mathcal{R}_\infty \propto \gamma\Delta$.



(b) Population risk dynamics for $\kappa = 0$ and $\delta = -3/8$. Dots represent simulations, while the solid line is obtained by integration of the ODEs given by Eqs. (25).

Figure 5: Network parameters $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Noise level $\Delta = 10^{-3}$. Activation: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \mathbb{1})$.

compare the cases $\delta = 0, 1/4, 3/8, 1/2$ in Figure 5a. All simulations are rendered on the scale $\delta t_0 = 1/d$ to illustrate the trade-off between asymptotic performance and training time.

4.3 Bad learning for $\kappa = 0$

We now quickly discuss the uncommon case of γ growing with d within the orange region. In Figure 5b we compare simulations varying d with the solution of the ODEs given by Eqs. (25). Both lead to poor results compared to the green and blue regions. Moreover, this regime presents strong finite-size effects, making it harder to observe the asymptotic ODEs at small sizes. However, the trend as d increases is very clear from the simulations. As discussed in Section 3, the more the learning term is attenuated on the ODEs, the worse they describe the dynamics.

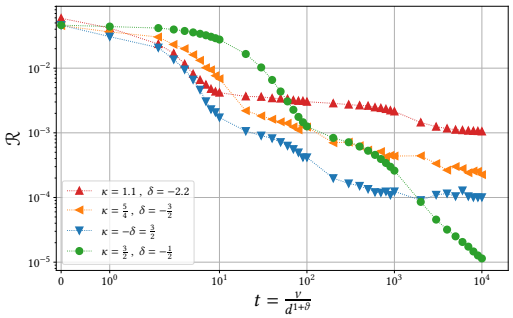


Figure 6: Simulations across different regions of Figure 1a. Networks parameters $d = 100$, $p = d^{\kappa}$, $\gamma = d^{-\delta}$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Noise: $\Delta = 10^{-3}$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \mathbb{1})$. Time scaling: $\vartheta = \kappa + \delta$ for green and blue and $\vartheta = 2(\kappa + \delta)$ for orange. The colors match Figure 1a.

4.4 Large hidden layer: $\kappa > 0$

Finishing our voyage through Figure 1a with examples, we briefly discuss the case where both input and hidden layer widths are large. Although Theorem 3.1 provides non-asymptotic guarantees for $\kappa > 0$, the number of coupled ODEs grows quadratically with p, making the task of solving them rather challenging. Thus, we present simulations that illustrate the regions of Figure 1a. Fixing $d = 100$ we show in Figure 6 learning curves for different values of κ and δ. The colors are chosen to match their respective regions in the phase diagram.

Due to the relatively small sizes used in Figure 6, the green dots seem to decrease towards perfect learning, even when $\delta < 0$, provided that κ is large enough, as is predicted by the phase diagram in Figure 1a. Moreover, since d is not large enough, when the parameters are within the orange region the finite-size effects actually dominates, similarly to Figure 5b. The learning contribution still plays a role and the asymptotic population risk is similar to the case $\kappa = \delta = 0$. Within the red region, which is out of scope of our theory, the simulation gets stuck on a plateau with larger population risk.

5 Conclusion

Building up on classical statistical physics approaches and extending them to a broad range of learning rate, time scales, and hidden layer width, we rendered a sharp characterisation of the performance of SGD for two-layer neural networks in high-dimensions. Our phase diagram describes the possible learning scenarios, characterizing learning regimes which had not been addressed by previous classical works using ODEs. Crucially, our key conclusions do not rely on an explicit solution, as our theory allows the characterization of the learning dynamics *without* solving the system of ODEs. The introduction of scaling factors is non-trivial and has deep implications. Our generalized description enlightens the trade-off between learning rate and hidden layer width, which has also been crucial in the mean-field theories.

Acknowledgements

We thank Gérard Ben Arous, Lenaïc Chizat, Maria Refinetti and Sebastian Goldt for discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Pro-

gram Grant Agreement 714608- SMiLe. RV was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. RV is grateful to EPFL and IdePHICS lab for their generous hospitality during the realization of this project.

Appendix

A Deterministic scaling limit of stochastic processes

In order to show the deterministic scaling of online SGD under a proper chosen time scale, we will make use of a convergence result by [21, 31], which is adapted below in Theorem A.1.

Theorem A.1 (Deterministic scaling limit of stochastic processes)**.** *Consider a d-dimension discrete time stochastic process sequence, $\{\Omega^\nu \ ; \ \nu = 0, 1, 2, ..., [S\tau]\}_{S=1,2,...}$ for some $\tau > 0$. The increment $\Omega^{\nu+1} - \Omega^\nu$ is assumed to be decomposable into three parts,*

$$\Omega^{\nu+1} - \Omega^\nu = \frac{1}{S}\psi(\Omega^\nu) + \Lambda^\nu + \Gamma^\nu \,, \tag{A.1}$$

such that

Assumption A.1.1. The process $\tilde{\Lambda}^\nu \equiv \sum_{\nu'=0}^{\nu} \Lambda^{\nu'}$ is a martingale and $\mathbb{E}\|\Lambda^\nu\|^2 \le C(\tau)^2/S^{1+\epsilon_1}$ for some $\epsilon_1 > 0$.

Assumption A.1.2. $\mathbb{E}\|\Gamma^\nu\| \le C(\tau)/S^{1+\epsilon_2}$ for some $\epsilon_2 > 0$.

Assumption A.1.3. The function $\psi(\Omega)$ is Lipschitz, i.e, $\|\psi(\Omega) - \psi(\tilde{\Omega})\| \le C\|\Omega - \tilde{\Omega}\|$ for any Ω and $\tilde{\Omega}$.

Let $\Omega(t)$, with $0 \le t \le \tau$, be a continuous stochastic process such that $\Omega(t) = \Omega^\nu$ with $\nu = [St]$. Define the deterministic ODE

$$\frac{d}{dt}\bar{\Omega}(t) = \psi(\bar{\Omega}(t)) \,, \tag{A.2}$$

with $\bar{\Omega}(0) = \bar{\Omega}_0$.

Then, if assumptions A.1.1 to A.1.3 hold and assuming $\mathbb{E}\|\Omega^0 - \bar{\Omega}_0\| < C/S^{\epsilon_3}$ for some $\epsilon_3 > 0$ then we have for any finite S:

$$\mathbb{E}\left\|\Omega^\nu - \bar{\Omega}\left(\frac{\nu}{S}\right)\right\| \le C(\tau)e^{c\tau}S^{-\min\{\frac{1}{2}\epsilon_1,\epsilon_2,\epsilon_3\}} \,, \tag{A.3}$$

where $\bar{\Omega}(\cdot)$ is the solution of Eq.(A.2).

Proof. The reader interested in the proof is referred to the supplementary materials of [21, 31]. □

Although the theorem wasn't originally proven in the $p \to \infty$ setting, a glance at its proof shows that it still holds upon replacing $C(\tau)$ by $C(p, \tau)$ in Assumption A.1.1 and A.1.2, as well as Equation (A.3). We choose $\|\cdot\|$ to be the L^∞ norm, since it suits better the $p \to \infty$ scaling. The S in Theorem A.1 corresponds to $1/\delta t$, where δt is defined in Theorem 3.1.

Following [21], we define for $j, l \in [p]$

$$\Psi_{jl}(\Omega; \boldsymbol{x}) = \frac{\gamma}{pd\,\delta t}\left(\mathcal{E}_j^\nu \lambda_l^\nu + \mathcal{E}_l^\nu \lambda_j^\nu\right) + \frac{\gamma^2}{p^2\,d\,\delta t}\,\mathcal{E}_j^\nu\,\mathcal{E}_l^\nu,$$

and

$$\psi_{jl}(\Omega) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\mathbf{0},\mathbb{1})}\left[\Psi_{jl}(\Omega; \boldsymbol{x})\right].$$

The functions Ψ, ψ are similarly defined on $[p] \times [p+1, p+k]$. With that, we write

$$\Omega^{v+1} - \Omega^v = \frac{1}{S}\psi(\Omega) + \underbrace{\frac{1}{S}\left(\Psi(\Omega^v; \boldsymbol{x}) - \psi(\Omega^v)\right)}_{\Lambda^v} + \Gamma^v,$$

where for $j, l \in [p]$

$$\Gamma^v_{jl} = \frac{\gamma^2}{p^2 d^2}\left(\|\boldsymbol{x}\|_2^2 - d\right)\mathcal{E}^v_j\,\mathcal{E}^v_l\,.$$

The main obstacle to bounding $\boldsymbol{\Lambda}^v$ and $\boldsymbol{\Gamma}^v$ is the fact that the q_{jj} can a priori diverge to infinity. Our first task is therefore to show that this does not happen; as a proxy we show a subgaussian-like moment bound:

$$\mathbb{E}\left[(q^v_{jj})^t\right] \le \left(C(\tau) + \frac{ct}{S}\right)^t.$$

Equipped with the above bound, controlling $\mathbb{E}\|\boldsymbol{\Lambda}^v\|^2$ and $\mathbb{E}\|\boldsymbol{\Gamma}^v\|$ becomes fairly easy. All proof details are in the below sections.

A.1 Preliminaries: bounding the q_{jj}

Since σ is L-Lipschitz, we have by the Cauchy-Schwarz inequality

$$(\mathcal{E}^v)^2 \le \frac{3L^2}{k}\sum_{r=1}^{k}(\lambda^*_r)^2 + \frac{3L^2}{p}\sum_{j=1}^{p}(\lambda_j)^2 + 3\Delta\zeta^2 \equiv \Phi^v \tag{A.4}$$

Define

$$s^v = \mathbb{E}\,\Phi^v = \frac{3L^2}{k}\sum_{r=1}^{k}\rho_{rr} + \frac{3L^2}{p}\sum_{j=1}^{p}q^v_{jj} + 3\Delta$$

Assumption 1 in Theorem 3.1 implies that

$$|q^{v+1}_{jj} - q^v_{jj}| \le \frac{1}{S}\left(c_1(\lambda^v_j)^2 + c_2(\mathcal{E}^v)^2\right)$$

where c_1, c_2 are absolute constants. Summing those inequalities yield

$$|s_{v+1} - s^v| \le \frac{c_3}{S}\Phi^v,$$

and finally

$$\mathbb{E}_v[s^{v+1}] \le s^v\left(1 + \frac{c_3}{S}\right) \le s^v e^{c_3/S}.$$

As a result, we have for any $0 \le v \le S\tau$

$$\mathbb{E}[s^v] \le c_4 e^{c_3\tau}. \tag{A.5}$$

For simplicity, let q^v denote any of the q^v_{jj}. We have, for all $t \ge 0$,

$$(q^{v+1})^t - (q^v)^t = t(q^v)^{t-1}(q^{v+1} - q^v) + O\left(\frac{t^2}{S^2}\right),$$

where the remainder term has bounded expectation. Again, we write

$$\left|(q^{\nu+1})^t - (q^\nu)^t\right| \le t(q^\nu)^{t-1}\frac{1}{S}(c_1(\mathcal{E}^\nu)^2 + c_2(\lambda_i^\nu)^2) + \frac{c_5 t^2}{S^2}.$$

By Assumption 3, the q_{ii}^ν are bounded from below by a constant, hence

$$\mathbb{E}_\nu[(q^{\nu+1})^t] \le (q^\nu)^t\left(1 + \frac{c_6 t}{S}\right) + O\left(\frac{c_5 t^2}{S^2}\right)$$

This implies that for any $t \ge 0$ and $0 \le \nu \le S\tau$,

$$\mathbb{E}[(q^\nu)^t] \le \left(c_7 + \frac{c_5 t^2}{S}\right)e^{c_6\tau} \le \left(C(\tau) + \frac{c_5 t}{S}\right)^t \tag{A.6}$$

A.2 Assumption A.1.1

We have for all $i, j \in [p+k]$,

$$\left(\Omega_{ij}^{\nu+1} - \mathbb{E}_\nu[\Omega_{ij}^{\nu+1}]\right)^2 \le 2\left((\Omega_{ij}^{\nu+1} - \Omega_{ij}^\nu)^2 + (\Omega_{ij}^\nu - \mathbb{E}_\nu[\Omega_{ij}^{\nu+1}])^2\right) .$$

As a consequence,

$$\mathbb{E}\|\mathbf{\Lambda}^\nu\|^2 \le 4\max_{i,j}(\Omega_{ij}^{\nu+1} - \Omega_{ij}^\nu)^2 .$$

Now, by definition,

$$(q_{ij}^{\nu+1} - q_{ij}^\nu)^2 \le \frac{L}{S^2}\left(c_1(\mathcal{E}^\nu)^2 + c_2|\,\mathcal{E}^\nu\,|(|\lambda_i| + |\lambda_j|)\right)^2 \le \frac{L}{S^2}\left(c_3(\mathcal{E}^\nu)^4 + c_4(\max_\ell \lambda_\ell^\nu)^4\right),$$

The term in $(\mathcal{E}^\nu)^4$ is bounded by the same techniques as the last section. For the second term,

$$\mathbb{E}_\nu\left[(\max_\ell \lambda_\ell)^4\right] \le c_5\log(p)^2\left(\max_\ell q_{\ell\ell}^\nu\right)^4,$$

and we can write for any $t \ge 0$

$$\max_\ell (q_{\ell\ell}^\nu)^4 \le \left(\sum_\ell (q_{\ell\ell}^\nu)^t\right)^{4/t}.$$

By Jensen's inequality, for $t \ge 4$

$$\mathbb{E}\left[\left(\max_\ell q_{\ell\ell}^\nu\right)^4\right] \le \left(\sum_\ell \mathbb{E}[(q_{\ell\ell}^\nu)^t\right)^{4/t} \le p^{4/t}\left(C(\tau) + \frac{c_6 t}{S}\right)^4,$$

using (A.6). Choosing $t = 4\log(p) \ll S$ shows that

$$\mathbb{E}\left[\max_{i,j}(q_{ij}^{\nu+1} - q_{ij}^\nu)^2\right] \le \frac{C(\tau)\log(p)^2}{S^2}$$

A similar bound holds for the m_{ij}, and hence

$$\mathbb{E}\|\mathbf{\Lambda}^\nu\|^2 \le \frac{c_5\log(p)^2}{S^2},$$

which implies Assumption A.1.1 with $\epsilon_1 = 1$ and $C(p,\tau) = C'(\tau)\log(p)$.

A.3 Assumption A.1.2

Since σ is Lipschitz, for any $i, j \in [p]$

$$\mathcal{E}_i^\nu \, \mathcal{E}_j^\nu \leq L^2 (\mathcal{E}^\nu)^2.$$

Hence,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] &\leq \frac{L^2\gamma^2}{d^2p^2} \mathbb{E}\left[\left(\|\boldsymbol{x}\|_2^2 - d\right)\Phi^\nu\right] \\
&\leq \frac{L^2\gamma^2}{d^2p^2}\left(\frac{1}{2\sqrt{d}} \mathbb{E}\left[\left(\|\boldsymbol{x}\|_2^2 - d\right)^2\right] + \frac{\sqrt{d}}{2}\mathbb{E}\left[(\mathcal{E}^\nu)^4\right]\right).
\end{aligned}$$

The first expectation is the variance of a χ_d^2 random variable, which is equal to $2d$, and the second expectation is bounded by the same methods as the above sections. The term in brackets is therefore bounded by $c_1\sqrt{d}$, and

$$\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] \leq c_2 \frac{\gamma^2}{d^{3/2}p^2}$$

Finally, since for any $y > 0$ we have $y^2 \leq \max(y, y^2)^{3/2}$, letting $y = \gamma/p$ we find

$$\mathbb{E}[\|\mathbf{\Gamma}^\nu\|_\infty] \leq c_2 \max\left(\frac{\gamma}{pd}, \frac{\gamma^2}{p^2 d}\right)^{3/2} \leq c_3(\delta t)^{3/2},$$

hence Assumption A.1.2 is true with $\epsilon_2 = 1/2$.

A.4 $\sqrt{}$-Lipschitz property

Let $\Omega, \Omega' \in \mathbb{R}^{(p+k)\times(p+k)}$, we can write the (i, j) coefficient of $\psi(\Omega)$ as $f_{ij}(\sqrt{\Omega})$, where

$$\begin{aligned}
f : \mathbb{R}^{(p+k)\times(p+k)} &\to \mathbb{R} \\
A &\mapsto \mathbb{E}_{x\sim\mathcal{N}(0,I_{p+k})}[g_{ij}(Ax)]
\end{aligned}$$

The same arguments as above show that the function f is Lipschitz, and hence for some constant L'' we have

$$\|\psi(\Omega) - \psi(\Omega')\| \leq L''\|\sqrt{\Omega} - \sqrt{\Omega'}\|.$$

B A lemma on ODE perturbation

In this section, we prove a proposition that bounds the difference between an ODE solution and a perturbed version, for a bounded time t.

Theorem B.1. *Let $f, g : \mathbb{R}^n \to \mathbb{R}^n$ be two L-Lipschitz functions, and consider the following differential equations in $\mathbb{R}^n$:*

$$\begin{aligned}
\frac{d\boldsymbol{x}}{dt} &= f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x}), \\
\frac{d\boldsymbol{y}}{dt} &= f(\boldsymbol{y}),
\end{aligned}$$

where $\epsilon > 0$, and with the initial condition $\boldsymbol{x}(0) = \boldsymbol{y}(0)$. Then, if $\tau > 0$ is fixed, we have

$$\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|_2 \leq c\epsilon e^{L\tau}$$

for any $0 \leq t \leq \tau$, with c a constant independent from ϵ, τ.

Before proving this proposition, we begin with a small lemma:

Lemma B.2. *Let $a, b > 0$, and $z : \mathbb{R}^+ \to \mathbb{R}^+$ a function satisfying*

$$\frac{dz}{dt} = az + b\sqrt{z}$$

with $z(0) = 0$. Then, for some constant $c > 0$, we have

$$z(t) \leq c\frac{b^2 e^{at}}{a^2} \quad \textit{for all} \quad t \geq 0$$

Proof. Upon considering the function $a^2 z(t/a)/b^2$ instead, we can assume that $a = b = 1$. Then, we have

$$\frac{dz}{dt} \leq \max(z, 1) + \max(\sqrt{z}, 1),$$

and the RHS is an increasing function. Hence, if $\tilde{z}$ is a solution of

$$\frac{d\tilde{z}}{dt} = \max(z, 1) + \max(\sqrt{\tilde{z}}, 1),$$

with $\tilde{z}(0) = 0$, then $z(t) \leq \tilde{z}(t)$ for all $t \geq 0$. Since the RHS of the above equation is Lipschitz everywhere, we can apply the Picard–Lindelöf theorem, and check that the unique solution to this equation is

$$\tilde{z}(t) = \begin{cases} 2t & \text{if } t \leq \frac{1}{2} \\ (c_1 e^t - c_2)^2 & \text{otherwise} \end{cases},$$

where c_1 and c_2 are ad hoc constants. The lemma then follows from adjusting the constant c as needed. □

We are now in a position to show Theorem B.1:

Proof. Assume for simplicity that $\boldsymbol{x}(0) = \boldsymbol{y}(0) = \boldsymbol{0}$. We begin by bounding $\boldsymbol{x}(t)$; we have

$$\frac{d\|\boldsymbol{x}\|^2}{dt} = 2\boldsymbol{x}^\top \frac{d\boldsymbol{x}}{dt} \leq 2\|\boldsymbol{x}\| \, \|f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x})\| \, .$$

By the Lipschitz condition,

$$\|f(\boldsymbol{x}) + \epsilon g(\boldsymbol{x})\| \leq \|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\| + \frac{L}{2}\|\boldsymbol{x}\| \, ,$$

so that

$$\frac{d\|\boldsymbol{x}\|^2}{dt} \leq L\|\boldsymbol{x}\|^2 + 2\|f(\boldsymbol{0}) + \epsilon g(\boldsymbol{0})\| \, \|\boldsymbol{x}\| \, .$$

Applying Lemma B.2 and taking square roots on each side,

$$\|\boldsymbol{x}(t)\| \leq c\frac{\|f(\mathbf{0})+\epsilon g(\mathbf{0})\|}{L}e^{Lt/2} \leq c\frac{\|f(\mathbf{0})+\epsilon g(\mathbf{0})\|}{L}e^{L\tau/2}\,, \tag{B.1}$$

for any $0 \leq t \leq \tau$. Now, similarly,

$$\begin{aligned}\frac{d\|\boldsymbol{x}-\boldsymbol{y}\|^2}{dt} &\leq 2\|\boldsymbol{x}-\boldsymbol{y}\|\left\|\frac{d(\boldsymbol{x}-\boldsymbol{y})}{dt}\right\| \\ &\leq 2\|\boldsymbol{x}-\boldsymbol{y}\|\,\|f(\boldsymbol{x})-f(\boldsymbol{y})+\epsilon g(\boldsymbol{x})\| \\ &\leq L\|\boldsymbol{x}-\boldsymbol{y}\|^2 + 2\epsilon\|g(\boldsymbol{x})\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \\ &\leq L\|\boldsymbol{x}-\boldsymbol{y}\|^2 + \epsilon\left(\|g(\mathbf{0})\| + c\|f(\mathbf{0})+\epsilon g(\mathbf{0})\|e^{L\tau/2}\right)\|\boldsymbol{x}-\boldsymbol{y}\|\,,\end{aligned}$$

having used (B.1) on the last line. This is again the setting of Lemma B.2, which gives

$$\|\boldsymbol{x}-\boldsymbol{y}\| \leq c_1\epsilon e^{L\tau/2}\frac{e^{Lt/2}}{L} \leq c_2\epsilon e^{L\tau}\,.$$

□

C Expectations over the local fields

In this appendix we present the explicit expressions from the expectations of the local fields used to compute the population risk and the ODE terms.

C.1 Population risk

We write the population risk (12) as

$$\begin{aligned}\mathcal{R}(\boldsymbol{\Omega}) &= \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\left(\hat{f}(\boldsymbol{\lambda})-f(\boldsymbol{\lambda}^*)\right)^2\right] \\ &= \mathcal{R}_{\mathrm{t}}(\boldsymbol{P})+\mathcal{R}_{\mathrm{s}}(\boldsymbol{Q})+\mathcal{R}_{\mathrm{st}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{M})\,,\end{aligned} \tag{C.1}$$

with

$$\mathcal{R}_{\mathrm{t}} \equiv \mathbb{E}_{\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{P})}\left[f(\boldsymbol{\lambda}^*)^2\right] = \frac{1}{k^2}\sum_{r,s=1}^{k}\mathbb{E}_{\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{P})}\left[\sigma(\lambda_r^*)\sigma(\lambda_s^*)\right] \tag{C.2a}$$

$$\mathcal{R}_{\mathrm{s}} \equiv \mathbb{E}_{\boldsymbol{\lambda}\sim\mathcal{N}(\boldsymbol{\lambda}|\mathbf{0},\boldsymbol{Q})}\left[\hat{f}(\boldsymbol{\lambda})^2\right] = \frac{1}{p^2}\sum_{j,l=1}^{k}\mathbb{E}_{\boldsymbol{\lambda}\sim\mathcal{N}(\boldsymbol{\lambda}|\mathbf{0},\boldsymbol{Q})}\left[\sigma(\lambda_j)\sigma(\lambda_l)\right]\,, \tag{C.2b}$$

$$\mathcal{R}_{\mathrm{st}} \equiv \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\hat{f}(\boldsymbol{\lambda})f(\boldsymbol{\lambda}^*)\right] = -\frac{2}{pk}\sum_{j=1}^{p}\sum_{r=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\sigma(\lambda_j)\sigma(\lambda_r^*)\right] \tag{C.2c}$$

Define the vector $\boldsymbol{\lambda}^{\alpha\beta} \equiv \left(\lambda^\alpha,\lambda^\beta\right)^\top \in \mathbb{R}^2$, where the upper indices on the components indicate they may refer to student or teacher local fields. Consider the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta}$:

$$\boldsymbol{\Omega}^{\alpha\beta} \equiv \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\boldsymbol{\Omega})}\left[\boldsymbol{\lambda}^{\alpha\beta}\left(\boldsymbol{\lambda}^{\alpha\beta}\right)^\top\right] \in \mathbb{R}^{2\times 2}\,. \tag{C.3}$$

For $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (C.2) are in general given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma(\lambda^\alpha)\sigma(\lambda^\beta)\right] = \frac{1}{\pi}\arcsin\left(\frac{\Omega_{12}^{\alpha\beta}}{\sqrt{\left(1+\Omega_{11}^{\alpha\beta}\right)\left(1+\Omega_{22}^{\alpha\beta}\right)}}\right). \tag{C.4}$$

where $\Omega_{jl}^{\alpha\beta} \equiv (\Omega^{\alpha\beta})_{jl}$ is an element of the covariance matrix given by Eq. (C.3).

Explicitly, the population risk contributions are

$$\mathcal{R}_{\mathrm{t}}(\boldsymbol{P}) = \frac{1}{k^2}\sum_{r,s=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{\rho_{rs}}{\sqrt{(1+\rho_{rr})\,(1+\rho_{ss})}}\right), \tag{C.5a}$$

$$\mathcal{R}_{\mathrm{s}}(\boldsymbol{Q}) = \frac{1}{p^2}\sum_{j,l=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{q_{jl}}{\sqrt{(1+q_{jj})\,(1+q_{ll})}}\right), \tag{C.5b}$$

$$\mathcal{R}_{\mathrm{st}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{M}) = -\frac{2}{pk}\sum_{j=1}^{p}\sum_{r=1}^{k}\frac{1}{\pi}\arcsin\left(\frac{m_{jr}}{\sqrt{(1+q_{jj})\,(1+\rho_{rr})}}\right). \tag{C.5c}$$

C.2 ODE contributions

From the update equations, we first consider the expectations linear in $\mathcal{E}_j$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\lambda_l\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_{r'}^*)\right]\\ &-\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_{l'})\right],\end{aligned} \tag{C.6a}$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\,\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)}\left[\mathcal{E}_j\,\lambda_r^*\right] =& \frac{1}{k}\sum_{r'=1}^{k}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma'(\lambda_j)\lambda_r^*\sigma(\lambda_{r'}^*)\right]\\ &-\frac{1}{p}\sum_{l'=1}^{p}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma'(\lambda_j)\lambda_r^*\sigma(\lambda_{l'})\right].\end{aligned} \tag{C.6b}$$

Define the vector $\boldsymbol{\lambda}^{\alpha\beta\gamma} \equiv \left(\lambda^\alpha,\lambda^\beta,\lambda^\gamma\right)^\top \in \mathbb{R}^3$, where the upper indices on the components indicate they may refer to student or teacher local fields. Consider the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta\gamma}$:

$$\Omega^{\alpha\beta\gamma} \equiv \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\boldsymbol{\lambda}^{\alpha\beta\gamma}\left(\boldsymbol{\lambda}^{\alpha\beta\gamma}\right)^\top\right] \in \mathbb{R}^{3\times3}. \tag{C.7}$$

For $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (C.6) are given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\left[\sigma'(\lambda^\alpha)\lambda^\beta\sigma(\lambda^\gamma)\right] = \frac{2}{\pi}\frac{\Omega_{23}^{\alpha\beta\gamma}\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)-\Omega_{12}^{\alpha\beta\gamma}\Omega_{13}^{\alpha\beta\gamma}}{\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)\sqrt{\left(1+\Omega_{11}^{\alpha\beta\gamma}\right)\left(1+\Omega_{33}^{\alpha\beta\gamma}\right)-\left(\Omega_{13}^{\alpha\beta\gamma}\right)^2}}, \tag{C.8}$$

where $\Omega_{jl}^{\alpha\beta\gamma} \equiv (\Omega^{\alpha\beta\gamma})_{jl}$ is an element of the covariance matrix given by Eq. (C.7). As examples, we write explicitly:

$$\Omega^{jlr'} = \begin{bmatrix} q_{jj} & q_{jl} & m_{jr'} \\ q_{jl} & q_{ll} & m_{lr'} \\ m_{jr'} & m_{lr'} & \rho_{r'r'} \end{bmatrix} \quad , \quad \Omega^{jrr'} = \begin{bmatrix} q_{jj} & m_{jr} & m_{jr'} \\ m_{jr} & \rho_{rr} & \rho_{rr'} \\ m_{jr'} & \rho_{rr'} & \rho_{r'r'} \end{bmatrix} . \tag{C.9}$$

The quadratic contribution in $\mathcal{E}_j$ is given by

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)}\, \mathbb{E}_{\zeta\sim\mathbb{P}(\zeta)} \left[\mathcal{E}_j\, \mathcal{E}_l\right] =& \frac{1}{k^2} \sum_{r,r'=1}^{k} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda_r^*)\sigma(\lambda_{r'}^*)\right] \\ &+ \frac{1}{p^2} \sum_{j',l'=1}^{p} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda_{j'})\sigma(\lambda_{l'})\right] \\ &- \frac{2}{pk} \sum_{l'=1}^{p} \sum_{r=1}^{k} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda_j)\sigma'(\lambda_l)\sigma(\lambda_r^*)\sigma(\lambda_{l'})\right] \\ &+ \Delta\, \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda_j)\sigma'(\lambda_l)\right] \end{aligned} \tag{C.10}$$

The solution of the noise-dependent term can be constructed with the covariance matrix (C.3) and is given by [6]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda^\alpha)\sigma'(\lambda^\beta)\right] = \frac{2}{\pi} \frac{1}{\sqrt{1+\Omega_{11}^{\alpha\beta}+\Omega_{22}^{\alpha\beta}+\Omega_{11}^{\alpha\beta}\Omega_{22}^{\alpha\beta}-\left(\Omega_{12}^{\alpha\beta}\right)^2}} \tag{C.11}$$

Similarly, one can define the vector $\boldsymbol{\lambda}^{\alpha\beta\gamma\delta} \equiv \left(\lambda^\alpha, \lambda^\beta, \lambda^\gamma, \lambda^\delta\right)^\top \in \mathbb{R}^4$ and write the covariance matrix on the subspace spanned by $\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}$:

$$\Omega^{\alpha\beta\gamma\delta} \equiv \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\boldsymbol{\lambda}^{\alpha\beta\gamma\delta} \left(\boldsymbol{\lambda}^{\alpha\beta\gamma\delta}\right)^\top\right] \in \mathbb{R}^{4\times4} . \tag{C.12}$$

For $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$ the expectations in Eqs. (C.10) are given by [5]

$$\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*\sim\mathcal{N}(\boldsymbol{\lambda},\boldsymbol{\lambda}^*|\mathbf{0},\Omega)} \left[\sigma'(\lambda^\alpha)\sigma'(\lambda^\beta)\sigma(\lambda^\gamma)\sigma(\lambda^\delta)\right] = \frac{4}{\pi^2} \frac{1}{\sqrt{\bar{\Omega}_0^{\alpha\beta\gamma\delta}}} \arcsin\left(\frac{\bar{\Omega}_1^{\alpha\beta\gamma\delta}}{\sqrt{\bar{\Omega}_2^{\alpha\beta\gamma\delta}\bar{\Omega}_3^{\alpha\beta\gamma\delta}}}\right) , \tag{C.13}$$

with

$$\bar{\Omega}_0^{\alpha\beta\gamma\delta} \equiv \left(1+\Omega_{11}^{\alpha\beta\gamma\delta}\right)\left(1+\Omega_{22}^{\alpha\beta\gamma\delta}\right) - \left(\Omega_{12}^{\alpha\beta\gamma\delta}\right)^2 , \tag{C.14a}$$

$$\begin{aligned} \bar{\Omega}_1^{\alpha\beta\gamma\delta} \equiv& \bar{\Omega}_0^{\alpha\beta\gamma\delta}\Omega_{34}^{\alpha\beta\gamma\delta} - \Omega_{23}^{\alpha\beta\gamma\delta}\Omega_{24}^{\alpha\beta\gamma\delta}\left(1+\Omega_{11}^{\alpha\beta\gamma\delta}\right) - \Omega_{13}^{\alpha\beta\gamma\delta}\Omega_{14}^{\alpha\beta\gamma\delta}\left(1+\Omega_{22}^{\alpha\beta\gamma\delta}\right) \\ &+ \Omega_{12}^{\alpha\beta\gamma\delta}\Omega_{13}^{\alpha\beta\gamma\delta}\Omega_{24}^{\alpha\beta\gamma\delta} + \Omega_{12}^{\alpha\beta\gamma\delta}\Omega_{14}^{\alpha\beta\gamma\delta}\Omega_{23}^{\alpha\beta\gamma\delta} , \end{aligned} \tag{C.14b}$$

$$\begin{aligned} \bar{\Omega}_2^{\alpha\beta\gamma\delta} \equiv& \bar{\Omega}_0^{\alpha\beta\gamma\delta}\left(1+\Omega_{44}^{\alpha\beta\gamma\delta}\right) - \left(\Omega_{24}^{\alpha\beta\gamma\delta}\right)^2\left(1+\Omega_{11}^{\alpha\beta\gamma\delta}\right) - \left(\Omega_{13}^{\alpha\beta\gamma\delta}\right)^2\left(1+\Omega_{22}^{\alpha\beta\gamma\delta}\right) \\ &+ 2\Omega_{12}^{\alpha\beta\gamma\delta}\Omega_{13}^{\alpha\beta\gamma\delta}\Omega_{23}^{\alpha\beta\gamma\delta} , . \end{aligned} \tag{C.14c}$$

$$\begin{aligned} \bar{\Omega}_3^{\alpha\beta\gamma\delta} \equiv& \bar{\Omega}_0^{\alpha\beta\gamma\delta}\left(1+\Omega_{44}^{\alpha\beta\gamma\delta}\right) - \left(\Omega_{24}^{\alpha\beta\gamma\delta}\right)^2\left(1+\Omega_{11}^{\alpha\beta\gamma\delta}\right) - \left(\Omega_{14}^{\alpha\beta\gamma\delta}\right)^2\left(1+\Omega_{22}^{\alpha\beta\gamma\delta}\right) \\ &+ 2\Omega_{12}^{\alpha\beta\gamma\delta}\Omega_{14}^{\alpha\beta\gamma\delta}\Omega_{24}^{\alpha\beta\gamma\delta} . \end{aligned} \tag{C.14d}$$

C.3 From gradient flow to local fields

Consider the gradient flow approximation

$$\begin{aligned}\frac{d\boldsymbol{w}_j}{dt} &= -\nabla_{\boldsymbol{w}_j}\mathcal{R}(\boldsymbol{W},\boldsymbol{W}^*)\\ &= -\frac{1}{p\sqrt{d}}\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\mathbb{1})}\left[\boldsymbol{x}\sigma'(\lambda_j)\,\mathcal{E}\right].\end{aligned}$$

Now, since for any $\boldsymbol{x}^\top\boldsymbol{y}$, we have

$$\frac{d\,(\boldsymbol{x}^\top\boldsymbol{y})}{dt} = \boldsymbol{x}^\top\frac{d\boldsymbol{y}}{dt} + \boldsymbol{y}^\top\frac{d\boldsymbol{x}}{dt},$$

we find

$$\frac{dq_{jl}}{dt} = -\frac{1}{pd}\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\mathbb{1})}\left[\left(\sigma'(\lambda_j)\lambda_l + \sigma'(\lambda_l)\lambda_j\right)\mathcal{E}\right].$$

Recalling the definition $\mathcal{E}_j = \sigma'(\lambda_j)\,\mathcal{E}$, the terms present inside the expectation are exactly those in the learning term of Eq.(14).

D Initial conditions and symmetric teacher

In this work we have constructed teacher matrices $\boldsymbol{W}^* \in \mathbb{R}^{k\times d}$ in order to have

$$\rho_{rs} = \frac{\boldsymbol{w}_r^{*\top}\boldsymbol{w}_s^*}{d} = \delta_{rs}\,, \tag{D.1}$$

where $\boldsymbol{w}_r^* \equiv [\boldsymbol{W}^*]_r \in \mathbb{R}^d$ is the r-th row of the matrix $\boldsymbol{W}^*$. We have started by sampling k vectors of dimension d uniformly on a ball of radius $\sqrt{d}$. Then we constructed an orthonormal basis using singular value decomposition.

The initial student weights $\boldsymbol{W}^0 \in \mathbb{R}^{p\times d}$ were taken as

$$\boldsymbol{W}^0 = \boldsymbol{A}\boldsymbol{W}^*\,, \tag{D.2}$$

with each row of $\boldsymbol{A} \in \mathbb{R}^{p\times k}$ sampled uniformly on a ball of radius one. We acknowledge choosing initial student weights as linear combinations of the teacher can be artificial and shrinks the first plateau, but our focus on this work was the specialization phase. Nevertheless, this choice and Eq. (D.1) are particularly suitable to theoretical analysis. Once k and p are fixed, the dimension d can be varied without changing $\boldsymbol{Q}^0$, $\boldsymbol{M}^0$ and $\boldsymbol{P}$, thereby removing any influence of different initial conditions for different d and providing the reader better visualization on the learning curves. To clarify this point, consider the j-th row $\boldsymbol{w}_j^0 \equiv [\boldsymbol{W}^0]_j \in \mathbb{R}^d$ of $\boldsymbol{W}^0$:

$$\boldsymbol{w}_j^0 = \sum_{r=1}^{k} a_{jr}\boldsymbol{w}_r^*\,, \tag{D.3}$$

with $a_{jr} \equiv [\boldsymbol{A}]_{jr}$. Using Eq. (D.1) one can write

$$q_{jl}^0 = \frac{\boldsymbol{w}_j^{0\top}\boldsymbol{w}_l^0}{d} = \sum_{r,r'=1}^{k} a_{jr}a_{jr'}\underbrace{\frac{\boldsymbol{w}_r^{*\top}\boldsymbol{w}_{r'}^*}{d}}_{=\delta_{rr'}} = \sum_{r=1}^{k} a_{jr}a_{lr}\,. \tag{D.4}$$

Similarly,

$$m_{jr}^0 = \frac{\boldsymbol{w}_j^{0\top}\boldsymbol{w}_r^*}{d} = a_{jr} \ . \tag{D.5}$$

Thus once $\boldsymbol{A}$ is fixed, the input dimension d can be varied without affecting the initial conditions. We chose to sample $\boldsymbol{a}_j \equiv [\boldsymbol{A}]_j \in \mathbb{R}^k$ on a ball of radius one both to introduce some randomness on the initialization and to keep the initial parameters bounded by one.

We stress that we use these initial conditions to make the data comparable for varying dimension d in the numerical illustrations. Our conclusions do not depend on this particular choice of initial conditions. If one simply takes random initialization $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{w}_j|\boldsymbol{0}, \mathbb{1})$ for each j, the full picture we have presented in this manuscript remains unchanged. In Figure 7 we present an example of curves within the blue region (see Section 3 for the characterization of this regime) with unconstrained Gaussian initialization. Dots represent simulations, while solid lines are obtained by integration of the ODEs given by Eqs. (23), with initial conditions adjusted to match simulations.

Although varying the initial population risk with d slightly changes the exact position where the specialization transition starts, the particular initial conditions adopted in this work do not affect whether the specialization transition takes place or not, comparing to unconstrained Gaussian initialization.

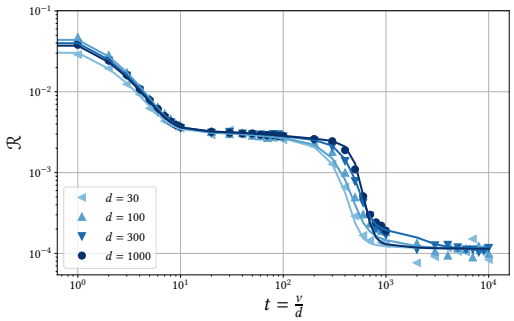


Figure 7: Population risk dynamics for $\kappa = \delta = 0$ (Saad & Solla scaling) : $p_0 = 8$, $k = 4$, $\rho_{rs} = \delta_{rs}$. Initialization: $\boldsymbol{w}_j \sim \mathcal{N}(\boldsymbol{w}_j|\boldsymbol{0}, \mathbb{1})$ for $j = 1, ..., p_0$. Activation function: $\sigma(x) = \mathrm{erf}(x/\sqrt{2})$. Data distribution: $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \mathbb{1})$. Dots represent simulations, while solid lines are obtained by integration of the ODEs given by Eqs. (23), with initial conditions adjusted to match simulations. Observe the difference on the initialization for different d.

References

[1] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. E7665–E7671, 2018.

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

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

[4] J. Sirignano and K. Spiliopoulos, "Mean field analysis of neural networks: A central limit theorem," *Stochastic Processes and their Applications*, vol. 130, no. 3, pp. 1820–1852, 2020.

[5] D. Saad and S. A. Solla, "On-line learning in soft committee machines," *Phys. Rev. E*, vol. 52, pp. 4225–4243, Oct 1995.

[6] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[7] W. Kinzel and P. Ruján, "Improving a network generalization ability by selecting examples," *Europhysics Letters (EPL)*, vol. 13, no. 5, pp. 473–477, nov 1990.

[8] O. Kinouchi and N. Caticha, "Optimal generalization in perceptions," *Journal of Physics A: Mathematical and General*, vol. 25, no. 23, pp. 6243–6250, dec 1992.

[9] M. Copelli and N. Caticha, "On-line learning in the committee machine," *Journal of Physics A: Mathematical and General*, vol. 28, no. 6, pp. 1615–1625, mar 1995.

[10] M. Biehl and H. Schwarze, "Learning by on-line gradient descent," *Journal of Physics A: Mathematical and General*, vol. 28, no. 3, pp. 643–656, feb 1995.

[11] P. Riegler and M. Biehl, "On-line backpropagation in two-layered neural networks," *Journal of Physics A: Mathematical and General*, vol. 28, no. 20, pp. L507–L513, oct 1995.

[12] D. Saad and S. Solla, "Dynamics of on-line gradient descent learning for multilayer neural networks," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. C. Mozer, and M. Hasselmo, Eds., vol. 8. MIT Press, 1996.

[13] R. Vicente, O. Kinouchi, and N. Caticha, "Statistical mechanics of online learning of drifting concepts: A variational approach," *Machine learning*, vol. 32, no. 2, pp. 179–201, 1998.

[14] D. Saad, Ed., *On-Line Learning in Neural Networks*, ser. Publications of the Newton Institute. Cambridge: Cambridge University Press, 1999.

[15] S. Mei, T. Misiakiewicz, and A. Montanari, "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. PMLR, 25–28 Jun 2019, pp. 2388–2464.

[16] D. Saad and S. A. Solla, "Exact solution for on-line learning in multilayer neural networks," *Phys. Rev. Lett.*, vol. 74, pp. 4337–4340, May 1995.

[17] M. Refinetti, S. Goldt, F. Krzakala, and L. Zdeborova, "Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8936–8947.

[18] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, "The gaussian equivalence of generative models for learning with two-layer neural networks," in *Proceedings of Machine Learning Research*, vol. 145. 2nd Annual Conference on Mathematical and Scientific Machine Learning, 2021, pp. 1–46.

[19] H. Hu and Y. M. Lu, "Universality laws for high-dimensional learning with random features," 2020.

[20] A. Montanari and B. Saeed, "Universality of empirical risk minimization," 2022.

[21] C. Wang, Y. C. Eldar, and Y. M. Lu, "Subspace estimation from incomplete observations: A high-dimensional analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1240–1252, 2018.

[22] Y. Yoshida, R. Karakida, M. Okada, and S.-i. Amari, "Statistical Mechanical Analysis of Online Learning with Weight Normalization in Single Layer Perceptron," *Journal of the Physical Society of Japan*, vol. 86, no. 4, p. 044002, Apr. 2017.

[23] P. Del Moral and A. Niclas, "A taylor expansion of the square root matrix function," *Journal of Mathematical Analysis and Applications*, vol. 465, no. 1, pp. 259–266, 2018.

[24] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová, "The committee machine: computational to statistical gaps in learning a two-layers neural network," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124023, dec 2019.

[25] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[26] L. Chizat, E. Oyallon, and F. Bach, "On lazy training in differentiable programming," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[27] F. Bach and L. Chizat, "Gradient descent on infinitely wide neural networks: Global convergence and generalization," *arXiv preprint arXiv:2110.08084*, 2021.

[28] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized Kaczmarz: theoretical guarantees," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 97–123, 04 2018.

[29] G. B. Arous, R. Gheissari, and A. Jagannath, "Online stochastic gradient descent on non-convex losses from high-dimensional inference," *Journal of Machine Learning Research*, vol. 22, no. 106, pp. 1–51, 2021.

[30] ——, "Algorithmic thresholds for tensor PCA," *The Annals of Probability*, vol. 48, no. 4, pp. 2052 – 2087, 2020.

[31] C. Wang, H. Hu, and Y. Lu, "A solvable high-dimensional model of gan," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.